

CLOUD COMPUTING



Prof. Soumya Kanti Ghosh

Department of Computer Science and Engineering
IIT Kharagpur

INDEX

<u>S.NO</u>	<u>TOPICS</u>	<u>PAGE.NO</u>
Week 1		
1	Lecture 1:	4
2	Lecture 2:	17
3	Lecture 3:	40
4	Lecture 4:	59
5	Lecture 5:	79
Week 2		
6	Lecture 6	101
7	Lecture 7	127
8	Lecture 8	154
9	Lecture 9	171
10	Lecture 10	189
Week 3		
11	Lecture 11:	219
12	Lecture 12:	240
13	Lecture 13 :	259
14	lecture 14 :	284
15	Lecture 15 :	305
Week 4		
16	Lecture 16	326
17	Lecture 17:	345
18	Lecture 18:	371
19	Lecture 19:	381
20	Lecture 20:	393

Week 5

21	Lecture 21 : SLA-Tutorial	408
22	Lecture 22 : Cloudonomics-Tutorial	424
23	Lecture 23 : MapReduce-Tutorial	439
24	Lecture 24 : ResourceMgmt-I	455
25	Lecture 25 : ResourceMgmt-II	476

Week 6

26	Lecture 26: Cloud Computing: Security I	494
27	Lecture 27: Cloud Computing: Security II	518
28	Lecture 28: Cloud Computing: Security III	540
	Lecture 29: Cloud Computing: Security Issues in Collaborative SaaS	
29	Cloud	563
30	Lecture 30; Cloud Computing: Broker for Cloud Marketplace	591

Week 7

31	Lecture 31: Mobile Cloud Computing -I	613
32	Lecture 32: Mobile Cloud Computing -II	637
33	Lecture 33: Fog Computing-I	654
34	Lecture 34: Fog Computing-II	670
35	Lecture 35:Use Case-Geo-spatial Cloud	691

Week 8

36	Lecture 36 : Introduction to DOCKER Container	721
37	Lecture 37 : Green Cloud	741
38	Lecture 38 : Sensor Cloud Computing	761
39	Lecture 39 : IoT Cloud	787
40	Lecture 40 : Course Summary and Research Areas	807

Week 9

41	Lecture - 41 Cloud–Fog Computing - Overview	823
42	Lecture - 42 Resource Management - I	833

43	Lecture 43 Resource Management - II	842
44	Lecture 44 Cloud Federation	851
Week 10		
45	Lecture 45 "VM Migration - Basics Migration strategies"	859
46	Lecture 46 "VM Migration - Basics Migration strategies"	867
47	Lecture 47 "Containers Container based Virtualization Kubernetes Docker Container "	875
48	Lecture 48 "Docker Container – Overview Docker – Components Docker – Architecture"	883
49	Lecture 49 Docker Container - Demo	892
50	Lecture 50 Docker Container - Demo	898
Week 11		
51	Lecture 51 Dew Computing	905
52	Lecture 52 Serverless Computing - I	913
53	Lecture 53 Serverless Computing - II	921
54	Lecture 54 Sustainable Cloud Computing - I	929
55	Lecture 55 Sustainable Cloud Computing - II	937
Week 12		
56	Lecture 56 Cloud Computing in 5G Era	946
57	Lecture 57 CPS and Cloud Computing	954
58	Lecture 58 Case Study I (Spatial Cloud Computing)	963
59	Lecture 59 "Case Study II (Internet of Health Things) (Part-A)"	973
60	Lecture 60 "Case Study II (Internet of Health Things) (Part-B)"	981

Cloud Computing
Prof. Soumya Kanti Ghosh
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Lecture – 01
Cloud Computing – Overview

Welcome to the course on cloud computing. Today we will have our first lecture. So, as you might have seen the broad overview of the course. So, in this particular series of lectures, we will try to give an overall picture of what cloud computing is, and what are its major components, and what are the recent trends. And at the end may be; what are the different types of research opportunities, or this trends of future trends in the cloud computing, right. So, before going to the details of cloud computing, we will try to have a quick overview of course, and the basic paradigm of computing.

(Refer Slide Time: 01:04)

Introduction

- The ACM *Computing Curricula 2005* defined "computing" as

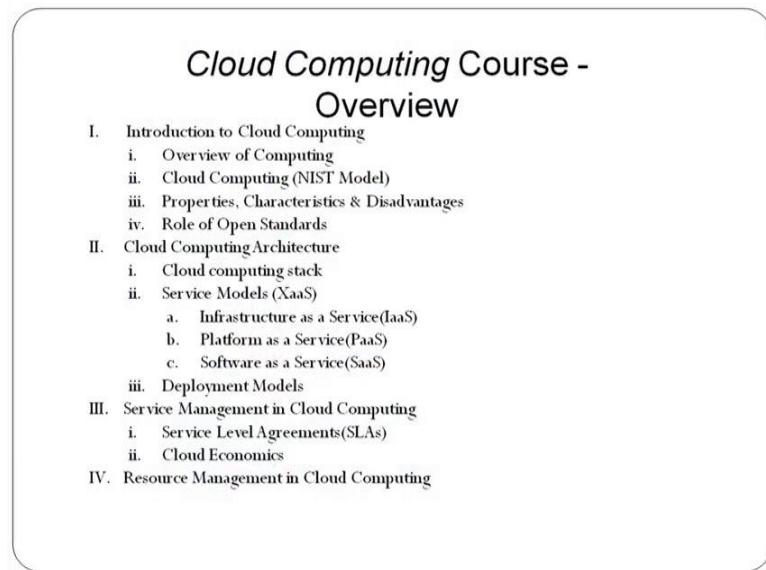
"In a general way, we can define computing to mean any goal-oriented activity requiring, benefiting from, or creating computers. Thus, computing includes designing and building hardware and software systems for a wide range of purposes; processing, structuring, and managing various kinds of information; doing scientific studies using computers; making computer systems behave intelligently; creating and using communications and entertainment media; finding and gathering information relevant to any particular purpose, and so on. The list is virtually endless, and the possibilities are vast."

2

Now, if you look that as defined by ACM computing curricula in 2005; as they defined computing, it is a general way we can define computing to mean; as a mean to solve any goal oriented activity, right. So that means, it can include starting from hardware software system for a wide range of purposes, and also making computing systems intelligent using communications, finding gathering information from relevant to any particular purpose and so on. So, if you look at it has anything where some sort of a computing is needed, it falls under the computing paradigm. So, this gives us broad

spectrum of thing; not only in terms of resources, also in terms of the terms are category a level of people who can who are going to use it. Starting from a high end researcher or a professional, to a student, to even to a housewife, or a citizen in general, look one to use it for its benefit or something which serves particular purpose.

(Refer Slide Time: 02:18)



Before going to other overview of this computing, I will; let us try to reiterate the type of a codes or type of things, we will like to cover.

So, initial lectures, we will have a more things most likely today and may be something on the next day, that introduction to cloud computing which gives overview what that NIST models says, what are the typical properties, characteristics advantage and disadvantages of cloud computing or role of open standards. So, whether there is a standardization need or things, then we will look at more on a cloud computing architecture, like what is the typical computing cloud computing stack moving towards a service oriented architecture. So, what sort of service models are available in cloud like typically infrastructure as a service, platform as a service, or software as a service, or anything as a service; later we will see that anything as a service, whether we can realize; what are the different deployment models, right in case of a cloud how; when I want to deploy, whether it is in a; what all be the different deployment models of a cloud.

Then one of the another major aspect of cloud is the service management. Like as we whenever we try to purchase any service, or whenever I want to leverage any service,

there is a need of service management, like from the say consumer end, I would like to have what is the guarantee of minimal services, from the provider end, the cloud provider or cloud service provider CSP one to see that may be the profit or may be may be that how this guarantee; what are the resource requirement at the back end to surf so much computers.

So, from the; if you look at the provider consumer for any type of services, not only cloud services, any type of services in our day today life, we require some sort of a agreement between the service provider and the consumer. What in; what we say something call; service level agreement, like I want to say that my availability will be 100 percent or near 100 percent, based on my thing like I say when a exam is going on, I want to have redundant services, So that the availability of the resources is 100 percent or very near to 100 percent; whereas, when my practice session is going on, requirement of availability may come down to 90 percent.

Now based on the availability, the resource pooling or resource management will be done by the at the provider end. And the provider will charge based on the type of type of his resources type of availability etc. Then there are issues of down time what will be there are issues of quality of services, there are several other issues that we will try to discuss under the paradigm of service level agreements and other things. There is another important person another important aspects is cludonomics, or economy of using cloud computing. It may not be whether it is always good that if I use cloud, it will be beneficial, whether it is true like it is as we see like suppose if you want to commute 220 kilometer per day for your office or work. Then it may be economical to purchase a car, right, but if you are commuting say even 50 kilometer or 100 kilometer, once in a month it may not be economical then purchasing a car, right it may be more economical than hiring a car, right. Similarly, when I suit hire when I suit purchase whether there is a relationship, whether there is a economic model behind it or what if at all how to how do I from my say organization point of view, may be from a particular say even point of view, whether I can see that whether purchasing or hiring a resources or is economical or what is the what is the economic model of the things.

So, that type of things economics in cloud or economy in cloud we have to see, another aspect is the resource management. Like this is more in the service provider end right, or cloud service provider, how these resources will be manage right like I. So, what I what I

see that I have I need to surf so many people. So, what sort of resources I need to manage at the things. This is true for anything like if I say if I have a stationary shop or who which takes care of stationary is related to say academic things like I says a note books pen and etcetera, etcetera. So, how much I need to stock or it depend on how much maybe myself projection etcetera or so, I do not have some a situation when I am starving for my store or I should not have a situation when I am say my shop is full and I need to keep some thing outside the soft type of things, right.

So, it is I should not have a overloading. So, I have a proper resource management. So, it is very tricky when we have a computing as a resource, or when I provide computing as a resource, then I have to manage several type of resources. Like typically of I look at a in a typical computing system forget about cloud or anything. So, what are the things we are basically looking for? Maybe one maybe the processor or the CPU popularly, or maybe or one maybe the your working memory or popularly the RAM or hard disk and maybe network connectivity and there are other several other resources which are they are right.

So, how much resources I need to maintain, manage, etcetera, right. So, any resources has a inherent costing into the things. So, if I need to manage huge volume of resources without utilization, then I have to incur what we say more cost on the resources, or than in maintaining the things. So, this appropriate or optimally management of resources is a serious challenge. And they are here like to see that; what are the different type of resource management issues in these particular cloud computing thing. So, other aspects of these cloud computing one is the data management, right.

(Refer Slide Time: 08:40)

Cloud Computing Course *(contd.)*

- V. Data Management in Cloud Computing
 - i. Looking at Data, Scalability & Cloud Services
 - ii. Database & Data Stores in Cloud
 - iii. Large Scale Data Processing
- VI. Cloud Security
 - i. Infrastructure Security
 - ii. Data security and Storage
 - iii. Identity and Access Management
 - iv. Access Control, Trust, Reputation, Risk
- VII. Case Study on Open Source and Commercial Clouds, Cloud Simulator
- VIII. Research trend in Cloud Computing, Fog Computing

So, data is the very tricky thing, like we look to look at this how these data will be stored manage, scalability, and cloud services over this data services over that. If it is a not only data of it is a data bases and data stores. So, there is a separate type of looking at the data things in the cloud. And if it is a large scale data processing, then I need to look at the how this data management will be there. So, our conventional way of approaching normal data or data base management system whether it is still good for cloud, or whenever I want to give as data storage or type of services. So, what type of things; I need to do for the things like we are popularly using different type of storage as a service stuffs and in our day today life like one of the popular thing may be the dropbox. So, how things at the background need to be managed? So, it is not like that I need to build a data services all the qualities, but at least looking at that what are the architecture and what are the issues in data management type of things.

Another major aspects of a cloud is security right. So, your data is in some other place, you are computing in some other others domain. So, what will be the different type of security aspects? So, at it has like what will be the security which has a infrastructure as thing, or what are data related or storage related security, because data is a important aspects of our all things like, what many people say that you can regenerate a or reinstall an application, but you cannot reinstall your data right, write a report of 100 or 10 pages and it gets system get crashed and then data is lost along with the application I can reinstall the OS, I can reinstall the Word processing tool, but reinstalling the data is not

possible. That particular report is not possible if not; you have a recovery mechanism expect them, right.

Whenever I have on my personal system like it may be personal computing desktop or in a laptop or wherever, then it is my responsibility to take backups or I have redundancy things in that things. But whenever I store the data into the others place then one is one is how things are saved and type of things in case of loss or another typical thing comes up whether my data is being accessed or over retained ride somebody else right. So that means, whether this; what is the security of this particular data. There are issues of identity and access management. So, this is this is another important aspects where particular identity and access management of the collaborating parties need to be there. There are issues of access control trust reputation risk.

So, there are several issues like how access control will be there whether it is a our standard access control mechanism. So, whether role base access control mechanism or whatever things we can we work on the security, how much trust on having I have a cloud service provider whether I trust service provider one more than the service provider 2 or whether it is how to calculate a particular service provider. So, there are issues of the reputation I want to look at a reputation there are issues of risk of losing data losing application losing your, because your own customers like you are purchasing cloud to surf somebody, right.

So, you may intern things. So, this trust reputation risk goes somewhere what we say 3 nodes of a triangle. So, they are interlinked have a in for any systems; they have a lot of what we say; a lot of influence on working of the whole systems, right. So, I need to assure that how it is assured in the in cloud computing paradigm that you see. Then we will try to look at some of case studies or some what we say demo type of things on open source and on commercial cloud may be some cloud simulator. There are various commercial cloud in things in the in the market. So, we will try to see that; what are the basic property or how they work etcetera, then there are open source cloud.

So, we are try to see that how we open source things are there, even we will if time permits we will try to see that; what are the different type of installing a open source cloud. And there are few cloud simulators also. So, we like to; if it is time permits we like to see the simulator. And at the end of the things as one of our major motivation of

this academic world to take things in a in future, right. We want to see that something more in the future. So, we will try to look at the recent trend in cloud computing.

So, those who are interested in research or even some of projects in pg level ug level. So, they can they can have a pointer that what are the different aspects of sentence in cloud computing, there are there are people are talking about fog computing and other different technologies. So, we like to see that; what are the different aspects of those things. So, this is broadly the overall code structure we will try to give a proper weightage based on the importance of the course we will and we will give more details as an when we will we will basically going through those lectures, right.

(Refer Slide Time: 14:30)

Trends in Computing

- Distributed Computing
- Grid Computing
- Cluster Computing
- Utility Computing
- Cloud Computing

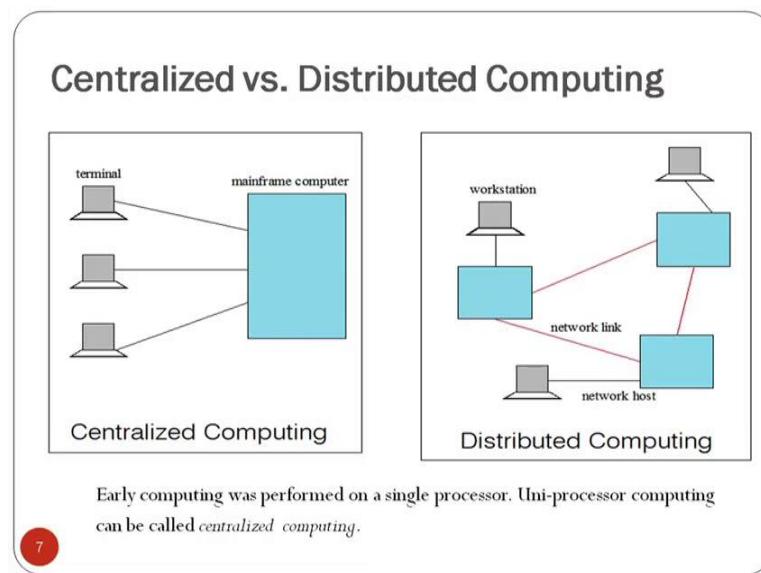
5

So, with this; we will try, we will have a quick overview of; what are the different computing trend which actually made this cloud computing a reality. So, it is not that it is from the day one something was there. So, it as we say that all invention or all any type of development is primarily come up with some necessity or requirement of the what we say community scientific community, or even general citizen at large. So, that drives that to what think is there one of the things there are definitely sky searches where we it is driven by that in own things, but we will like to see that so much computing is already was in place or is in place while still it has the importance. Whether it is a totally new baby or new stuff or it is a amalgamation or evolve through the things, right.

So, what we see it is in different literatures or even in if you look at the cloud computing as all. It is not a suddenly new stuff which came into the play. It has evolved and it has different other development which is already in place which has basically helped in bringing this into play. So, if we look at that different type of computing paradigm which are or which were they are for long time and still in a it is there in a big way.

So, one first of all; the mother of the things all those things is known as the distributed or people say that it is a distributed computing, right. So, distributed computing, then we have differ and other computing, it is not that it came in the sequence like one after another, but it is more of the these are the different aspects what we look at the things. So, it is a distributed computing we have grid computing, we have cluster computing, we have utility computing and we are talking about cloud computing. Now if we see that these different development where different needs, everybody has advantages; some disadvantages, and they helped in making some other things in a feasible way. So, we will go quickly because these are some of the things already known to you, and are available in the literature just to have that why what is the how it came up this cloud computing may be things.

(Refer Slide Time: 17:07)



So, if you look at distributed computing, so we started with or still; we are; we work with centralized computing, like primarily in previous days, we used to use mainframe, where

different terminals are there. So, jobs are submitted to the mainframe that get executed and being used by the or being viewed by the user.

So, primarily it is something which has a logically single processing thing, right. So, or what we say some sort of a uni processor computing or centralized computing type of things. Now also in different places is there; it is not like that we need to throw out the things there is a particular necessity of the things and if these are still useful in several places and being used at several areas. So, the other thing which evolved is the distributed computing, where you have different systems distributed over a particular geographical space. Typically it may vary from a lab type of a scenario, to a scenario where you have large geographical boundaries also. Again depends on the type of requirements are there, right. And one important aspect came up is that network link availability of seamless network connectivity during between this collaborating systems, like they or what we say different network systems, right.

(Refer Slide Time: 18:43)

Distributed Computing/System?

- Distributed computing
 - Field of computing science that studies distributed system.
 - Use of distributed systems to solve computational problems.
- Distributed system
 - Wikipedia
 - There are several autonomous computational entities, each of which has its own local memory.
 - The entities communicate with each other by message passing.
 - Operating System Concept
 - The processors communicate with one another through various communication lines, such as high-speed buses or telephone lines.
 - Each processor has its own local memory.

The diagram illustrates a distributed computing system. It shows three separate nodes, each consisting of a 'Processor' (blue box) and a 'Memory' (red box). Arrows point from the Processor of one node to the Memory of another, representing communication between the nodes. A bracket on the right side of the diagram is labeled 'Distributed Computing'.

So, it is basically a field of computing science that studies distributed systems, it was there for long time; use of distributed system to solve computational processes; right. There are different other type of other definitions which come up if you look at the internet several definitions come.

So, there are several; it is one says that there are several autonomous computational entities each of has its own local memory. So, it is separate autonomous independent

computing entities having their own local memory. The entities communicate with each other by message passing over a backbone communication network, right. So, that is one thing if we look at the operating system point of or way of the concept the processor communicates with each other through various communicational line say at high speed busses or even telephone lines where the things each processor has it is own local memory, right.

(Refer Slide Time: 19:42)

Example Distributed Systems

- Internet
- ATM (bank) machines
- Intranets/ Workgroups
- Computing landscape will soon consist of ubiquitous network-connected devices

9

So, there are several type of example people put different things in the distributed computing paradigm, starting from over internetworking is a distributed system. Or this ATMs, bank machines, different branches of the banks or even different collaborating. And doing executing different functions that can be a things. Intranet or workgroups within the Internets may be a distributed system. Computing landscape will soon consist of ubiquitous network connected devices, right or rather not will be it is already we have ubiquitous network connected devices. So, or what we say it is something ad hoc type of establishment which comes and type of things.

And these days we see different type of networks which are which form as a adhoc network they are different volatile like one example is vehicular adhoc networks right like vehicles smart vehicles with their own on board units communication path once they come together they form ad hoc network, and it executes different type of function like

maybe safety related things may be entertainment related or infotainment related type of things and different type of stuff are there.

(Refer Slide Time: 21:02)

Computers in a Distributed System

- *Workstations:* Computers used by end-users to perform computing
- *Server Systems:* Computers which provide resources and services
- *Personal Assistance Devices:* Handheld computers connected to the system via a wireless communication link.

10

So, if we look at the broad type of computers in distributed systems. So, they are primarily what we say workstation, server systems and personal assistance devices like it may so. workstation is computers which are in the end user to perform computing. Server systems which works on a which give some provide some services per say. So, computers which provide resources and services right there can be personal assistance devices like handheld computers connected to systems where wireless communication network, I can be any type of things like any type of communication paradigms which helps in communicating with a things. So, these are the different what we say typical end nodes in a distributed systems, right. There can be other type of nodes also, like which has more network capabilities network processing type of things etcetera, but this what we can say broadly these are the typical loads in a in a typical distributed system.

(Refer Slide Time: 22:06)

Common properties of Distributed Computing

- Fault tolerance
 - When one or some nodes fails, the whole system can still work fine except performance.
 - Need to check the status of each node
- Each node play partial role
 - Each computer has only a limited, incomplete view of the system.
 - Each computer may know only one part of the input.
- Resource sharing
 - Each user can share the computing power and storage resource in the system with other users
- Load Sharing
 - Dispatching several tasks to each nodes can help share loading to the whole system.
- Easy to expand
 - We expect to use few time when adding nodes. Hope to spend no time if possible.
- Performance
 - Parallel computing can be considered a subset of distributed computing

11

So, if we look at the why such of thing some common properties or common advantages, or what we say benefits of distributed system one is fault tolerant like you have one or more means several systems are working. So, even with some node failures it works faithfully, right or may be at a lower performance, but it is not totally out of service right, had it been a centralized system. So, if down the whole thing is down what do you get to do something in a lower thing. So, it also to make it fault tolerant, there are different mechanism etc. Many of you may be knowing and to make the things. So, there are other thing that each node, another typical aspect is each node play it is partial role, right.

So, each node in the distributed system plays it is partial role, there is another aspects of or a property of resource sharing the share resources among themselves, there is a load sharing. So, what is not only resource sharing that computing resource sharing, but also load sharing like if it is a load or what we say that load balancing among the things can be realized, easy to expand, so, usually systems may be like that that we can easy to expand like I can have I can add distributed system more system into the network as and when as and when I have it or use it. Performance is a issue. So, parallel computing can be considered as a subset of distributed systems where I can have higher performance and need to be monitor.

(Refer Slide Time: 23:48)

Why Distributed Computing?

- Nature of application
- Performance
 - Computing intensive
 - The task could consume a lot of time on computing. For example, Computation of Pi value using Monte Carlo simulation
 - Data intensive
 - The task that deals with a large amount or large size of files. For example, Facebook, LHC(Large Hadron Collider) experimental data processing.
- Robustness
 - No SPOF (Single Point Of Failure)
 - Other nodes can execute the same task executed on failed node.

12

So, what we will, so another aspect of distributed system is that why we require maybe the nature of application demands it, maybe the different performance like I have computing intensive, data intensive type of things. And in some of the cases I require robustness into the system that should be no single point failure. I don't want any single point failure I may be doing a miss and critical things which may not be very computing intensive or memory intensive, but I can not afford to do any failure on the system, right. So, in the several cases there is a need of the things or in other sense this need primarily one of the primary what we say motivation of developing or development of this distributed systems. So, we will we will break for now and we will continue our discussion in the subsequent in the next lecture.

Thank you.

Cloud Computing
Prof. Soumya Kanti Ghosh
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Lecture – 02
Cloud Computing – Overview (Conld.)

Hi. So, we will continue our discussion on overview of computing and evolution of cloud computing. So, where we end in the last talk is where we are talking about the distributed systems, right.

(Refer Slide Time: 00:32)

Common properties of Distributed Computing

- Fault tolerance
 - When one or some nodes fails, the whole system can still work fine except performance.
 - Need to check the status of each node
- Each node play partial role
 - Each computer has only a limited, incomplete view of the system.
 - Each computer may know only one part of the input.
- Resource sharing
 - Each user can share the computing power and storage resource in the system with other users
- Load Sharing
 - Dispatching several tasks to each nodes can help share loading to the whole system.
- Easy to expand
 - We expect to use few time when adding nodes. Hope to spend no time if possible.
- Performance
 - Parallel computing can be considered a subset of distributed computing

11

So, I believe that we are in this particular site or in this slide, why we require these distributed system. So, as we were discussing; so, the nature of the application is one of the driving force, performance is another major driving force like why if the some of the applications, some of the requirements are computing intensive or some of the things can be data intensive.

(Refer Slide Time: 00:49)

Why Distributed Computing?

- Nature of application
- Performance
 - Computing intensive
 - The task could consume a lot of time on computing. For example, Computation of Pi value using Monte Carlo simulation
- Data intensive
 - The task deals with a large amount or large size of files. For example, Facebook, LHC(Large Hadron Collider) experimental data processing
- Robustness
 - No SPOF (Single Point Of Failure)
 - Other nodes can execute the same task executed on failed no

12



So, this requires distributed computing. So, another aspect is the robustness, right like as we were discussing that there should not be any single point of failure, my application does not want any single point of failure; that means, always it should be on. Even at a low performance level, I cannot say it should be a failure. In case of a centralized systems if the system fails, everything drops down, but in this case, we still work on a on the things.

Another other nodes can execute the same does executed on the failed nodes. So that means, I evolve as a system that if it is a failed, then the task which is working on the failed node can be executed or shared the load by the other node. So, this type of technology, algorithms are possible to develop or are being used in case of a distributed system.

(Refer Slide Time: 01:54)

Distributed applications

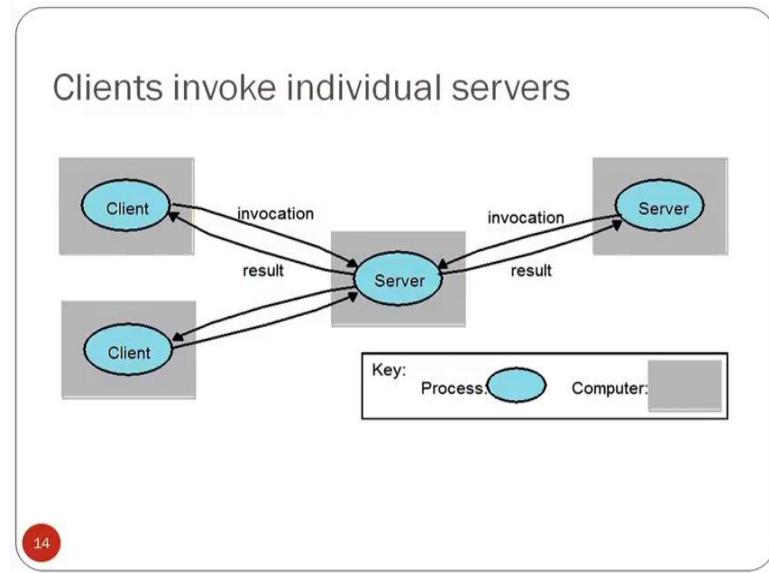
- Applications that consist of a set of processes that are distributed across a network of machines and work together as an ensemble to solve a common problem
- In the past, mostly “client-server”
 - Resource management centralized at the server
- “Peer to Peer” computing represents a movement towards more “truly” distributed applications

13

So, there are several another several distributed applications. So, application that consist of set of processes that are distributed across network of machines and work together as an ensemble to solve a common problem; In other sense, I have a several applications which coordinate among themselves to address a particular problem. So, this is; as if is a group of people working to address a particular problem. This is primarily useful for a large scale application development or large scale operational need, right where you have different operation being serve as different, and we have we realize a particular overall job at the end, right. So, there are not only computing aspect, there are several other aspects into the things as those who have worked or those who are read other things, there is a need of or orchestration of this processes, or orchestration of the services like who will do, etcetera. So, we have to form there. So, there are different aspects, but there is one of the major applications of this type of distributed computing.

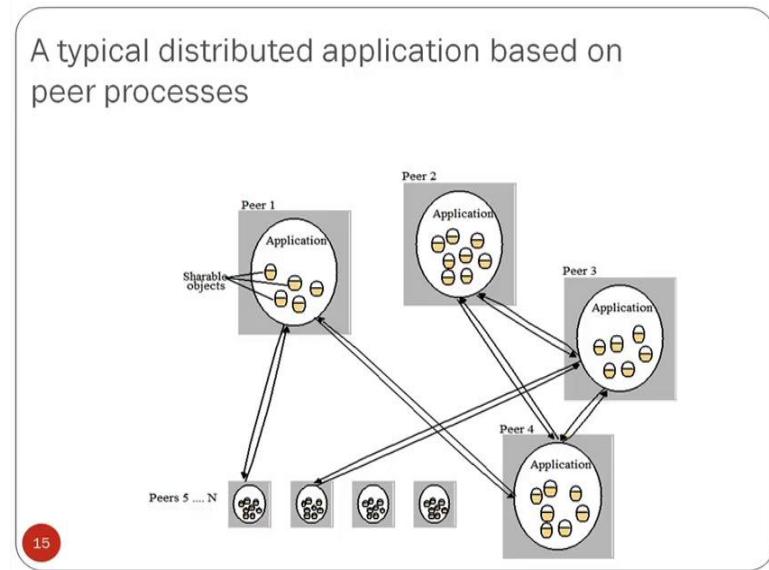
And not only in the past; now also, it is mostly, several application are client server type of things, resource management centralized at the server. So, we want to make it an a distributed fashion. There are peer to peer computing which represents a movement towards truly distributed applications, right. So, there are other types of motivating forces or motivations towards the distributed system.

(Refer Slide Time: 03:34)



So, this is typically clients invoke individual servers we are used to that; like client invoke a particular server. So, invocation and results, the sever can as a client for another server, right and there are there be can be there. So, I can say that all those things works in a individually may be clients sever mode, but they are basically trying to realize a overall particular job.

(Refer Slide Time: 04:00)



So, I can have a typical distributed application based on peer processes. So, there are different peers, there are different applications which are running in the peers, and it can

so happen that; these applications talk to each other to realize a particular job right. So, there are applications based on this peer processes. So, these are different types of distributed computing paradigm, another computing paradigm which became popular rather still it is very much popular is the grid computing, right.

(Refer Slide Time: 04:42)

Grid Computing?

- Pwwebopedia.com
 - A form of networking unlike conventional networks that focus on communication among devices, grid computing harnesses unused processing cycles of all computers in a network for solving problems too intensive for any stand-alone machine.
- IBM
 - Grid computing enables the virtualization of distributed computing and data resources such as processing, network bandwidth and storage capacity to create a single system image, granting users and applications seamless access to vast IT capabilities. Just as an Internet user views a unified instance of content via the Web, a grid user essentially sees a single, large virtual computer.
- Sun Microsystems
 - Grid Computing is a computing infrastructure that provides dependable, consistent, pervasive and inexpensive access to computational capabilities

17

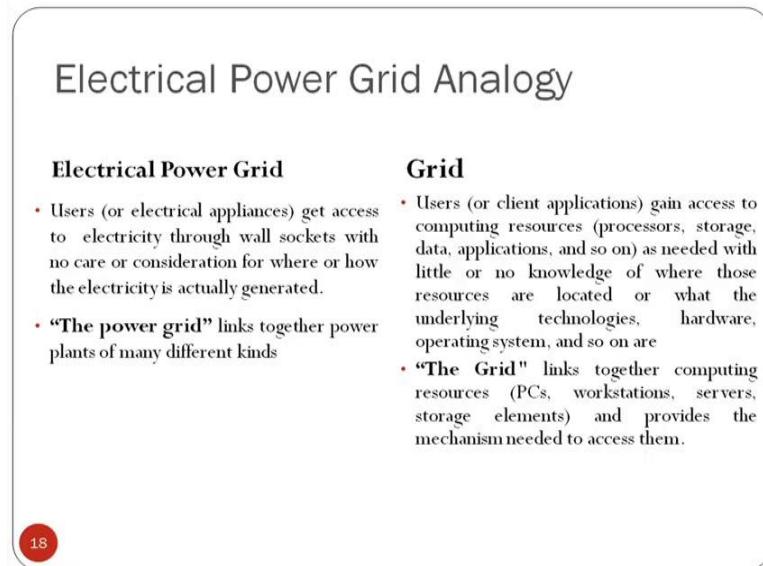
There are again different definitions of grid computing, and if you look at some of them. So, a form of networking unlike conventional network, that focus on communication among devices, grid computing harnesses unused resource cycle of all computers in the network, can solve problems intensive for standalone systems. So that means, I have a network of systems much stronger couple, solved some particular problems. So, if you look at other type of definitions grid computing enables virtualization of distributed computing, and data resources such as processing, network, bandwidth, storage capacity to create a single system image granting user and application, seamless access to the vast IT applications.

This is another important way of looking at it; that means, I have a distributed computing paradigm and then I want to realize or virtualize another system over the things, right which is one of the prime over of this cloud computing paradigm, right. I have several systems over the network, they have different processes etcetera, then I want to have a system evolved out of the things like I or popularly I can say that it is a I want to have

one or more virtual machines, with defined resources based on my requirement, right, this is one way of looking at it.

So, there are these are the different aspects and then grid computing is a computing infrastructure, that provides dependable, consistent, pervasive, inexpensive access to computing capability. In other sense I have resources which are available over the network.

(Refer Slide Time: 06:22)



And one popular thing we have in our day today life is or hear about the things or experience the thing is the electrical grid, right. So, one good analogy is an electrical grid. So, electrical grids are running and we tap power, we in the sense that organizations or households or the power distribution systems tap power from the grid. So, it is something available. So, in case of a; you can have computing analogies like users or client applications gain access to the computing resources, processors storage data applications so on as needed with little or more no knowledge of where those resources are located.

That means I it is a way of looking I have resources, I tap those resources and use it for my thing say I want to run something on a computing grid. So, I am more bothered about my own algorithm, methodology and I what I think that; what is the resource requirement? Now if I that particular grid particular computing grid allow me. So, I need to hook into the grid and run the resources, run my program or my; run my processes and

it detunes some the resources. So, the grid links together computing resources like PC, workstation, etcetera, and provides the mechanisms needed to accessing.

So, if I have agreed. So, what I require from the user point of view? A mechanism accesses this grid.

(Refer Slide Time: 07:52)

Grid Computing

1. Share more than information: Data, computing power, applications in dynamic environment, multi-institutional, virtual organizations
2. Efficient use of resources at many institutes. People from many institutions working to solve a common problem (virtual organisation).
3. Join local communities.
4. Interactions with the underneath layers must be transparent and seemless to the user.

19

So if you look at grid computing characteristics. So, say are more than information it is not only information, it is something data computing per so and so, efficient use of resources at many organization or institutes. So, I can have an efficient use of resources like if you look at our typical computing lab of a UG-PG classes, they accept the lab classes as they are mostly underutilized. So, I could have form a computing grid during the off office hours, specially off academic hours. So, that other researches could have used as the as a computing platform. So, my PCs set of 100 PCs everything is fine.

I basically I virtualize with a layer of another middleware. So, that people can work and as a enduser I do not care that what is going at the other end. So, long my program my processes runs faithfully with a particular performance level, which is desired by me. So, this is a good chance of using underutilized resources. There is a join local communities. So, I can have different type of communities like I can say some grid resources for some biological sciences, some arts sciences some genetic research. So, there are not only computing, but they also give some basic processes to do that right. So, those things are available.

So, in other sense I want those resources without having to purchase install maintain at a my end. Interaction with underlying layers, must be transferred and seamless to the user. So that means, I have a interface as a end user, I should not be bothered about what is going on at the other end of the things like how you are using the other leverage resources etcetera. You may ask me for money like if may ask that you need to pay so much amount for so much utilization that is fine, but I may not be willing to I do not want to go to the nitty-gritty of how you maintain your resources, how you maintain your network, how what is the services etcetera that is your thing.

(Refer Slide Time: 10:07)

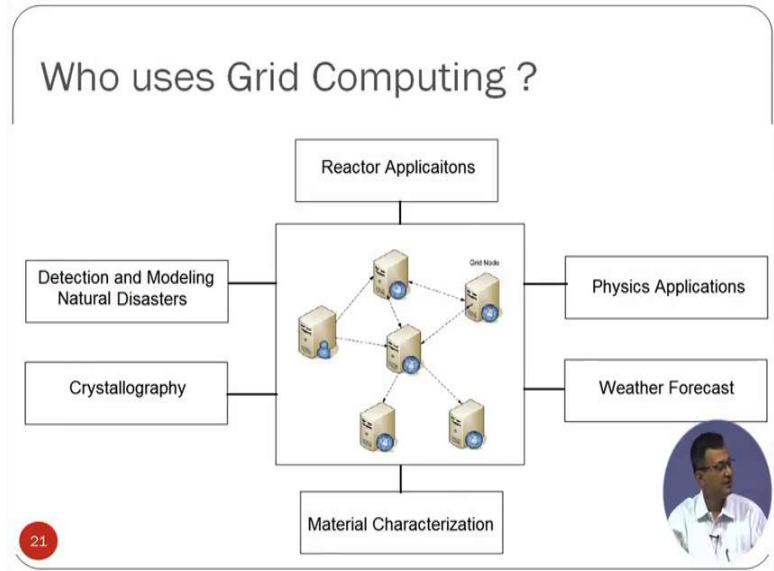
Need of Grid Computing?

- Today's Science/Research is based on computations, data analysis, data visualization & collaborations
- Computer Simulations & Modelling are more cost effective than experimental methods
- Scientific and Engineering problems are becoming more complex & users need more accurate, precise solutions to their problems in shortest possible time
- Data Visualization is becoming very important
- Exploiting under utilized resources

20

So, need of grid computing is well known. So, especially scientific research community is a big user of the things, specially data analysis, data analytics, visualization collaboration between the scientific wall, this days across continent people collaborate. Computer simulation and modeling are other common use; scientific and engineering problems are becoming more complex and need more accurate precise solution for their problem. Data visualization is becoming an important aspects of the thing. Exploiting underutilized resources maybe one of the motivating forces right. I have lot of resources which are underutilize; I am want to do that. Again this is another stepping stone for towards your cloud computing.

(Refer Slide Time: 10:56)



So, who uses grid? There are as we say that finally, there can physics application; weather application, material science reactor application and the so on and so forth; very few of them are there, but it you can see that all paradigm or of research or all paradigm of different scientific activities are there where who may be need grid.

(Refer Slide Time: 11:21)

Type of Grids

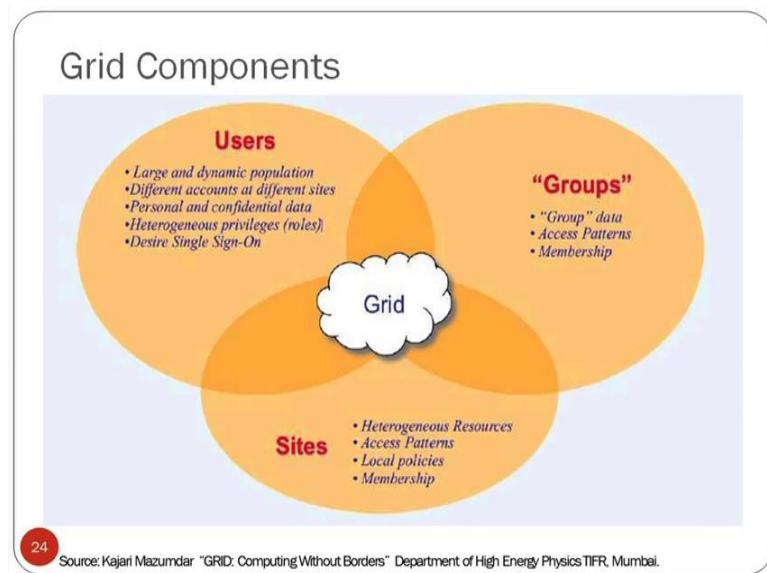
- **Computational Grid:** These grids provide secure access to huge pool of shared processing power suitable for high throughput applications and computation intensive computing.
- **Data Grid:** Data grids provide an infrastructure to support data storage, data discovery, data handling, data publication, and data manipulation of large volumes of data actually stored in various heterogeneous databases and file systems.
- **Collaboration Grid:** With the advent of Internet, there has been an increased demand for better collaboration. Such advanced collaboration is possible using the grid. For instance, persons from different companies in a virtual enterprise can work on different components of a CAD project without even disclosing their proprietary technologies.

22

So, there are different type of grids. We will not go to much nitty-gritty, there is one maybe the computational grid more on the computing, one may the data grid which serves as a more of a data storage.

One maybe collaboratively grid with that means, so that it helps in doing collaborate physics; there are other grids like network grids. So, providing fault tolerant high performance communication services over the things, there can be utility grid in this is a utility form of grid in which not only data and computation cycles are shared, but software or just about any resources are shared of the things. So, I say this is a utility grid.

(Refer Slide Time: 11:59)



So, there can be different players into the thing, definitely grid is a central thing that is the; which provides the thing. There are users who uses grid there are groups which has a group activity, there are different sides or what we resource availability locations of a different heterogeneous sites, etcetera. Not only that; there are issues of policy maintaining and other things like which can be share which cannot be shared, which can be computed where a central things are there.

Then another important aspects towards computing which is pretty popular is the cluster computing. So, what is a cluster?

(Refer Slide Time: 12:42)

What is Cluster Computing?

- A cluster is a type of parallel or distributed computer system, which consists of a collection of inter-connected stand-alone computers working together as a single integrated computing resource.
- Key components of a cluster include multiple standalone computers (PCs, Workstations, or SMPs), operating systems, high-performance interconnects, middleware, parallel programming environments, and applications.

26



A cluster is a type of parallel or distributed computing platform, which consists of collection of interconnected stand alone computing computers working together in a single integrated computing resources, right. So, key components are stand alone computers may be PC workstation or SMPs operating system, hyper performance interconnects, middleware parallel programming environment and applications.

So, there are different these are there are different these are different components of the cluster computing.

(Refer Slide Time: 13:16)

Cluster Computing?

- Clusters are usually deployed to improve speed and/or reliability over that provided by a single computer, while typically being much more cost effective than single computer the of comparable speed or reliability
- In a typical cluster:
 - Network: Faster, closer connection than a typical network (LAN)
 - Low latency communication protocols
 - Loosely coupled than SMP

27



So, clusters are usually deployed to improve speed or reliability over that provided by the single computer, right typical clusters are like having the properties of the network faster, closer connection then a typical LAN, low latency of communication protocol and loosely coupled than our SMPs.

(Refer Slide Time: 13:41)

Types of Cluster

- High Availability or Failover Clusters
- Load Balancing Cluster
- Parallel/Distributed Processing Clusters



28

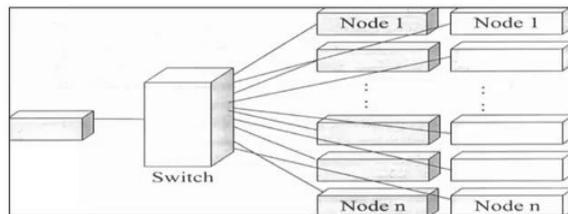
So, there are several type of clusters, right, few of them are high availability or failover clusters, like it say the resources are so be highly available if there is a any failure of the cluster node, the other things takes care etcetera. There are things are load balancing cluster like if I have particular processes. So, the load balancing is done by the cluster.

There are parallel distributed processing clusters, right, these are these helps or this facilitates parallel and distributed processing. So, these are the clusters which are there.

(Refer Slide Time: 14:17)

Cluster Components

- Basic building blocks of clusters are broken down into multiple categories:
 - Cluster Nodes
 - Cluster Network
 - Network Characterization



29

So, if you look at typical cluster nodes. So, there is a there is several cluster node. So, and a interconnect of network which connect the clusters, and there are different network characterization of the things like how it will be connected, what will be the different type of mechanisms so on the things.

(Refer Slide Time: 14:35)

Key Operational Benefits of Clustering

- System availability: offer inherent high system availability due to the redundancy of hardware, operating systems, and applications.
- Hardware fault tolerance: redundancy for most system components (eg. disk-RAID), including both hardware and software.
- OS and application reliability: run multiple copies of the OS and applications, and through this redundancy
- Scalability: adding servers to the cluster or by adding more clusters to the network as the need arises or CPU to SMP.
- High performance: (running cluster enabled programs)

30

So, there are several operational benefit; one is the system availability is one of the thing, fault tolerance is one of the thing, scalability another aspects, OS and application reliability what we expect from the cluster and high performance. Usually cluster what

we try to use it as a high it will provide us high performance, right. Then there is with all this paradigm, we started with the distributed computing and different outsource like grid cluster and so and so forth, we this has motivated to a system or to a realization of a things what we say utility computing, right.

What is utility computing or what is utility? Utility typically means that if I have something I need something, I should be able to get it as and when I require it, right. It is as if you go to the market and purchase the thing or get the services like I want to book a railway or flight ticket. So, I require some interface it maybe a interface calling a broker or a travel agent, right.

So, what I do? I do not care about how things happens, I want a utility to be there I want a even a plumbing job in my home call up in my residential complex, the utility office and then I say this is the thing right. I do not care or I even that the person who is taking the call may not be knowing that how to any plumbing job, but he connects or he or she connects to the particular things. But as a user I want a utility thing there. So, as if resources are there as and when we require trying to that. If you look at distributed computing strum that a cluster and other type of a computing, they also some sort of provide try to provide some utility based on the user need. And try to I should not say disconnect try to basically make the enduser least bothered about the nitty-gritty of maintaining the whole resources, right.

(Refer Slide Time: 16:45)

“Utility” Computing ?

- Utility Computing is purely a concept which cloud computing practically implements.
- Utility computing is a service provisioning model in which a service provider makes computing resources and infrastructure management available to the customer as needed, and charges them for specific usage rather than a flat rate.
- This model has the advantage of a low or no initial cost to acquire computer resources; instead, computational resources are essentially rented.
- The word *utility* is used to make an analogy to other services, such as electrical power, that seek to meet fluctuating customer needs, and charge for the resources based on usage rather than on a flat-rate basis. This approach, sometimes known as *pay-per-use*

So, utility computing is that it is a purely a concept which cloud computing practically implements what we will see later on, its a service provisioning model right model, in which a service provider makes computing resources and infrastructure management available to the customer, as needed and charges them for specific uses rather than the flat rate, right. So, the word utility is used to make an analogy to other services like electrical power, right. That seek to meet fluctuating customer needs, charge the resources based on the uses rather than the flat rate basis, and approach this approach sometimes called pay per use or what we say pay as you go model.

So, if we look at our electrical services electrical things. So, we have a meter. So, and we whenever I switch on any electrical appliances, it maybe power, it can be my electrical light or microwave oven or air conditioning or computing a computer whatever I resources the as much as I consume, that much I pay based on the rates etcetera what is decided by the electrical that electric authority or electric power authority and type of things.

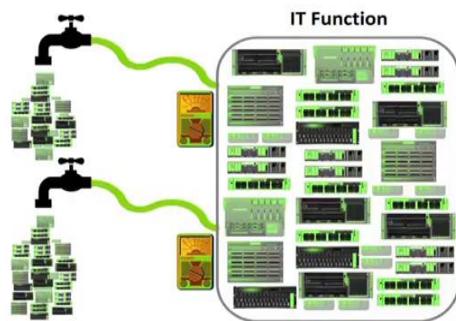
Now, I really do not care or do not know how these powers are generated, how these are coming to up to my home. Only what I care that I have metered service right whatever I use it is a service which is meter. Another popular uses is our telecom services especially the mobile services right. So, that also we use as utility and we take a connection from a again telephone or mobile service provider and then, I want, I use, I pay as I use, right, it may be postpaid prepaid and type of things, and based on that whatever the model of payment model but do not care that whether the service provider one mobile service provider 1, whom I am using another mobile service provider to connect me to somebody else, etcetera, etcetera.

What I need that facility like calling, sending messages and may be data services right data video services all the things and I am ready to pay the things right. So, based on that I select the service so, this is also utility services. So, this typically it is metered and pay as you go model or pay as you use model type of things. So, this is a paradigm shift or how we are looking at the computing though it is suddenly definitely it suddenly one day utility computing or term computing in come knocking at the door. So, it evolved from these all distributed systems to computing to cluster to type of other things, right.

(Refer Slide Time: 19:54)

“Utility” Computing ?

- "Utility computing" has usually envisioned some form of virtualization so that the amount of storage or computing power available is considerably larger than that of a single time-sharing computer.



Source: Alan McSweeney, "The Economics of Utility Computing"

33

So, that is a nice picture nice analogy, like assume that there are different it resources and I if I open up the tap it, I can have this applications resources falling from the tap when I do not require I close it. So, utility computing as a usually, what we view as form of virtualization, so that the amount of storage or computing power available is considerably larger than a single system sharing computer. So, what we are thinking it is a enormous amount of resources are there which is which I can use as I. It can be unlimited computing thing which is there, it is a unlimited networking unlimited application which is there.

Like think about if we try to make a analogy, when we using power right unless there is a restriction by the things that you cannot use more than so much kilowatt or so much megawatt that say one thing, otherwise what virtually we are thinking that there is a enormous resources available at the other end somewhere other, and I can tap anything which I want to use and I pay for the thing and I do not want to use I close the thing.

Similarly, for the computing also or even if for telecom or mobile services, what we say whenever I am downloading a particular say file, I say that the bandwidth is something which is available whatever I need, right. It may be 10 KB or 100 KB file or it may be 100 MB file, right so, but the resources are there I need to pay for it I need to pay for the whole thing, but I am able to use it. So, that there is another way of looking at it.

So, in the incase of utility computing what we are trying to see, that the computing itself starting from infrastructure computing resource may be that is hardware or the services

like may be a particular platform, we where we compile etcetera or may be that some particular data storage capability I can store any type of data or it can be some software like I want to run particular simulation or mathematical simulation tool when I say that whatever extend I want to use I can use it I need to pay away. So, it is as of a huge enormous storage of things which are there in the thing.

(Refer Slide Time: 22:10)

“Utility” Computing ?

- a) Pay-for-use Pricing **Business Model**
- b) Data Center Virtualization and **Provisioning**
- c) Solves **Resource Utilization Problem**
- d) **Outsourcing**
- e) **Web Services Delivery**
- f) **Automation**



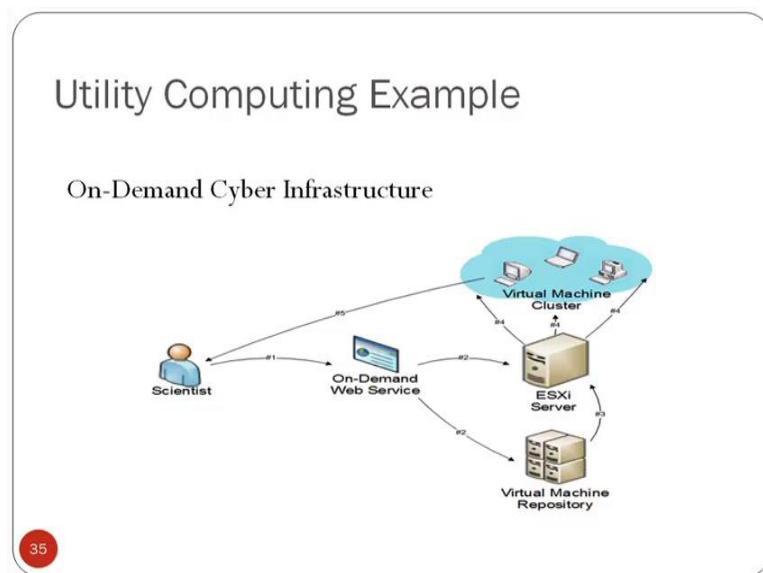
34

So, it has typical pricing model like our power grid pay for use pricing model so that is the one of the major thing there is another aspects of data center virtualization or provisioning right that will come to that; that means, I virtualize the resource at my end. Say I want to emulate a particular system which has I want say a particular processor speed so much 8 GB, 16 GB or 128 GB RAM; so much hand disk and I want to run a particular scientific simulation for my purpose, right.

So, in from the whole volume of resource I virtualize a resource for my or customize the resource, which is as if as a virtual instance. Another user from the same pool of resources may be virtualize for some other things, right. So, it is a we say that the whole thing has the data center virtualization and provisioning, solves resource utilization problem, right, I may having lot of resource at my thing and this allows me to basically supplied this resources, I can do outsource some my thing like say whenever I am doing this, I am outsourcing say software maintenance to something else or data storage thing something. So, I can do outsourcing web service level delivery, right.

So, that we will see that is what we these days what we are or rather last couple means rather last one decade mode primarily, we have shifted from data driven architecture to a service driven architecture. So, what we look at more as services than the data; right. So, that is it utilization and then I have a enormous scope of automation, right. If I have different resources etcetera then the whole I can have; I can build a work flow which allows me to automate the whole resources. So, there is another aspect of this utility computing.

(Refer Slide Time: 24:19)

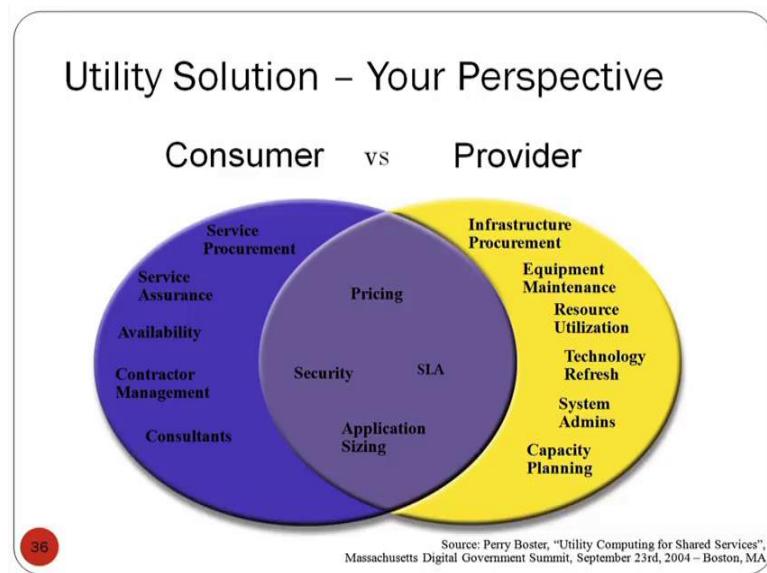


So, one example may be on demand cyber infrastructure; like I want to do a particular say examination for my for recruiting some recruiting or test of my students, and I want to have a separate infrastructure and they are they may be are in different physical location may be in different departments or if they different from their house, I want a cyber infrastructure which allows me to have a virtual resources which they can take their exam, right.

So, this I can have this type of cyber infrastructure, I can have a typical cyber infrastructure in case of a say some event management, typical event management like I want to manage a conference, and for that period I want to have service cyber infrastructure, I can have a infrastructure for disasters management which is more of a real time, that resource need maybe may vary from one to another, but I do not want to keep those resources like I say here at IIT Kharagpur we organizes conference say, it

may be coming in a one or two in a month or type of things workshop, etcetera, and I do not want to keep a separate infrastructure for the things, I want to make a infrastructure as and when I require, I may pay for the things. So, that is one demand cyber infrastructure may be the one thing.

(Refer Slide Time: 25:41)



There is another aspects of things whenever so what we see, there is a provider or service provider, there is a consumer, those two category of things has different requirement and we have different models or different what we say management issues, which ties them together. Like one maybe pricing what price you give right. So, that say I want to if you take analogy, I want to take a service provider which gives me a favorable price for my services like I want more of a data service than calling. I want data services more. So, I want to invest on those I want to select those service providers, mobile service provider who gives me a better rate or better performance, I do not care about the rate performance on this data services somebody may be more interested in the calling or messaging. So, they will do on their own.

So, that is the pricing model there, there should be a service level agreement like if I do this type of things, what should be the service level agreement what should be the availability down time etcetera there can be a security aspects right; if I say I want to use and store data, etcetera. So, somebody else would not be using or looking at my data. So, there would a security aspects even security aspects like denial of services type of things

should not be there, that also I can look at the security aspects there can be access mechanisms, etcetera.

And there may be a need for application sizing; the application need to be sized for me some application may be for doing data churning for terabytes of data, but I do want to analyze with some megabytes of data. So, I do not want that large application neither I want to purchase the thing, but I want to size the application based on my need right or in like reverse in the other way.

(Refer Slide Time: 27:37)

Utility Computing Payment Models

- Same range of charging models as other utility providers: gas, electricity, telecommunications, water, television broadcasting
 - Flat rate
 - Tiered
 - Subscription
 - Metered
 - Pay as you go
 - Standing charges
- Different pricing models for different customers based on factors such as scale, commitment and payment frequency
- But the principle of utility computing remains
- The pricing model is simply an expression by the provider of the costs of provision of the resources and a profit margin

37

So, there can be different type of payment models, right. So, it can be flat rate, tiered and type of things. Different pricing model for different customer based on factors like scale commitment payment frequency, etcetera, I can have customized things.

Principle of utility computing remains same, the pricing model is simply an expression of the providers of the cost; that means, it is how much provisioning of resources etcetera will come into the things.

(Refer Slide Time: 28:06)

Risks in a UC World

- Data Backup
- Data Security
- Partner Competency
- Defining SLA
- Getting value from charge back

38

There are several risk; we say or disadvantages based on this utility computing one, one may be the data backup. My data is somewhere, if there is a crash what will happen, right, you store the data in third party and if there is a crash or that service provider itself goes out of the business then what will happen.

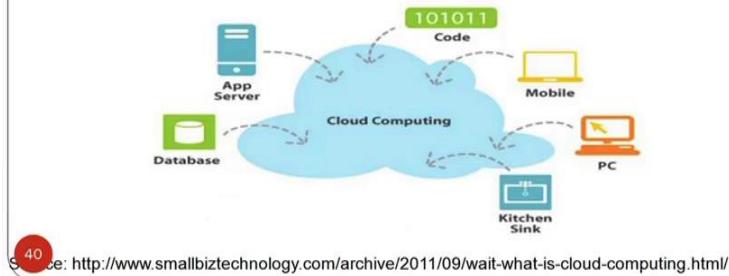
There is a data security aspect somebody else is reading or my data or not, there is a competence issue that what I am working with whether that organization is competent enough. Defining SLA is another big problem, because everybody want to define that agreements on their favor and there is still not very standard approach of doing that across the different provider and consumer, getting value from charge back that is also things which we look at that how I get values of my charge backs and these all will evolved as a cloud computing, all right.

(Refer Slide Time: 29:10)

Cloud Computing

US National Institute of Standards and Technology defines Computing as

- * Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g. networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. ”



40

Source: <http://www.smallbiztechnology.com/archive/2011/09/wait-what-is-cloud-computing.html/>

So, cloud computing is a model, for enabling ubiquitous convenient on demand network access to a sheared pool of resources or configurable resources, such as network server storage application services, that can be rapidly provision and released with minimal management effort or service provider interaction. So, it says everything; right. So, this is the typical NIST definition, which we all everybody try to what we say follow or respect. So, it says that any the huge pool of resources can be provisioned with minimal management referred, and provision and de-provision when I use it and when I have to release it both can be done in a very seamless manner.

So, this is this if you look at what we have discussed or from your background computing knowledge, this is something which is not came up suddenly on the table. What it there it evolved from some of the things; like rather we look we want to look the computing world in a different way and try to facilitate computing to different type of purposes with some basic modeling basic model. Like the pricing models, SLA model, say there is a security model management thing; that means, we want to place different things and so that a computing as service can be provided, like I want to provide computing as a service.

So, what we will do with this we will end our today's talk and we will continue with this thing with the different other aspects of the things. So, this we will do another aspect I want to mention that some of these materials, some of these figures etcetera are taken from different resources, we have I have tried to put all the references and acknowledge

those things we want to do it specifically for academic purpose only for that for our landing mechanisms no where it anything commercially using all.

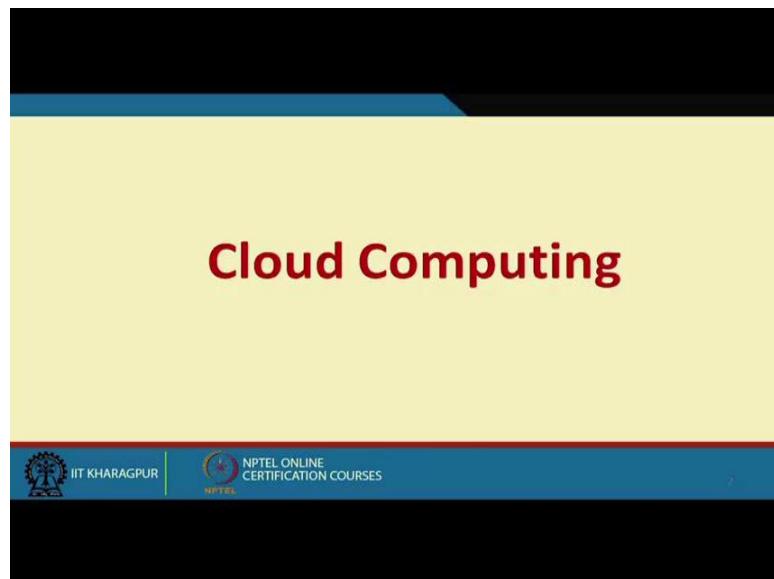
So, what we do, we with this we end our today's talk and in the coming lectures we will look at other aspects or different aspects not other aspects, different aspects of cloud computing how it goes and what are the different what are the different properties advantages disadvantages and like that.

Thank you.

Cloud Computing
Prof. Soumya Kanti Ghosh
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

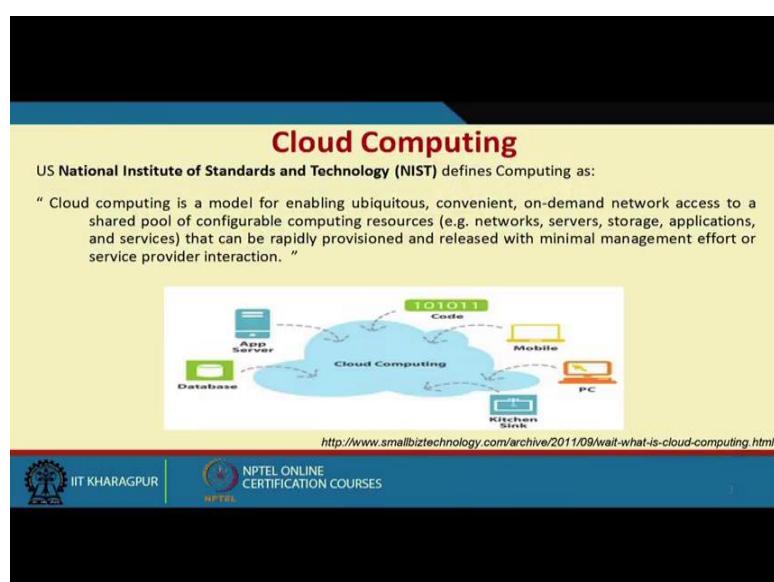
Lecture – 03
Cloud Computing – Introduction

(Refer Slide Time: 00:21)



Hi, welcome to our cloud computing course. We will continue our discussion on initially on basic or foundation of these cloud computing and then go into the more detail.

(Refer Slide Time: 00:32)



So, as we were discussing yesterday or in the last class rather, that NIST defined cloud computing how NIST define the cloud computing model, as a enabling ubiquitous convenient on demand network access to a shared pool of configurable computing resources this is point to be noted.

So that means, there is a shared pool of resource which can be configured as per demand. So, that can be rapidly provision; that means it can be provisions to realize something and released or when I do not require it can be released with minimal management effort or service provider interaction. That means, whenever a customer or the user or the service consumer needs it can provision whenever it does the; if the no requirement is there, it can deprovisioned and that is exactly the pay as you go model.

(Refer Slide Time: 01:32)

Essential Characteristics

- **On-demand self-service**
 - A consumer can unilaterally provision computing capabilities, such as server time and network storage, as needed automatically without requiring human interaction with each service provider.
- **Broad network access**
 - Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (e.g., mobile phones, tablets, laptops, and workstations).
- **Resource pooling**
 - The provider's computing resources are pooled to serve multiple consumers using a multi-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to consumer demand.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

If you look at the essential characteristics as we have already seen some of the things, they what we required; it is on-demand self-service that is a service which is on demand and it is a on-demand self-service. Broad network access, this is important thing; that means, I should have a appropriate network access because resources are distributed, they are at different geographical location, they are being pooled and provision to the user there is another thing is the resource pooling; that means, I can basically pooled resources into the for my needs and that gives us a multi tenant model with different physical and virtual resource dynamical assign like I from the same service provider, I

want to have a say, windows system windows subsystem with so many hardware specification and this are the different software specification.

From the same service providers somebody is having say a Linux say Ubuntu subsystem and one to realize those resources on the things. So, this sort of resource pooling and management of the resources is important aspects of the thing.

(Refer Slide Time: 02:46)

Cloud Characteristics

- **Measured Service**
 - Cloud systems automatically control and optimize resource use by leveraging a metering capability at some level of abstraction appropriate to the type of service (e.g., storage, processing, bandwidth, and active user accounts). Resource usage can be monitored, controlled, and reported, providing transparency for both the provider and consumer of the utilized service.
- **Rapid elasticity**
 - Capabilities can be elastically provisioned and released, in some cases automatically, to scale rapidly outward and inward commensurate with demand. To the consumer, the capabilities available for provisioning often appear to be unlimited and can be appropriated in any quantity at any time.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

There is another characteristics any type of services are measured services, right in a say; that means, I pay for what I use. So, whatever I am using is being measured. So, this is a measured services and another important aspects is rapid elasticity; that means, I can go up down as and when requirement is there, like a if I take a example like I am running a simulation, and I need a something a 4 GB sort of RAM, and when it is when we are when while running the things more data comes or more application load is there; I want to increase the RAM to 4GB to 8GB, right.

So, I need to for some portion of the time; I need to have bring more resources. So, that should I will be able to rapidly provisioning and not only that this type of elasticity is should be there; it should be I can go up come down when I requirement is there. So, that should be one of the important characteristics of property of cloud.

(Refer Slide Time: 03:50)

Common Characteristics

- Massive Scale
- Resilient Computing
- Homogeneity
- Geographic Distribution
- Virtualization
- Service Orientation
- Low Cost Software
- Advanced Security

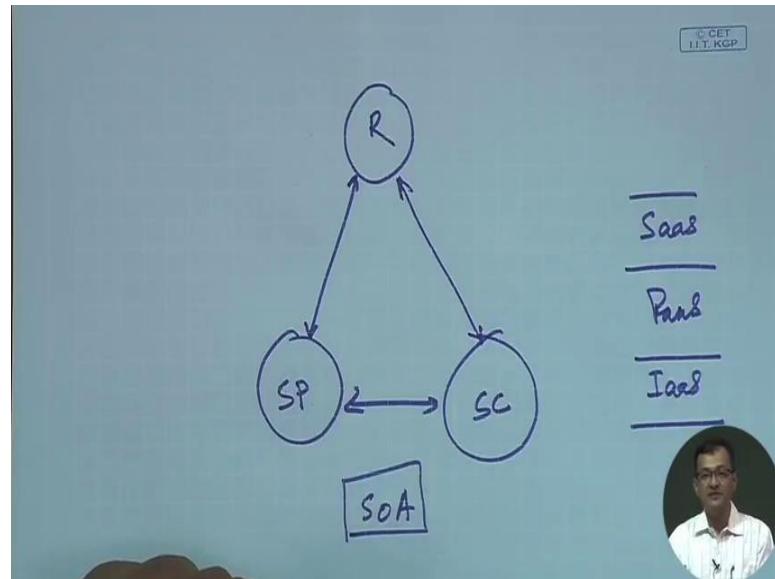
IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES



There are several common characteristics which are what we say that typically a cloud computing platform or cloud service provider. So, provide one is as we have discussed massive scale up and down, resilient computing, homogeneity like I do not care that how weather at the back in that heterogeneous resources are there I want to have a homogeneous thing, typical one characteristics is that they are usually the resources many geographically, are geographically spread or geographically geographic distribution.

One major aspect is virtualization. So, I virtualize the think. So, I virtualize a windows over a Linux say there is a that sort of thing or I have a infrastructure over there virtualize several virtual machines and work on it, I can have different applications I work on it and type of things. So, that is one important aspect rather in subsequent lectures we will look little bit more on this virtualization aspect; there is a important aspect of service orientation. So, it is the services which are being communicating with each other rather than it is a data driven. So, it is not data orientation more of a service orientation, there can be different type of services. So, if we will see that in service oriented architecture.

(Refer Slide Time: 05:18)



So, like we can have a service provider there can be a service consumer and there is a registry service or what we say somewhere some catalogue. So, the consumer talks to the or look at look up to the registry and find that what are the services are there, appropriately collect to the provider to access the service, right and service provider whenever it is having more or this is different service provider, it is a basically a bunch of represent a bunch of service provider jump bunch of consumer and it is a distributed, but some of logically centralized registry or catalogue and this service provider with basically upload update the registry in the things. It is as if I can have a analogy like this is the telephone directory type of thing, where I have different type of yellow pages and other thing and if whenever I required some to something to search starting from a particular person to particular type of activity, like I want to look for a say home delivery services.

So, I look at the things are over the home delivery services and then connect to that thing by a telephone call, right. So, this is somewhere a catalogue or registry service consumes or service provider and consumer, this over all it is a some sort of a service oriented architecture, right. So, that will again we will see little bit more that; what are the different type of service oriented architecture. So, one of the major aspects of this cloud computing is to realize service oriented architecture.

There is it may it usually have a low cost software, in the sense like at as this is being used and it is a multi tenant type of things the overall costing of the software comes down for the consumer end, right and expected to give a advance security. There are different concerned about security we will discuss some of the things, but expect to have a constant on the security. It is not only security in terms of whether somebody is bridging the security on things, some of the things like I say that the if the data is in cloud it is expected that the data is not lost right, but it is in my own system.

So, it is my responsibility to set the system keeps the system up and things. Similarly, some application I am; say a mail application on the cloud. So, it is expected that this is more stable and type of things whereas, if I am having my own mail server and mail relay etcetera then it is my maintenance and it goes on things. So, there are different aspects of the things. So, it is like it is supposed to provide a advance security features with respect to access control with respect to different resilient or blocking of different attacks, in a with respect to data preservation or no data lost type of situation and type of things, right.

(Refer Slide Time: 08:18)

Cloud Services Models

- **Software as a Service (SaaS)**
 - The capability provided to the consumer is to use the provider's applications running on a cloud infrastructure. The applications are accessible from various client devices through either a thin client interface, such as a web browser (e.g., web-based email), or a program interface.
 - The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, storage, or even individual application capabilities, with the possible exception of limited user-specific application configuration settings.
 - e.g: *Google Spread Sheet*
- **Cloud Infrastructure as a Service (IaaS)**
 - The capability provided to provision processing, storage, networks, and other fundamental computing resources
 - Consumer can deploy and run arbitrary software
 - e.g: *Amazon Web Services and Flexi scale*.

 IIT KHARAGPUR |  NPTEL ONLINE CERTIFICATION COURSES



So, if you look at the typical cloud service model, we will go into more details subsequently. So, one as we immediately look at is the software as a service, right. So, what we see the capability to provide consumer to use provider's application running on cloud infrastructure.

Like we are nowadays using different type of things different type of aspects like say word processing, a spread sheet on the internet, right even there are different if you those who are use to scientific technical writing etcetera using tools likely I take etcetera we are having things called which are on the web; so that means, those applications are somewhere yields on the cloud and we want to hook into that application and then use those application and get my data either stored in my own the local machine or at times the data is stored in some other data service provider, right.

So that means, what you require is a basic minimal way of interfacing to the external cloud right and gets that the software as a service. Now the softwares must be running as some system or must be compiling or compiled on some platform, as a user I am not bothered about it right and what I say that for using the system what I require a basic interface may be a in sort of a web browser like say Mozilla, internet explorer, chrome and etcetera, etcetera.

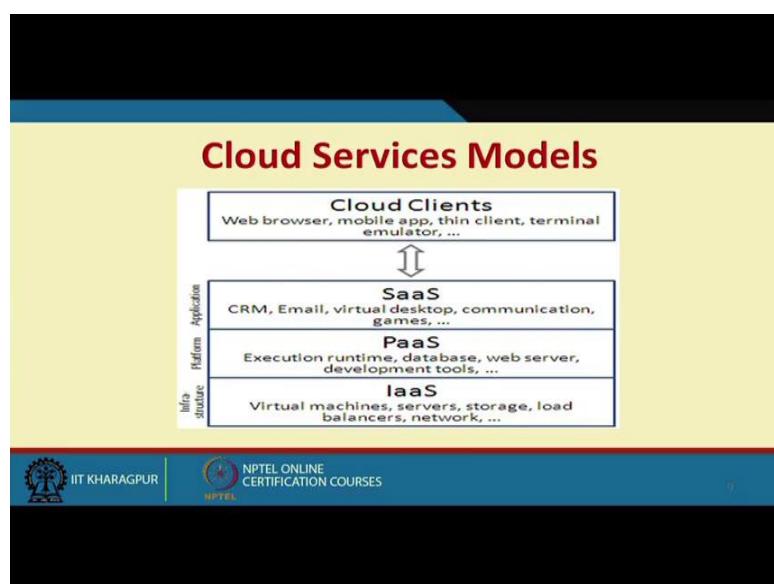
So, and I have a basic and I have the knowledge to how the use the software, it can be word processer, it can be a scientific simulator, it can be a text processing, it can be some other type of applications right; a mathematical application or type of things. So, this is as a software as a service. So, what I am getting; the service I am getting as a software that. So, that makes lot of convenience at my end I do not have to purchase, that software neither I have to maintain the updates, neither I have to look at that how much my hardware requirement to run the software in some of the software there may be hardware requirement much higher, but my application is not that large. So, that scaling of the software becomes the responsibility.

So, this is the software as a service. Another thing is that infrastructure as a service so; that means, I want to realize some infrastructure at my end, right. So, that is the infrastructure as a service. Like, I want to realize a machine of particular specification, like I say it should be a it should have some 4GB RAM so much megahertz processor and so much hard disk, right. So, this is a infrastructure, as if I am realizing the whole system on my desk, right. So, that if I again if I what I have I can have different a simple and I can basically have this infrastructure as service. Today I require a 4GB RAM, tomorrow I require a 64GB RAM for some work next day, I require a 2 GB RAM, I do not know I want to only to some basic operation.

So, usually when I have this type of very variable requirements or type of things, I can basically provision that the infrastructure. So, there are provider who provides infrastructure as a service, right. In between this 2 IAS, infrastructure as a service and software as a service, we have another stuff call platform as a service, right. So, it provides a platform for some development work, some using some test cases and other thing. So, I require one is that software I am using another things is that I have the infrastructure, I load OS I or the OS can come with the infrastructure, I load other application then run and test and other things; otherwise I require that platform as a service, right.

So, there are service provider, who provide platform as a service type of models. So, this allows me to use as a platform. So, these are three very prominent what we say service model of cloud there are other various service model rather like data as a service which keep the data. Even people are talking about whether I can have something like science as a service like I give different scientific tools etcetera and then you work on a particular avenue on the type of things and looking other things or what we say it is a something anything as a service or XaaS type of things, it typical cloud philosophy wants to make this XaaS is a realization anything as a service type of model.

(Refer Slide Time: 13:09)



So, if we look at this over all stack. So, what we have at the below is the basic infrastructure, this is and over that this is platform as a service and over there are

software as a services, right. And there can be different type of infrastructure like we can say that virtual machine many be one infrastructure; servers, storage, load balancer, network these are the different type of infrastructures, right and there can be different other type of infrastructure as you can feel so. I can have external devices and other things have a infrastructure and type of things which can be pulled in. Over that when we look at the platform as a service there is more of a something execution platform on runtime, database, web server, development tool, development platform there can be some sort of a sand box type of environment and type of things.

So, these are a more as a platform as a service, and over that I have a SaaS or the software as a service which basically realize the this softwares likes there can be CRM, email, different virtual desktop, there are different games and so on things which runs over the over and above the this cloud infrastructure and PaaS. Now how where the? This part is more of the cloud provider end. So, some provider may provide SaaS, some may provide PaaS, some may provide SaaS, IaaS or what we say infrastructure as a service; some may provide a mix of the things right, but if we look at the client perspective. So, how client access it? So, the best or the what we say the universal way is the web browser, right or what we say web client there are other things like an I can have mobile apps right which can access this type of a thing, even there are thin clients the terminal emulator and different type of other thing.

So, these are the interface by which the client interface to this cloud, right.

(Refer Slide Time: 15:14)

Types of Cloud (Deployment Models)

- **Private cloud**
The cloud infrastructure is operated solely for an organization.
e.g. Window Server 'Hyper-V'.
- **Community cloud**
The cloud infrastructure is shared by several organizations and supports a specific goal.
- **Public cloud**
The cloud infrastructure is made available to the general public
e.g Google Doc, Spreadsheet,
- **Hybrid cloud**
The cloud infrastructure is a composition of two or more clouds (private, community, or public)
e.g Cloud Bursting for load balancing between clouds.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Now if we with this thing if we want to look at the different type of deployment models right. So, what are the different type of deployment model? How to realize those cloud or type of clouds right. One is the private cloud, I build a cloud and I use it like in our IIT Kharagpur with we indigenously with open source systems, we build a cloud called Meghamala. So, it is name of the cloud is Meghamala, it is a very not so large cloud, but is privately build on over source system primarily used by researchers, both faculty and research scholars, for mostly a infrastructure as a service over the things.

Over that we try to give other services like data services or other type of big data services and type of things, but primarily it is a infrastructure as a service. So, it is a private cloud it is accessible within the IIT, Kharagpur network itself and not accessible outside the network right and it has a different charging model like as it is a institute funded infrastructure. So, we do not charge anything on things. There can be others other end a what we say public cloud right. So that means, which is public. So, at different type of things, I can have public clouds as infrastructure, I can have public clouds as like amazon clouds are for infrastructure for a one of the popular example; there can be platform as a service type of cloud like what we look at some sort of a what we say that Microsoft azure or we can have things like software as a service type of cloud like Google doc tools spreadsheets type of things some of the examples.

So, these are the public clouds, right. There are another things what I can have a hybrid cloud right. So, it is a mix of things of public and private as such. There is a typical category of cloud which is a, what we say community cloud. So, it is something semi semi-public semi private type of things like I can have it is shared by several organization and support specific goals or what we say coat-un-coat it say some sort of a likeminded or like I means same type of goal achieving organization.

Like I say bank sector may have a banking cloud, right. So, that core banking things whatever they exchange with the things in our country like RBI, may be the nodal agency to handle or the things over that there is a cloud. I can have a institutional cloud like what we have say government funded institutional cloud, which can work on the things. I can have there are things like if I working with some other type of data like bank etcetera like one like what we work on in the spatial data we can have a Geospatial cloud which with the type of things.

So, this is addressed to a community, may not be a not in a one organization multiple organization, but it has some purpose to do right. So, this is a community cloud. Now I can have a hybrid or the things some private, some public, some community or public private only and type of things, right. So, these are the different type of what we say deployment models of the cloud.

(Refer Slide Time: 18:37)

Cloud and Virtualization

- **Virtual Workspaces:**
 - An abstraction of an execution environment that can be made dynamically available to authorized clients by using well-defined protocols,
 - Resource quota (e.g. CPU, memory share),
 - Software configuration (e.g. OS).
- **Implement on Virtual Machines (VMs):**
 - Abstraction of a physical host machine,
 - Hypervisor intercepts and emulates instructions from VMs, and allows management of VMs,
 - VMWare, Xen, KVM etc.
- **Provide infrastructure API:**
 - Plug-ins to hardware/support structures

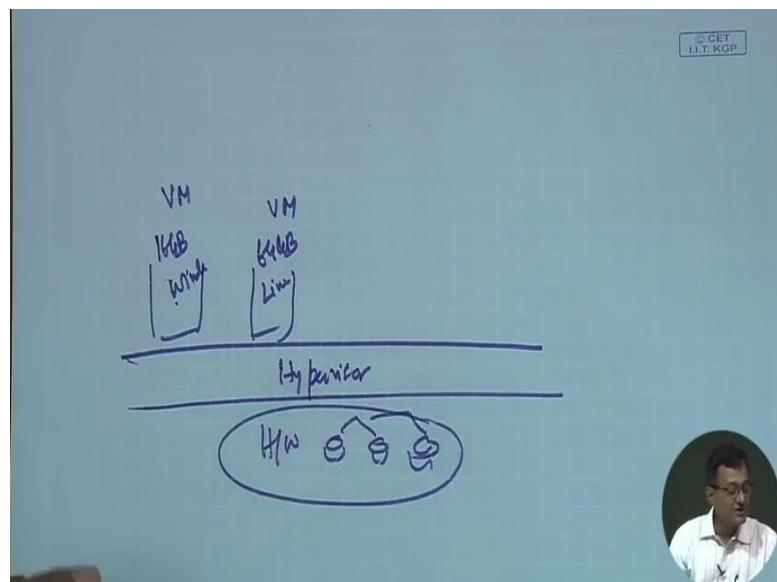
Virtualized Stack

```
graph TD; App[App] --- OS[OS]; OS --- Hypervisor[Hypervisor]; Hypervisor --- Hardware[Hardware]
```

Now, though we will go little more differ into the things is subsequent lectures, but we can do see that the cloud and this virtualization somewhat come hand in hand right. So, if we look at this picture which we say there virtualized or virtualization stack. So, we have a bare metal of the hardware at the end, right, over that I require a something a middleware which allows me to create different virtual things.

So, what we are trying to see like what we are saying that we are having different Hardware, right.

(Refer Slide Time: 19:16)



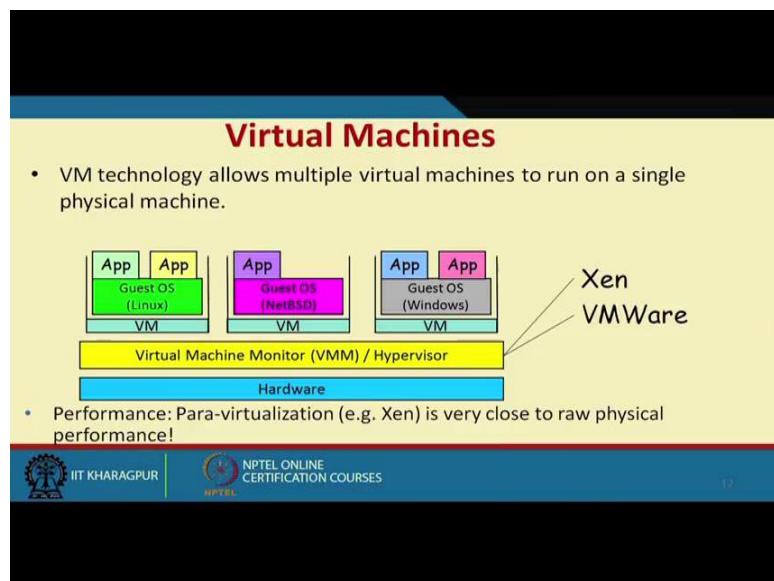
It can be different systems across different network etcetera and connected through different sort of network etcetera over that what I am end goal is to realize different virtual machine like I can say this is a 16GB machine of running windows. This is a maybe 64GB machine RAM with running something some orient of Linux etcetera right. Now, this collectively being realized to make this type of different virtual machines right I require something in between what we say VMM or virtual machine monitor or hypervisor.

So, this hypervisor basically allows me to emulate this different type of machines. So, if we look at the virtualization stack. So, one or virtualization how I want to what we want to virtualize one is that virtual works spaces I can have different virtual works spaces and abstraction of an execution environment that can be made dynamically available to the authorized clients by using well defined protocol. So, I suppose for my particular project

I require an environment for this may be development environment visualization environment, computing environment and then I generate dynamically or the organization generate dynamically to the for the team to work on it, right. And that may be a short term project of 14 days or something and it basically realized. Purchasing a whole stuff for that 14 days maybe a costly affair, but if I provision that from external world on a pay you or from public cloud and maybe a solution, right.

So, there is only another thing aspect is the virtual machine on to realize into the into different virtual machines; and then there are there can be different category of virtual machines which can be used by the user like as a talking about Meghamala, we realize different category of virtual machines with basically three category of virtual machines and whenever use a needs come we provision it based on the availability of the things. And it can provide infrastructure API, plugins to hardware support infrastructure etcetera. So, it can provide it provides API to connect to the different infrastructure. So, we will basically look at visualization there are different aspect of virtualization. So that we will look at in the subsequent lectures.

(Refer Slide Time: 21:50)



So, if we just look at this; whatever you are talking about. So, virtual machine technology allows multiple virtual machine to run a single physical machine.

So, this is my hardware. So, it is not that single physical machine means not only one machine there can be couple of machine clubbed together and realized, like I say that I

have ten 4GB systems or so that I can have some effectively 40GB things, and then I realize say three or two 16GB and one 8GB actually you cannot do all division there are some requirement to for this virtualization to happen. So, it 40GB can you break into two 16GB and so, I realize two 16 GB machine using underlining this hardware, right. So, what we require is one is the virtual machine monitor, which allows me to emulate this hardware and I can have different virtual machines over the things and there are a concept of guest OS which resize on the things, right.

So, there the guest OS here may be Linux, there can be other guest OS, there can be other guest OS and different type of things and this applications running on the things are different; and this virtual machine by may be used by somebody at some geographical location, this virtual machine somebody at the some other geographical location and so on and so for. So, everybody for individual consumer it is say machine for it is purpose. So, there are different category of virtualization, we will see the things like a performance for virtualization close to raw physical performance we will see that why and how and later on. So, some of the popular virtual machine monitor or hypervisor is one is Xen another is VM ware.

(Refer Slide Time: 23:40)

Virtualization in General

- ***Advantages of virtual machines:***
 - Run operating systems where the physical hardware is unavailable,
 - Easier to create new machines, backup machines, etc.,
 - Software testing using “clean” installs of operating systems and software,
 - Emulate more machines than are physically available,
 - Timeshare lightly loaded systems on one host,
 - Debug problems (suspend and resume the problem machine),
 - Easy migration of virtual machines (shutdown needed or not).
 - Run legacy systems

So, virtualization in general there are advantages of the virtual machines, like run operating system where the physical hardware is unavailable, easier to create new

machine backup machine etcetera, software testing using clean install etcetera, there are of number of things which are which can be realized with this virtual machine.

(Refer Slide Time: 24:04)

Cloud-Sourcing

- **Why is it becoming important ?**
 - Using high-scale/low-cost providers,
 - Any time/place access via web browser,
 - Rapid scalability; incremental cost and load sharing,
 - Can forget need to focus on local IT.
- **Concerns:**
 - Performance, reliability, and SLAs,
 - Control of data, and service parameters,
 - Application features and choices,
 - Interaction between Cloud providers,
 - No standard API – mix of SOAP and REST!
 - Privacy, security, compliance, trust...

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES
NPTEL



So, there are issues of cloud sourcing when we are, why it is becoming important? First of all high scale low cost providers. So, I can have lot of resources as a thing. So, economy party play a important role right when you get the much cheaper services on the things, etcetera, we will see in a separate lecture that whether this economy is always what this cloud economics of cloud says and whether always it is a beneficial or we need to think about there when the things are beneficial etcetera.

Anytime anyplace access via web browser. So, it is my whole computing my hold services whatever I am exercising is anytime anyplace type of services right. So, I do not bother about that where the services are leverage and type of things, and there are rapid scalability incremental cost and load sharing. So, there is scalability is pretty high right I can rapidly scalable. That is cannot forget need to focus on local ITs. So, that the whatever your infrastructure is there, that need to be looked into; there are some of the serious concerns also whenever you are leveraging or resources from a things, one major concern is performance reliability and service level agreements right what will be the performance what will be the reliability the service is always on and type of things and there are issues of service reliable agreements like I signs some agreement with the things whether these those are honored, not honored and type of things.

There are issues of concern of data and service parameters like I say concern of the data my data is in cloud, even I write a later the latter is in the cloud, I write a report there is in the cloud then I am concerned that whether this data is secured, should not be tampered and so far, and another issue is there service parameter this is also there like how do they say that the availability is 90 percent or 95 percent. How to look at those parameter; how to audit that that whatever I looked for is given by the things, right.

There are issues or application features and choices. So, it is sometimes we say that it is provides a application which is fixall, right, whether I do not have a customized choice of the thing whether there a issues there whether I can have or not there issues of interaction between cloud providers, right. So, there are one provider purchasing services from the thing, like it may be a infrastructure provider purchasing some of the infrastructure like hardware etcetera from a data provider and so on and so for.

No standard or like issue of standardization of the API is like whether it is a soap or rest type of services, when we look at the service oriented architecture and of course, issues of security privacy, security compliance, trust, competence, risk these are the major concern whenever we purchase this sort of a service. Especially it is critical when we purchase those things for our some of the misson critical or my day today operational requirement.

(Refer Slide Time: 27:25)

Cloud Storage

- Several large Web companies are now exploiting the fact that they have data storage capacity that can be hired out to others.
 - Allows data stored remotely to be temporarily cached on desktop computers, mobile phones or other Internet-linked devices.
- Amazon's Elastic Compute Cloud (EC2) and Simple Storage Solution (S3) are well known examples

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Cloud storage why we keep it separate because it is place a pretty important role. We will if time permits we will look at that some of the storage aspects, but it say one of the important factor because data is extremely important for the consumer or for the service consumer or the users, right.

So, allow data stored remotely to be temporarily cached on the desktop computers, mobile phones or other internet linked device. So, you can have a sinking with the data also many of us are using this sort of services, for putting our data as a service. So, there are things like amazon EC2, S3 are well known examples where this data services are there.

(Refer Slide Time: 28:16)

Advantages of Cloud Computing

- **Lower computer costs:**
 - No need of a high-powered and high-priced computer to run cloud computing's web-based applications.
 - Since applications run in the cloud, not on the desktop PC, your desktop PC does not need the processing power or hard disk space demanded by traditional desktop software.
 - When you are using web-based applications, your PC can be less expensive, with a smaller hard disk, less memory, more efficient processor...
 - In fact, your PC in this scenario does not even need a CD or DVD drive, as no software programs have to be loaded and no document files need to be saved.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Some of the advantages just to re-iterate one is the lower computer cost, right one major aspects there is low I should say not computer cost lower computing cost.

Improved performance I expect that the overall performance will be improved, because I am purchasing something as a higher price; reduced software of course, like I do not have to purchase separate software etcetera. Instance software updates if there is a update this is getting updated; improve document format compatibility, if there are interoperability you should be a prop better you addressed.

Unlimited theoretically unlimited storage, I pay and I get the storage, increase data reliability, I it is stored in distributed things with proper redundancy, etcetera. So, it is

expected the data it will be high, universe and information access, I can access information at every point, latest version of the systems and software available; easier group collaboration when you are having one data different applications; device independent I can access from different type of devices.

(Refer Slide Time: 29:29)

Disadvantages of Cloud Computing

- **Requires a constant internet connection**
 - Cloud computing is impossible if you cannot connect to the Internet.
 - Since you use the Internet to connect to both your applications and documents, if you do not have an Internet connection you cannot access anything, even your own documents.
 - A dead Internet connection means no work and in areas where Internet connections are few or inherently unreliable, this could be a deal-breaker.
- **Does not work well with low-speed connections**
 - Similarly, a low-speed Internet connection, such as that found with dial-up services, makes cloud computing painful at best and often impossible.
 - Web-based applications require a lot of bandwidth to download, as do large documents.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | NPTEL

Required say there are; obviously, some these are all good points there are always some pit faults or some concerns or disadvantages, like requires a constant internet connection you are disconnected your gone.

Does not work well with low speed connection that is again network connectivity. Feature might be limited right. So, some of the things I may not get what I get on a customized personal thing at can be slow at time because this is provisioning etcetera finally, takes time. Stored data I am really we are nearly concerned about whether how secured it is; stored data can be lost, if there is a crash on the things or the provider goes out of the things that there may be chance of loss. Cloud is not truly a high performance computing systems. So, if we think that cloud will provide high performance with may not provide there HPC type of things, and there are several general concept or connectivity of a API's having particular database connectivity where the database will run, etcetera.

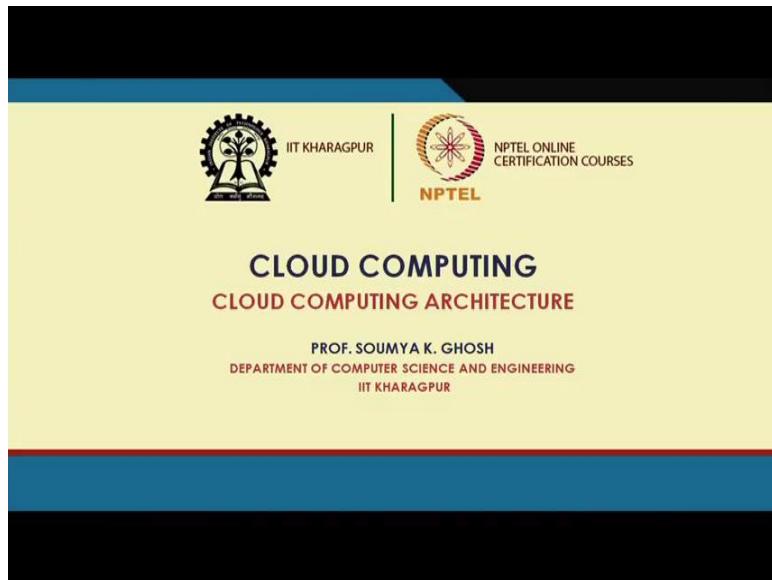
So, what will do we will stop here and for now and then we will look at the different architecture, and in my in our sub subsequent lectures.

Thank you.

Cloud Computing
Prof. Soumya Kanti Ghosh
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

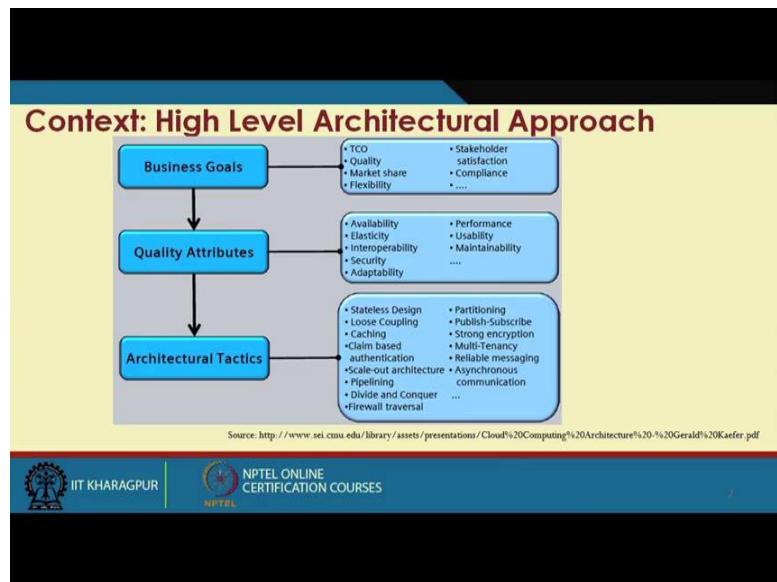
Lecture – 04
Cloud Computing Architecture

(Refer Slide Time: 00:27)



Welcome to the course on Cloud Computing. We will continue our discussion on different aspects of the cloud computing. Today, we will look at or we will start with Cloud Computing architecture, right. What are we will start with the basic introduction and we will go to more details in today's or subsequent lectures.

(Refer Slide Time: 00:43)



So, if we look at these high level architectural approach or high level view of this things, so what we have at the end of the things what we are looking for one is that from the user prospective or overall cloud computing, it is we want to achieve some business goal. It can be business goal with respect to consumers, it can be business goal with respect to the producer or it may be the overall framework of the cloud computing right. So, some of the business goals is the major thing. Next, we have the quality attributes and then the architectural tactics or the basic architectural framework.

So, if you look at that business goal at the top of the thing, what we looking for primarily, we are looking for some TCO like what is the total cost to the organization, who are the stake what about the stake holder satisfaction, compliance with different standards market shares, flexibility, etcetera. So, this is the overall business goal from the CSP point of view, and to achieve this business goal we have few more parameters we need to be looked into or which are which need to be monitored measured. And what we mean to say that need to be managed properly is that issue of availability, elasticity, interoperability, security, adoptability, performance, usability and maintainability.

Now, if you look at this if you just go back in one or two lectures, where we started with the basic definition of cloud computing. So, we were looking for these aspects of the things like that whether how much available, how it is elastic, how it is interoperability. So, these are major what we say major characteristics of the cloud. So, this

characteristics basically allows a service provider to define his business goal across these different cloud market.

And finally, in order to achieve; this different measured services or the business goals, we need to have basic architectural tactics or architectural view. So, like there were the considered essentials are whether it is a stateless design, whether we loosely coupled right what we are thinking that there are several devices which are geographical spread. So, they may be loosely coupled, heterogeneous, need to be what we want to bring into the thing is a some sort of a interconnect and homogeneity between the things over broad network access right. So, what is the caching mechanism. So, claim what should be the authentication, whether it is claim based authentication then there are several other aspects of the architectural techniques.

Now, if you look at this sort of things somewhere other are manifested at any type of service model, like if I say infrastructural service model, so some aspects will come into play, if I say PaaS some other aspects will come into play, and we have SaaS or any other type of model. So, this becomes important keep in mind while developing, proposing any architectural view. So, with this respect or with keeping this in view, we see that; what are the basic consideration to have in this cloud architecture.

(Refer Slide Time: 04:25)

Major building blocks of Cloud Computing Architecture

- **Technical Architecture:**
 - Structuring according to XaaS stack
 - Adopting cloud computing paradigms
 - Structuring cloud services and cloud components
 - Showing relationships and external endpoints
 - Middleware and communication
 - Management and security
- **Deployment Operation Architecture:**
 - Geo-location check (Legal issues, export control)
 - Operation and Monitoring

Ref: <http://www.iit.ac.in/library/assets/presentations/Cloud%20Computing%20Architecture%20-%20Gerald%20Koefler.pdf>

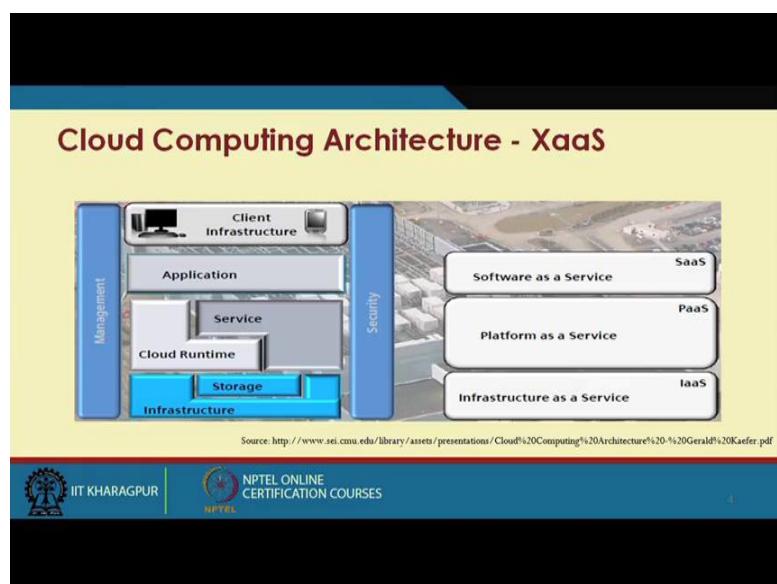
IIT KHARAGPUR | **NPTEL ONLINE CERTIFICATION COURSES**

So, if you look at the major building block of cloud computing architecture, one is that technical architecture, structuring according to the XaaS stack. Adopting cloud

computing paradigms, structuring cloud services and components, middleware and communication and management and security. So, these are more technically architect technical feature of the architectural things.

There are some of the things, which are deployment operation architecture like geo-location check, right. I may want to say that due to my federal requirement all my data, all my applications for any type of government related activities should be within the territory of the country. I do not host or any application or any data outside the things. Like, if I am having a mail service, the mail data, mail server etcetera should reside within the geographically bounded area of our nation, so that might be a thing. There may be other legal issues etcetera. And there are deployments issues like operation and monitoring. So, what should be the operational view, how it will be monitored, how this XaaS stacks looked in to those are the other deployment operation architecture.

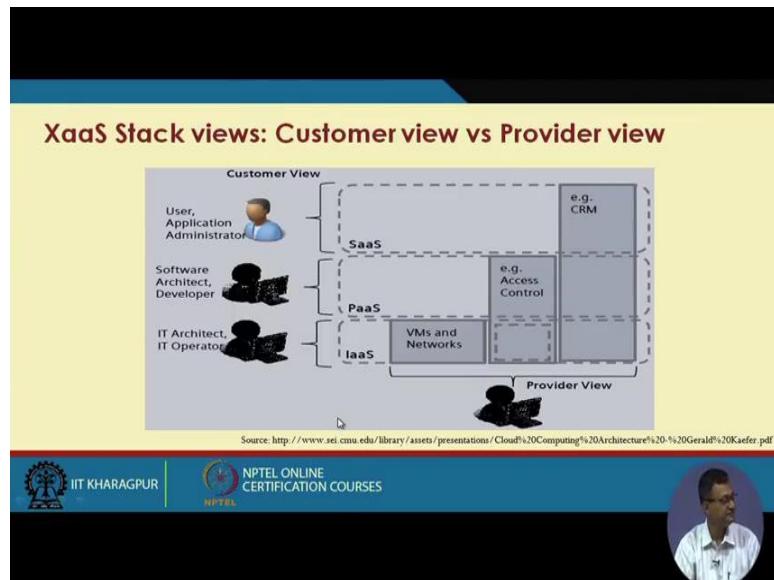
(Refer Slide Time: 05:52)



So, if you look at this typical XaaS thing, so what have at the down is this infrastructure as a service, then platform as a service, software as a service this is the typical stack of the or typical or most popular XaaS applications or services. Now, if you look at this architectural side of the things or the basic building block, so down the line is more of a storage and infrastructure. So, this is that is the colour in the blue; up in the thing is the application this right. And there are the client infrastructure which interacts with this application; in between is the middleware or what we say that the platform or the

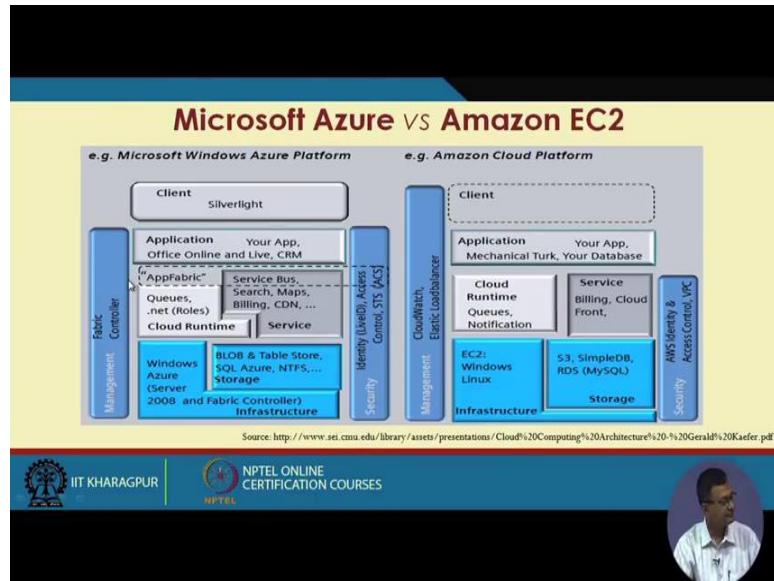
platform as the service type of things where the services cloud runtime libraries and other things come out. So, any realization of any cloud computing infrastructure whether it is a public, private, open sourced, customized, we need to look in to some where other is broadly has to have these type realization.

(Refer Slide Time: 06:58)



And if we look as view of consumer versus providers or the customer versus provider, so there are again for the IaaS this is primarily important for IT architect in the IT operation type of things. For the end user or the application user that is more important is the user application, administrator etcetera which overs around the top of the things like one of the popular application may be the CRM. And whereas for the software architect developer which comes in the middle which uses this hardware infrastructure or which is uses this infrastructure as a service, develop something and which are deployed or put support to these application above this. So, these are this categories like in the middle of the things. So, different people or different category of users are different view of this architecture.

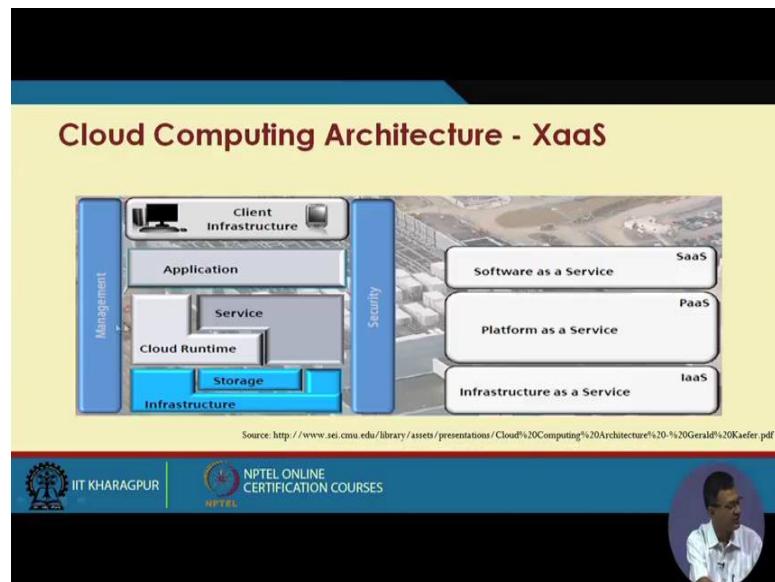
(Refer Slide Time: 07:59)



Now, we just try to map that with this our very two popular one of means two popular cloud services, there are several other popular services, but these are the things which are available which will put to get. So, one is the Microsoft windows azure platform, another is the Amazon cloud platform or Amazon EC2. Now, look at that also divided into more or less in that three stack. This is as more at the bottom of the things same as the incase of EC2; in case of Amazon it is windows servers and related stuff; where is in case of EC2 it is it can be windows Linux infrastructure along with Amazon S3 and simple DV and other type of storage.

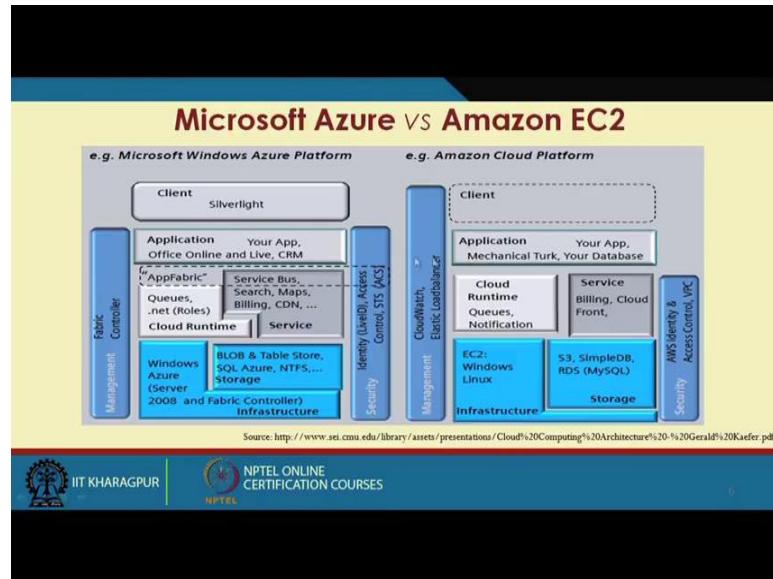
And in between we are having this middleware which emulates the platform this is has a analog of stuff there. And at the top it has the applications. These are some of the windows azure applications and these are the Amazon applications. And here the client can connect through either in case of a client this is silver light in case of this or we can connect it through the browser.

(Refer Slide Time: 09:29)



So, one thing I just look like to mention it was there in the previous slide also, there are two vertical stack right what we say management and security right. This is important because this management and security issues sometimes people who want to put the security is also a management aspect and sometimes we have other aspects like what we say quality of services. So, management quality of services and security in some of the references or material you will find that. So, all these are the scenarios where these are the stack which goes vertical that means, I cannot achieve security or management taking only one layer at a thing. So, it requires some sort of a cross layer considerations that is why these are vertical stacks.

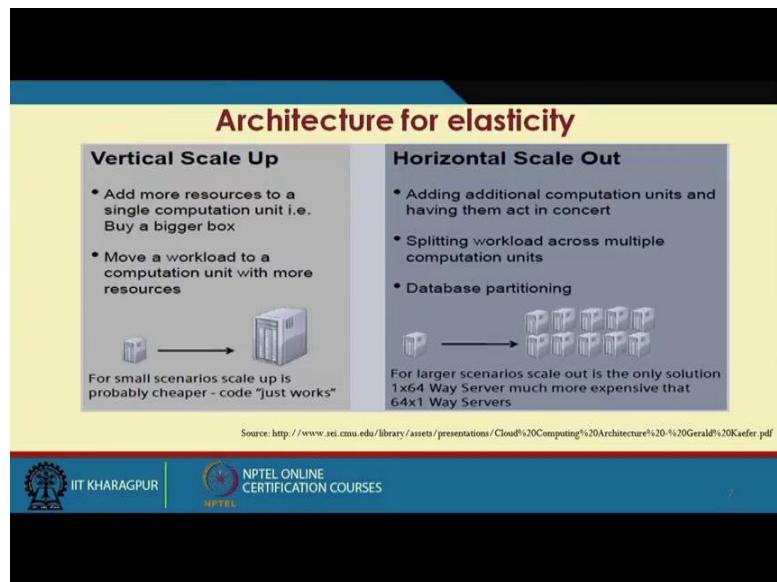
(Refer Slide Time: 10:27)



And if you see the same thing in case of a case of these azure or EC2, so these are also we have these vertical stacks. So, they have different component of security relevant component of management, but these goes on the vertical stack. Similarly, the security aspect is based on the we will see later on based on how much control we are having goes on to things. Like if I am having infrastructure as a service, so my security or management is more consideration on the as a infrastructure. So, once I rollout this infrastructure, one I provision this infrastructure virtual machine etcetera, so that is the responsibility of the user or the consumer to look into the thing.

If I having a up to a software as a service then all these issues are need to be provided up to that layer right. So, word processing service or CRM service which is a software service. So, all aspects up to that level of management and security it will provided, so that it depends that what sort of service provisioning or what sort of service deployment we are doing for this particular scenario.

(Refer Slide Time: 11:40)



There are several aspects or issues as we are talking about; one is elasticity. As we are mentioned or as we are know that elasticity is one of the most important characteristics of cloud like if I scale up, scale down into this computing paradigm. Now, in order to achieve that. So, there can be two broad approaches right. So, this is this two broad approaches which is two when without cloud also it is more of a elastic computing when do. One is that you can do it vertical scale up; that means, at more resources to the box right, we have a particular infrastructure say for example, I have a particular rack with different servers.

So, we initially we are having say I am having some 24 servers, then I scale up putting more server into the things or my chassis may be holding 16 blades and having initially 4 based on the requirement, the organization feels that the requirement go on another 4 or up to a 16 blade chassis like fully loaded chassis. Now this is vertically going up. So, this is has advantage where for in case of a small scenario, scale up is probably cheaper, so that it works just you be put on the things. So, add more resources to a single computational box, move the workload to the computation unit to more resources. So, I have more resources.

Whereas, when we have scenarios where for larger scenario scale out scale out is only solution 1 is to 64 way server is much more expensive than 64 into 1 way server, right.

So, having say 64 of 1 unit computational power and trying to achieve realize a 64 computational units is much cheaper than having a one unit of 64 computation.

So, what is happening that when I have a large scale scenario then expanding in that fashion may be pretty costly. So, in that cases not only expanding in terms of cost in terms of only financial aspects, it is maintainability rate of failure may be pretty high like a one if the single system, it is a single point of failure at time. Whereas if I have a multiple systems, so that if an if some systems fails, then I can have a way to work on in a little lower performance metric, but I can work on the things. So, this becomes whether it is a horizontal scale up or scale out or it is a vertical scale out, we need to consider based on our requirement.

(Refer Slide Time: 14:42)

Service Models (XaaS)

- Combination of Service-Oriented Infrastructure (SOI) and cloud computing realizes to XaaS.
- X as a Service (XaaS) is a generalization for cloud-related services
- XaaS stands for "anything as a service" or "everything as a service"
- XaaS refers to an increasing number of services that are delivered over the Internet rather than provided locally or on-site
- XaaS is the essence of cloud computing.

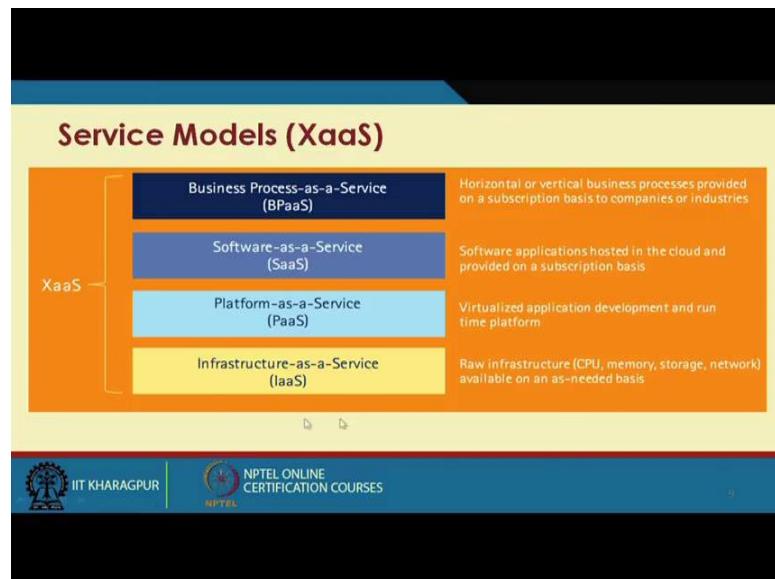
IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Now, if we again come back to our service model, so XaaS, what is happening is the combination of service-oriented infrastructure or a service-oriented architecture on for realization of infrastructure and cloud computing realize to a XaaS. So, what is the service oriented architecture, I believe most of you are acquire already knowing. So, what will do, we will have a; some sort of slides on service-oriented architecture in subsequent lectures. So, that those who are not accustomed or not very much familiar with this type of things, we will have immediately quick check up.

But nevertheless it is a service driven approach for infrastructure or PaaS or any type of things right. So, we are tried to combine SOI and cloud computing to realize this XaaS.

So, anything as a service is a generalization of cloud related services, refers to increasing number of services that are delivered over internet, rather than provided locally or on site. XaaS is the essence of cloud computing already we know. So anything as a service, so still we seen data is a service another popular we can have anything other type of services.

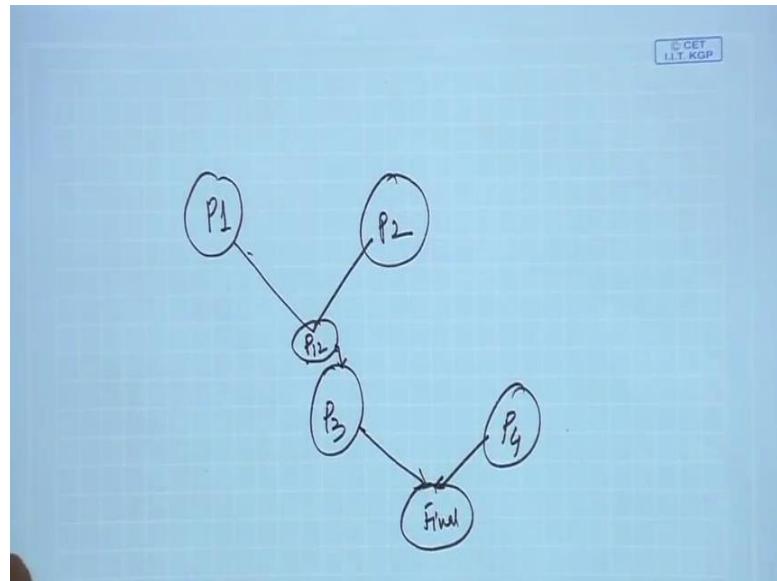
(Refer Slide Time: 16:01)



So, as some of these popular XaaS instance, one is infrastructure as a service, platform software service already known. There is another thing came up what is now becoming pretty popular what we say business process as a service. So, I have a business work flow which want to give as a service. So, horizontal or vertical business processes provided on a subscription basis to companies or industries. Like suppose for a particular operation, I have a business process. So, what are what does it mean, it may not be only one applications which is delivering the things.

So, I have different processes which has a orchestration between them, then allow me to realize a particular applications. It can be for something for the banking sector, it can be for some of the aspects of different development planning operations of the things. So, it is not only I run a application and get results. A application is dependent on other applications another etcetera.

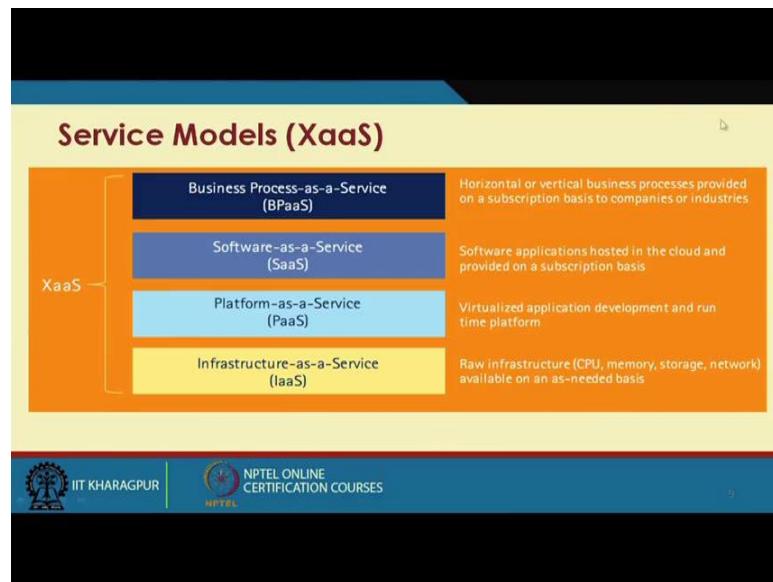
(Refer Slide Time: 17:16)



So, I have different processes or applications which allows me to realize the thing. So, it means as if P12 may lead to something called P3 which can be an input to the P3 and in turn P3 and P4 will come out with the final outcome of the things. Now, I have different may be different processes right. So, these are orchestration in the things. So, I can say I have business process which is basically division of the; or amalgamation or integration of different sub processes or sub applications. The interesting feature is that there is a timing relationship and what we say processes execution tree or process execution graph will be there, so that it follows a particular things.

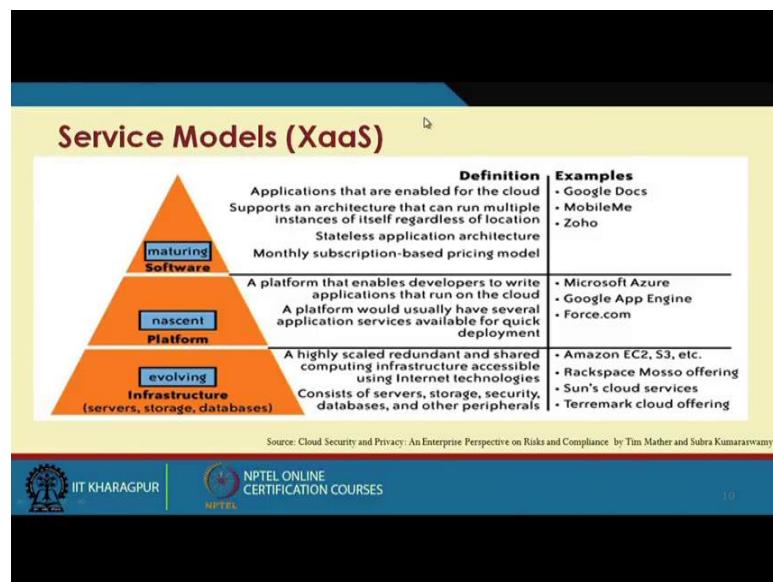
Now, for this sort of things, if it is defined, then whether I can say this I give a service with a business process as a thing. So, it is not now not only the application as a service or software as a service, so I basically have a something more than software as a service which talks with things; incidentally this different processes may come from different heterogeneous providers. So, I can basically now interact with or collaborate with different cloud service providers and then realize the thing. So, this business process is important because end of the day specially for the organization, this matters, rather many of you might have heard there is a very popular language what we say business process execution language where you can define your business process of outflow into the thing and so it realizes the thing. So, that is one of the major aspects.

(Refer Slide Time: 19:14)



So, IaaS is the RAW infrastructure CPU etcetera. PaaS is the virtualized application development or run time platform. SaaS is the software application hosted in the cloud provided in a subscription basis. And this business process as a service is a horizontal or vertical business process involved in basis on a subscription basis to companies or industries.

(Refer Slide Time: 19:40)



And if we again little relook, so it is something a tapering thing our popular IaaS, PaaS and thing, PaaS other infrastructure.

(Refer Slide Time: 19:52)

The slide has a yellow header and footer. The title 'Service Models (XaaS)' is in red. The content is organized into two main bullet points:

- **Most common examples of XaaS are**
 - Software as a Service (SaaS)
 - Platform as a Service (PaaS)
 - Infrastructure as a Service (IaaS)
- **Other examples of XaaS include**
 - Business Process as a Service (BPaaS)
 - Storage as a service (another SaaS)
 - Security as a service (SECaaS)
 - Database as a service (DaaS)
 - Monitoring/management as a service (MaaS)
 - Communications, content and computing as a service (CaaS)
 - Identity as a service (IDaaS)
 - Backup as a service (BaaS)
 - Desktop as a service (DaaS)

At the bottom, there are logos for IIT Kharagpur and NPTEL, along with the text 'NPTEL ONLINE CERTIFICATION COURSES'.

So, other than this three most popular examples, we have several other business process, several other XaaS instances, like we have already seen business process as a service, there is storage as a service, I can have security as a service like it provides different security future as a thing, I can have database as a service. So, I do not want to install etcetera, I want to leverage database into a service. Monitoring management as a service like I want to manage my IT infrastructure, like I want to I have a large say in our IIT infrastructure we have large labs which are used by different category different type of PG, UG courses and I want to basically manage the things, right.

Like manage the things up to a starting from software of the means what is the software going on, whether they are machine is basically having any problem from the hardware or type of things different things. So, whether I can have I have separate tools or somebody gives me a service which connects with this machines and do the things right or any large infrastructure in case of a organization. So, communication content computing as a service these are the some of the things, identity as a service, backup as a service, desktop as a service. So, I can have anything as a service and these are some of the things which are being used across the world.

(Refer Slide Time: 21:19)

Requirements of CSP (Cloud Service Provider)

- Increase productivity
- Increase end user satisfaction
- Increase innovation
- Increase agility

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES



So, end of the day what does CSP or the cloud provider is looking for increase productivity, end user satisfaction, some innovative services, so that it has it can hold its market position. And have different entry its infrastructure as a aisle. So, it is as aisle and it can give, it cores it can what we say configure itself along with the need of article.

(Refer Slide Time: 21:52)

Service Models (XaaS)

- Broad network access (cloud) + resource pooling (cloud) + business-driven infrastructure on-demand (SOI) + service-orientation (SOI) = **XaaS**
- XaaS fulfills all the 4 demands!



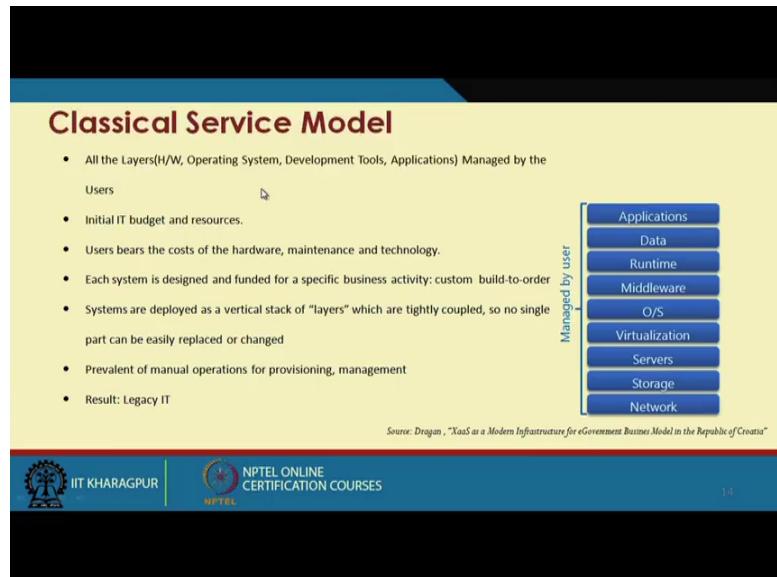
Source: Understanding the Cloud Computing Stack: PaaS, SaaS, IaaS © Diversity Limited, 2011

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, if we look at this again, if we come back to our basic definition and try to see, so one is the broad network access, resource pooling, business driven infrastructure on demand like what is there and service oriented orientation, service oriented architecture. So, all

this four feels us in realizing this XaaS type of services. So, XaaS fulfills all the four typical demands of service provider.

(Refer Slide Time: 22:27)



Now, like our network stack like our TCP/IP or OSI models, so if I want to have that what will be the typical classical stack of the things, stack of our service model or XaaS model. So, at the base of the things what so it is not like that, so it is a what we say, logical way of looking at the things, right, it is not like that something will done always over the other, but it is a logical representation of the things. Like at the core of the thing is the networking. So, you have a underlining broad network access which allows the things to talk each other right. Another core component is the storage it is also something omnipresent. Other aspect is the server.

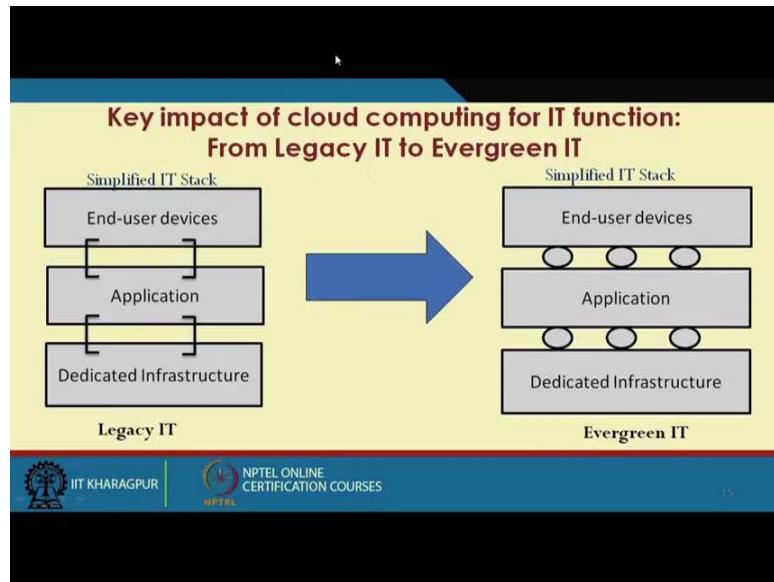
So, if you look at this three stuff are primarily is the major building block, what we say the bear metal things over which it works. So, one is the communication, one is this different servers and other physical infrastructure, and other one is the storage. Storage always we try to keep it as a separate aspect other than keeping one of the infrastructure, in some of the cases storage are kept as a infrastructure, but storage plays a different type of role storage management is different thing, so that is that those are a important aspects of the storage is the input aspect.

Over that what in order to realize are the things is that is aspect of virtualization. So, I need to virtualize all the things. So, I can have a virtualization of the whole infrastructure

I can have the virtualization only of the particular behavior virtual machine, I can even virtualizes the network. So, I have a infrastructure virtualization, I want to realize a particular virtualize network infrastructure, and of course, storage. So, there is a virtualization basically try to have these so called virtual machines or virtual network, virtual storage type of stuff. Over that if it is a virtual machine I need to put the particular OS over which I should run the thing right.

So, this over that this middleware then I have runtime library, the data over for which more closely with the application and finally, end of the top of the stack is the applications. So, these are the different components which make this IaaS that the feasible, it may be noted that all the components are not equally important may not be equally important for all type of XaaS type of thing, but nevertheless they play a important role in realization of the thing.

(Refer Slide Time: 25:11)



Now, if we try to because always it hover that all we are some somewhere other we are already getting this stuffs right, what is this big paradise shift right. So, as such we should we have already mentioned in our initial lectures that already we are having distributed systems and other things into place. We do realize lot of things etcetera. So, to make it more market viable and having a scalable computing infrastructure we as you go model measures services where real easily. So, if you look at that a simplified IT stack, so we have end user devices application and dedicated infrastructure.

So, this is in the legacy IT or still it is very much still it is very much present in or rather omnipresent in various organizations and other type of thing, so these are somewhat clipped together right or more strongly coupled right. So, that dedicated application runs over the applications dedicating infrastructure. So, application sizeing etcetera are made based on the infrastructure or other way around.

End user also have a some dedicated application to work with whereas, what we go for what we are trying to realize is more of a simplified IT infrastructure, these are little bit flexibilities are maintained. So, I have some dedicated infrastructure over application end user devices, but they are not very strongly bound. So, this is the thing what we are going to do and it can be shown also we can since see some of the subsequent lectures that this sort of things may allow us to have better utilization of the resources or better return on investment right. Or having a giving service different type of services on the same type of or a common resource based type of thing.

In some of the cases several organizations has several surplus resources, like if I talk again about our own UG or PG labs which may be having huge amount of computing resources, but then they may not be utilize across 24 hours right they are utilized may be 8 to 10 hours on a very loaded. But out of out of class hours or out of lab hours the those can be clubbed together to have to give the researchers the opportunity to run high simulation on the basic PC type of platform, so that may be one way of utilizing own resources in a better fashion.

(Refer Slide Time: 28:00)

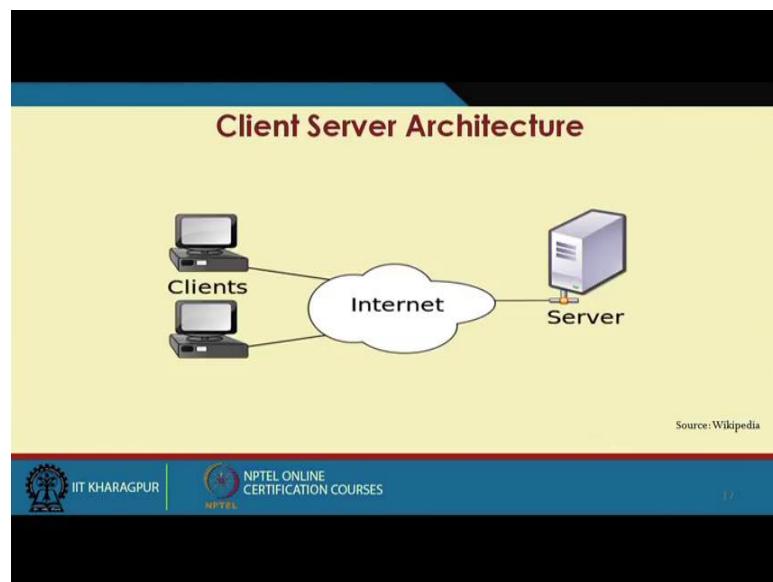
Classic Model vs. XaaS		
	Business Model	Definition/Example
Traditional	1 Licensed Software	Traditional Software Licenses (w/ upgrade + maintenance) Examples: Oracle; SAP; Microsoft
	2 Hardware Product	Hardware Product sale (e.g. PC, Server, Router) plus maintenance / support services Examples: Cisco, Dell, HP
	3 People-based Services	Professional Services Examples: IBM Global Services, Accenture, Wipro
New/Emerging	4 SaaS	Software functionality delivered as utility services Examples: Salesforce.com; Taleo; Workday; NetSuite
	5 IaaS	Storage-on-demand, compute capacity Examples: eVault; Amazon EC2; Dropbox
	6 PaaS	Provide entire web services dev. environment/ platform Examples: Force.com; Azure; Amazon Web Services

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

16

So, there are some of the other things like classical model plus XaaS model, there are different aspects like licensed software, hardware product and people based services these are the more popular aspects of our traditional cloud, traditional infrastructure models or classical infrastructure model. And we need to make the architecture in such a fashion that it maximizes this effort. Whereas, in is our imagine things we are going to a XaaS type of services like, so that is software IaaS, PaaS or any type of any type of XaaS type of services.

(Refer Slide Time: 28:46)



17

And rather if you look at our classical thing or still it is very much valid we have a client server architecture. So, there is a server which serves several clients; and based on the server programme is running on the server, and always waiting for the client to connect and we connect to the thing. And this is extremely popular and still very much in the use and will be in the use in several cases right. But from there we are migrated to a service oriented architecture, so instead of this very strongly coupled client server, we have service oriented to it, service talks to each other. And there are distinct advantages we can have heterogeneous things to talk to each other and we have lot of flexibilities. Very popular client server things are like FTP server, telnet, HTTP are a way particular demon of the server part which looks into this talks to this client.

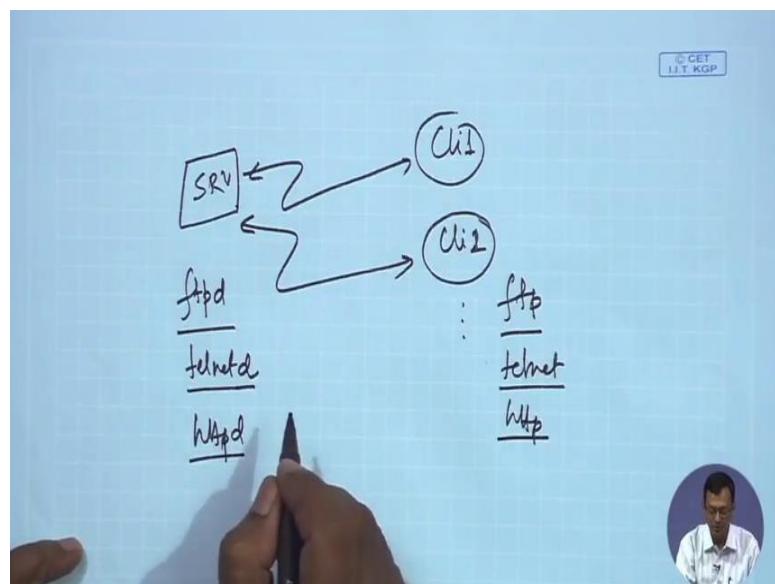
So, we will carry on our discussion in our subsequent lectures on this particular architectural aspects of a cloud computing and we will also see that different things of virtualization another thing. So, for today we will for now for this lecture we will stop here.

Thank you.

Cloud Computing
Prof. Soumya Kanti Ghosh
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Lecture – 05
Cloud Computing Architecture (Contd.)

(Refer Slide Time: 00:49)

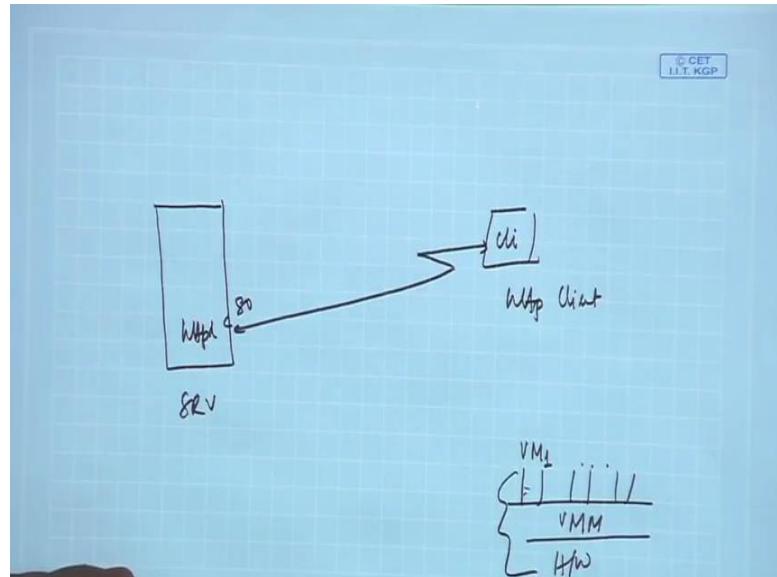


Hello, so we will continue our discussion with cloud computing architecture. So, in the last lecture, we are discussing about this client server - traditional client server paradigm, and what this cloud is offering from the service orientation point of view; we will continue that discussion with the things. So, what we have in case of a client server model and we have a server and we have several clients which are connecting to this server, right. So, we have several clients which are connecting to this particular server. So, this server and client logically or physically can be on the same machine, the same machine can be client server, different machine can be one client one server. A client can access server for some other things, etcetera.

So, what we do we have for every server things, we have a client like if I have a ftp client server. So, what we have ftpd daemon in a particular Linux flavor and there is a ftp client. So, ftp client look for a ftp server. So, in this case, what to try to realize a server basically the server process waits on a particular port. So, what we does like say if I have

ftp or telnet similarly telnet d, if the telnet daemon server then telnet or very popular thing httpd, then I have http server. So, these all these things are a way particular port.

(Refer Slide Time: 02:10)



That means, a in the in the server machine say for say httpd. So, a particular in the server thing. So, in the port 80 that logical port 80, the http daemon is listening. So, from somewhere from a client or http client that is any of the browser, it basically same say request to the things if the server is speed connects to the things and the connects in the rest of this, right. So, server does a always listening to a particular port is any client is there. Whenever the client needs a service, it basically hook into that port and it goes on things.

And we have seen that 2 type of that can be a concurrent server will give service is concurrently or there can be a iterative server one after another, but all this things I require a one component of this whole process like for http, httpd and http etcetera to talk to each other. So, somewhat it is more strongly bound, but nevertheless it serves many of our purposes and it is serving and we will be serving and it diffract to our standard in any type of sort of application where data and application communicates with each other.

(Refer Slide Time: 03:25)

Client Server Architecture

- Consists of one or more load balanced servers servicing requests sent by the clients
- Clients and servers exchange message in request-response fashion
- Client is often a thin client or a machine with low computational capabilities
- Server could be a load balanced cluster or a stand alone machine.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Now, client server architecture consists of one or more load balancing server send by the things. Client and servers exchange messages by request-response. Client is often may be a thin or a machine with low computational capability, not necessarily, usually. Server can be load balanced cluster or a standalone machine, etcetera. Like if you think about out http server using the server is much stronger and fatter systems whereas your client on anything from your mobile device to laptops to or some any type of devices which can be there, but nevertheless it can be some higher end devices also. So, it basically tries to emulate a three type of architecture, one is a presentation layer at the top level, then logical layer and the data layer at the bottom level. So, this; try to realize a three tier architecture.

(Refer Slide Time: 04:22)

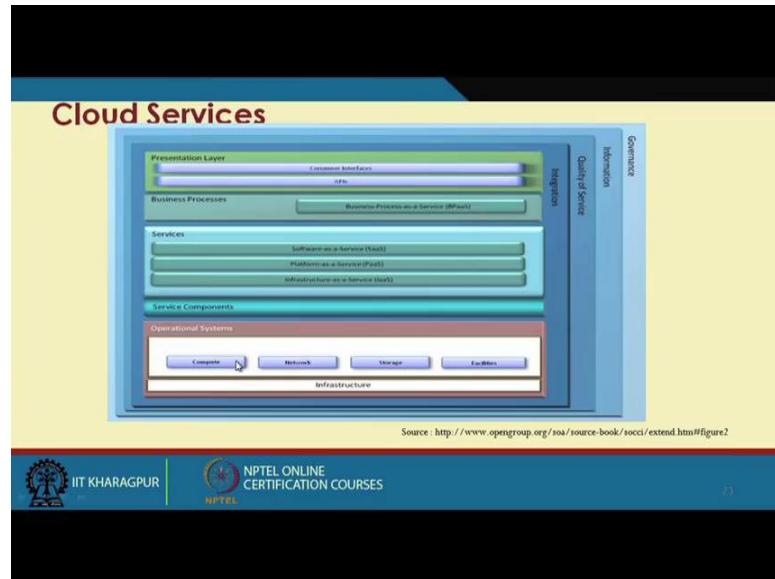
The slide compares the Client Server model and the Cloud computing model. It features two columns: 'Client server model' and 'Cloud computing model'. The 'Client server model' column lists four points: 'Simple service model where server services client requests', 'May/may not be load balanced', 'Scalable to some extent in a cluster environment.', and 'No concept of virtualization'. The 'Cloud computing model' column lists five points: 'Variety of complex service models, such as, IaaS, PaaS, SaaS can be provided', 'Load balanced', 'Theoretically infinitely scalable', and 'Virtualization is the core concept'. The slide is branded with IIT Kharagpur and NPTEL logos at the bottom.

Client server model	Cloud computing model
<ul style="list-style-type: none">Simple service model where server services client requestsMay/may not be load balancedScalable to some extent in a cluster environment.No concept of virtualization	<ul style="list-style-type: none">Variety of complex service models, such as, IaaS, PaaS, SaaS can be providedLoad balancedTheoretically infinitely scalableVirtualization is the core concept

And if we and this is a very popular things what we are which we are use to and if we try to look at that little bit comparison with the things what is the need of cloud computing or where it is different. So, in case of a client server, simple service models server service clients request, in case of a clouds variety of complex service models like IaaS, PaaS, SaaS can be provided. So, it is a so every service in case of a client server, I should have a sever program where the client connect, where in case of a cloud we have different type of XaaS type of service models. So, the way of looking at the service model it says different may or may not be load balanced in case of a client server. Whereas usually these clouds are load balanced that is one of the major aspects what we want to do.

Scalable to some extent in a cluster environment, so if the things are running over cluster this is scalable and in case of a cloud it is theoretically infinite scalable. So, anything can be there. The most important thing is there is no concept of virtualization right, in case of a client server. And the virtualization is our main concept or the core concept in cloud computing. So, these are some of the things and what we try to show or propose that this cloud computing is able to do much more justification or much more better utilization of the whole resources and able to cater a large number of stakeholders.

(Refer Slide Time: 05:57)



The picture is not pretty clear but so just divide the things one is that at the bottom is the infrastructure of the operational, operational systems like you see that compute network storage facilities and middle is the different type of services. So, I have infrastructure thing. So, we want to have little bit broader view of this of this cloud computing cloud services. So, these are our core bare metal. So, layer is realization of the services right. I have IaaS, PaaS or SaaS type of services. These services you use to have some business processes to run. So, it is not the services not only the services, I want something business my business was need to be realize. So, I work use the services to run different business processes. So, this is business processes.

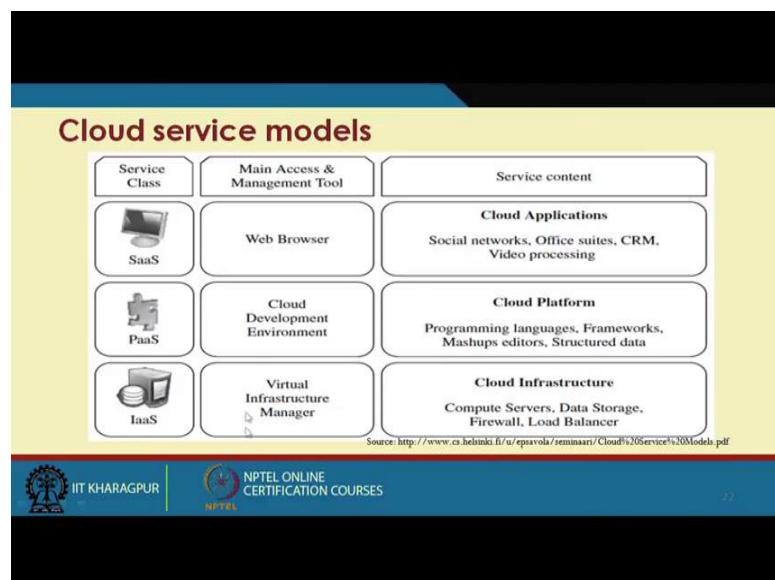
And I can have business process as a service to the thing. There is a presentation layer what we can have directly some interfaces with the users or I can have some API which allows me to connect to the connect to these different service model, like we are use to different type of APIs for different service provider is that I use that APIs to connect to the things. So, I do not I expose myself as a service provider this giving this API. So, or I can have customized user end some applications or what we say some particular interfaces to connect to the things.

Now whole thing should have a integration platform right, it is integration not that very strongly integration platform that is integration of the hardware it require some sort of a particular architectural point of view where I wants to have a service integration then I

should have a orchestration engine some sort of a execution three type of thing. So, service execution thing, so integration.

Then I whenever I do that then we require that some what is the quality of servicing doing all those things whether I can operate at the appropriate level of service what we want to confer. There is a dissemination of information or type of things right what sort of what services are provided how to hook in to the things that is information type of things. And finally, I should have a overall governance; this governance may be management of the infrastructure or the management of the whole system the governance we have also have legal issues and policies implemented, so that the legally valid as per the federal laws and regulations. So, this makes the governance.

(Refer Slide Time: 08:46)



And we basically realize those things we having different services; I am not repeating this because we have seen this in several in other form in several sides. So, finally, we are using this different service model.

(Refer Slide Time: 09:04)

Simplified description of cloud service models

- **SaaS** applications are designed for end users and are delivered over the web
- **PaaS** is the set of tools and services designed to make coding and deploying applications quickly and efficiently
- **IaaS** is the hardware and software that powers it all – servers, storage, network, operating systems

Source: http://broadcast.rackspace.com/hosting_knowledge/whitepapers/Understanding-the-Cloud-Computing-Stack.pdf

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, we will just see a little bit again relook into this service model little bit in a little more detail SaaS, PaaS and IaaS.

(Refer Slide Time: 09:17)

Transportation Analogy

- By itself, infrastructure isn't useful – it just sits there waiting for someone to make it productive in solving a particular problem. Imagine the Interstate transportation system in the U.S. Even with all these roads built, they wouldn't be useful without cars and trucks to transport people and goods. In this analogy, the roads are the infrastructure and the cars and trucks are the platform that sits on top of the infrastructure and transports the people and goods. These goods and people might be considered the software and information in the technical realm

Source: http://broadcast.rackspace.com/hosting_knowledge/whitepapers/Understanding-the-Cloud-Computing-Stack.pdf

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Like one may be the sometimes; they have a we used to have a studies in analogy with our transportation like roads are the infrastructure, cars and trucks are the platforms which carries the thing, and goods and people might be considered at the software or information in a technical terms or technical paradigm. So, as if I have this infrastructure

over that I have different platform and this different applications are running over the things if we look at the point of our cloud computing paradigm.

(Refer Slide Time: 09:48)

Software as a Service

- SaaS is defined as software that is deployed over the internet. With SaaS, a provider licenses an application to customers either as a service on demand, through a subscription, in a “pay-as-you-go” model, or (increasingly) at no charge when there is opportunity to generate revenue from streams other than the user, such as from advertisement or user list sales.

Ref: http://broadcast.rackspace.com/hosting_knowledge/whitepapers/Understanding-the-Cloud-Computing-Stack.pdf

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, software as a service is this definition, I am not repeating this, it is a software that deployed over the internet, use as a pay as you go model, and you can basically search and leverage that service for things.

(Refer Slide Time: 10:08)

SaaS characteristics

- Web access to commercial software
- Software is managed from central location
- Software is delivered in a ‘one to many’ model
- Users not required to handle software upgrades and patches
- Application Programming Interfaces (API) allow for integration between different pieces of software.

Source: http://broadcast.rackspace.com/hosting_knowledge/whitepapers/Understanding-the-Cloud-Computing-Stack.pdf

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Characteristics; web access to commercial software that is one of the typical features; software is managed from some sort of a logically centralized location, so it may be

distributed and deployment distributed fashion, but the management is logically some sort of a centralize location, right I say that I give a some sort of software as a service for computational support for different simulation, so mathematical simulation, etcetera. So, this over all simulation platform may be running on heterogeneous different platforms may be on a distributed systems, but the overall management is logically centralize type of things.

Software is delivered in a one-to-many model. So, it is to one-to-many. Users not required to handle upgrades etcetera there is one of the features. Application programming interface allow for integration between different pieces of software or APIs allows us to amalgamate or integrate different software to realize a larger thing. Where it is extremely useful or may be useful is application where there is a significant interplay between organization and outside world, like email, letter campaign software like. So, there is a lot of interface between the things.

(Refer Slide Time: 11:20)

Applications where SaaS may be useful

- Applications where there is significant interplay between organization and outside world. e.g. email newsletter campaign software
- Applications that have need for web or mobile access. E.g. mobile sales management software
- Software that is only to be used for a short term needs.
- Software where demand spikes significantly. E.g. Tax/Billing software.
- E.g. of SaaS: Sales Force Customer Relationship Management (CRM) software

Source: http://broadcast.rackspace.com/betting_knowledge/whitepapers/Understanding-the-Cloud-Computing-Stack.pdf

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Application that have a need for web and mobile access; software that is only to be use for a short term needs I do not require something always, but I only want to use for a short term needs. Like, I want to execute I want to run a particular examination or particular written test for interview and I want to run that thing for a particular thing, there is for not that day to day I am running the things. So, neither I want to deploy those everything on the thing. So, I have a somewhat somewhere I can get the services and I

realize the services over the say desktop on a particular lab and execute that examination and then forget that thing for the time being.

So, software where demand spikes significantly like tax, billing, etcetera, right. So, the demands are have spikes right, like usually the tax submission or the tax fillings has a particular time window during the years and that time the demands increases otherwise slows down and that is important. Then there are these are some of the popular or will use thing that sales force, CRM and type of things.

(Refer Slide Time: 12:36)

The slide has a dark blue header and footer. The main content area has a light yellow background. The title 'Applications where SaaS may not be the best option' is in bold red font at the top left. Below it is a bulleted list of four items. At the bottom left is the IIT Kharagpur logo, and at the bottom center is the NPTEL logo. The footer contains the URL 'Source: http://broadcast.rackspace.com/hosting_knowledge/whitepapers/Understanding-the-Cloud-Computing-Stack.pdf' and the number '78'.

Applications where SaaS may not be the best option

- Applications where extremely fast processing of real time data is needed
- Applications where legislation or other regulation does not permit data being hosted externally
- Applications where an existing on-premise solution fulfills all of the organization's needs

Source: http://broadcast.rackspace.com/hosting_knowledge/whitepapers/Understanding-the-Cloud-Computing-Stack.pdf

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

78

Applications, where SaaS may not be the best option right, the other side of the story application where extremely fast processing or real time data is used. So, where because what we have seen in that cloud computing may not be or cannot be considered as a high performance real type of operation, there can be delay as a times it may be slow down and type of things. So, it applications where extremely fast operations for real time data is needed is may not be the things may be say for example, disasters management etcetera. So, I cannot leverage something and which takes time and type of things. So, I too have something which is more concrete even I am provisioning it should be dedicated with a particular QoS maintenance.

Application; where legislation and other regulation does not permit data being hosted externally. Like, I can say that the my institutional work or in our institute say with that organization can say that all for a organization, the organization related communication

over the mail should be on the organization mail server or using the organizationally email id it cannot be on a public type of things. So, because that is the legal provision in the data where I do not know like to see. Application where an existing on-premise solution fulfills all the organizations needs it may be there that already you have a legacy applications which is fulfilling your organization need in may not be want to shift.

(Refer Slide Time: 14:04)

Platform as a Service

- Platform as a Service (PaaS) brings the benefits that SaaS bought for applications, but over to the software development world. PaaS can be defined as a computing platform that allows the creation of web applications quickly and easily and without the complexity of buying and maintaining the software and infrastructure underneath it.
- PaaS is analogous to SaaS except that, rather than being software delivered over the web, it is a platform for the creation of software, delivered over the web.

Ref: http://broadcast.rackspace.com/betting_knowledge/whitepapers/Understanding-the-Cloud-Computing-Stack.pdf

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Similarly, for platform as a service it brings benefits of SaaS bought over the application, but over a software development platform. So, it provides the things. It is bring software delivered over the web. It is a platform for creation of the software deliver of the things. So, it gives a platform to develop your own system, etcetera.

(Refer Slide Time: 14:26)

Characteristics of PaaS

- Services to develop, test, deploy, host and maintain applications in the same integrated development environment.
- Web based user interface creation tools help to create, modify, test and deploy different UI scenarios.
- Multi-tenant architecture where multiple concurrent users utilize the same development application.
- Built in scalability of deployed software including load balancing and failover.
- Integration with web services and databases via common standards.
- Support for development team collaboration – some PaaS solutions include project planning and communication tools.
- Tools to handle billing and subscription management

Source: http://broadcast.rackspace.com/hosting_knowledge/whitepapers/Understanding-the-Cloud-Computing-Stack.pdf

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, there are several characteristics like services to develop, deploy, host, maintain in the same integrated development environment. Web based user interface to create tools to create, modify, test deploy different user interface and other things. Multi-tenant architecture; where multiple concurrent user utilize the same deployment application. Built in scalability is there integration with web services support for development team collaboration, that is important. Then I have a different a development team which is spread over several regions, may be several countries and want to have a common development platform and want to work on it that, it gives a good solution. And tools to handle billing subscription management etcetera are other aspects of the things, that how do I manage the whole.

(Refer Slide Time: 15:15)

Scenarios where PaaS is useful

- PaaS is especially useful in any situation where multiple developers will be working on a development project or where other external parties need to interact with the development process
- PaaS is useful where developers wish to automate testing and deployment services.
- Popularity of agile software development, a group of software development methodologies based on iterative and incremental development, will also increase the uptake of PaaS as it eases the difficulties around rapid development and iteration of software.
- PaaS Examples: Microsoft Azure, Google App Engine

Source: http://broadcast.rockspace.com/hosting_knowledge/whitepapers/Understanding-the-Cloud-Computing-Stack.pdf

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, where is extremely useful or very useful is that especially useful in the situation where multiple developers will be working on development project that we were discussing. Where other external parties needs interact with the development processes etcetera. So, I have a development teams which is the core in my in our organization office. I have development team at the customer premises which interact with the customer day-to-day basis and there may be other subsequent development team in head office and type of things. So, when we want to coordinate between these two, this platform may be plat form is the service may be a ideal base for do it this.

PaaS is useful where developers wish to automate testing and development. So, if you want to automate then this are the things where you put something, which is being checked. Popularity for agile software development a group of software development methodology based on iterative and incremental development, in these cases also what we find that PaaS will be extremely useful. So, some of the popular PaaS and type of things are one is Azure, and Google app engine or other several stuff where people are using for developing their applications then making it is web enabled and available over webs.

(Refer Slide Time: 16:28)

Scenarios where PaaS is not ideal

- Where the application needs to be highly portable in terms of where it is hosted.
- Where proprietary languages or approaches would impact on the development process
- Where a proprietary language would hinder later moves to another provider – concerns are raised about vendor lock in
- Where application performance requires customization of the underlying hardware and software

Source: http://broadcast.rackspace.com/hosting_knowledge/whitepapers/Understanding-the-Cloud-Computing-Stack.pdf

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Scenarios where PaaS may not be ideal. So, it is may not be the good choice where application needs to be highly portable in terms of where it is hosted. So, if high portability is required then it may not be good choice. Where proprietary language or approaches would impact on the development process, so there may be proprietary languages approaches, especially in mission critical stuff, like in case of a defense in case of some of the cases of financial sector. You may not want that organization may not want to do in this all open system open I do not say that it is a open source only, but it is not on a public systems rather than on proprietary things and there are can be some proprietary approaches of looking at the problems. So, they are this type things may not be useful.

Where proprietary language would hinder later moves to another provider concerns raised it settle vendor locking. So, there is another issue of vendor lock in come into the thing. So, it is a vendor locking, then also we have a problem here. Where application performance require customize of underlying hardware and software. So, my objective is to have a application performing at its base then I need to customize that the underlying hardware and software that means, I want to have a customize stack, right. I do not have generic stack of the thing, but I have a customize stack then this type of things may not be PaaS may not be a good choice.

(Refer Slide Time: 18:02)

Infrastructure as a Service

- Infrastructure as a Service (IaaS) is a way of delivering Cloud Computing infrastructure – servers, storage, network and operating systems – as an on-demand service.
- Rather than purchasing servers, software, datacenter space or network equipment, clients instead buy those resources as a fully outsourced service on demand.

Source: http://broadcast.rackspace.com/hosting_knowledge/whitepapers/Understanding-the-Cloud-Computing-Stack.pdf

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

And finally, that IaaS, which provides a infrastructure in terms of servers, storage, network operating system on an on-demand bases right. So, in rather than purchasing or maintaining we can have a virtualization of those things like or I can have a virtual machines and virtual storage and type of thing. So, that is the IaaS, extremely popular and used.

(Refer Slide Time: 18:29)

Characteristics of IaaS

- Resources are distributed as a service
- Allows for dynamic scaling
- Has a variable cost, utility pricing model
- Generally includes multiple users on a single piece of hardware

Source: http://broadcast.rackspace.com/hosting_knowledge/whitepapers/Understanding-the-Cloud-Computing-Stack.pdf

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Characteristics are well-known that resources are distributed as services, allows dynamic scaling. Has variable cost like this I can have I can pay as you use utility pricing model,

includes multiple users on a piece of hardware. So, I can have number of user working on a same hardware stack or realizing different virtual machine on the same hardware stack.

(Refer Slide Time: 19:00)

The slide has a dark blue header and a light yellow body. At the top, it says 'Scenarios where IaaS makes sense'. Below that is a bulleted list of six points. At the bottom, there's a reference link and logos for IIT Kharagpur and NPTEL.

Scenarios where IaaS makes sense

- Where demand is very volatile – any time there are significant spikes and troughs in terms of demand on the infrastructure
- For new organizations without the capital to invest in hardware
- Where the organization is growing rapidly and scaling hardware would be problematic
- Where there is pressure on the organization to limit capital expenditure and to move to operating expenditure
- For specific line of business, trial or temporary infrastructural needs

Ref: http://broadcast.rockspace.com/hosting_knowledge/whitepaper/Understanding-the-Cloud-Computing-Stack.pdf

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Where it makes lot of sense, the demand is very volatile, right. I have some times a machine of a particular memory requirement sometime some other requirements. So, demands are very that means, there are spikes are there or new organizations without the capability of investment. In several organizations or several projects what we see that we immediately purchasing something on the hardware is may not be feasible or sometimes what we want to do some of the PoCs - proof of concept that time also these are not purchasing takes a lot of time so or I can have a kick start of the whole process. So, they are the IaaS is helpful where there is a pressure organization to limit capital expenditure may one of the things. For specific line of business trial or temporary infrastructural needs, so that means what we have say that proof of concept, etcetera.

(Refer Slide Time: 19:57)

The slide has a dark blue header and footer. The main content area is yellow. At the top, it says 'Scenarios where IaaS may not be the best option'. Below that is a bulleted list. At the bottom, it says 'Source: http://broadcast.rackspace.com/hosting_knowledge/whitepapers/Understanding-the-Cloud-Computing-Stack.pdf'.

- Where regulatory compliance makes the offshoring or outsourcing of data storage and processing difficult
- Where the highest levels of performance are required, and on-premise or dedicated hosted infrastructure has the capacity to meet the organization's needs

Source: http://broadcast.rackspace.com/hosting_knowledge/whitepapers/Understanding-the-Cloud-Computing-Stack.pdf

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Where some of the cases where it may not be the best option, where there may be regulatory compliance on all legal issues to make things you cannot make your data, your computing on the other premises or you cannot outsource the data storage or computing. There can be some of the legal issues or regulatory bindings on the things or where the highest level of performance are required on-premise or dedicated sources are required like I say some of the mission critical operations, where you require some of the highest level performance, where you want to customize the whole thing right that in those cases this may not be a good solution having a generic infrastructural realize.

After all when you are having these VMs, so you are having multi things like you have underlying hardware over that some virtual machine monitor and then you are realizing different VM. So, what happened all these stuff is basically takes some amount of the resources and computational time. So, when it is a mission critical where you want highest level of performance, so you may not be looking for this.

(Refer Slide Time: 21:22)

The slide has a dark blue header and footer. The main content area has a yellow background. At the top left, it says 'SaaS providers'. Below that is a table with three columns: 'Provider', 'Software', and 'Pricing model'. The table lists ten providers with their respective software types and pricing models. At the bottom of the slide, there is a source link and logos for IIT Kharagpur and NPTEL.

Provider	Software	Pricing model
Salesforce.com	CRM	Pay per use
Google Gmail	Email	Free
Process Maker Live	Business process management	Pay per use
XDrive	Storage	Subscription
SmugMug	Data sharing	Subscription
OpSource	Billing	Subscription
Appian Anywhere	Business process management	Pay per use
Box.net	Storage	Pay per use
MuxCloud	Data processing	Pay per use

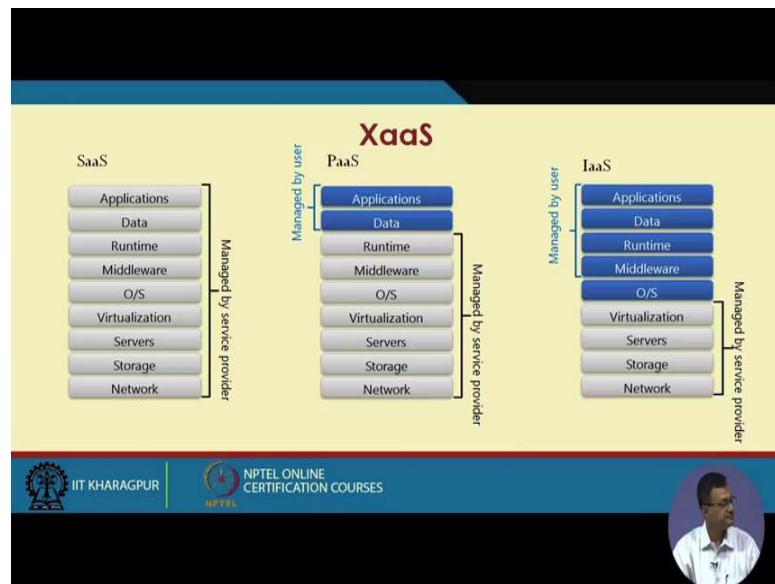
Source: <http://www.cs.helsinki.fi/u/epitavola/seminarit/Cloud%20Service%20Models.pdf>

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, these are some of the popular SaaS provider, we are not going much detail into the things. So, they have different pricing model like some are pay as you go, some of the provider are free, some are subscription base that amount of you subscribe for a year and then use it and type of things. These are feature of like PaaS provider. So, we are having different PaaS provider and the provide different type of thinks like some are on the using databases, big table on data bases. So, the different type of characteristics you can see that here different type of target to use, some dot NET platform, web development enterprise application, etcetera.

So, it has different type of flavors to offer to that. And there are IaaS provider major IaaS provider there are several other IaaS providers also. And they have different type of capacities like there you can have different type of hardware capacities, there are different types of operating systems and we have different type of billing mechanisms into the things. So, these are the different type of IaaS provider.

(Refer Slide Time: 22:41)



And as we have discussed already and discussing. So, if you if you relook at that again typical stack of these components then what we see that in case of a SaaS up to the application level the manage by the service provider. So, it is at the application level, the responsibility of the cloud service provider look in, whereas in case of a XaaS the responsibility is up to this runtime things, so that data and applications are the users. Whereas in case of infrastructure up to virtualization and providing VMs is the responsibility of the provider, whereas over that sometimes it will provide the provider will provide the guest OS or you can allow you to load the guest OS actually basically you can choose your guest OS to be loaded on the virtual machine, so that. And rest of the things is basically a responsibility of the IaaS consumer or the user.

(Refer Slide Time: 23:38)

Role of Networking in Cloud Computing

- In cloud computing, network resources can be provisioned dynamically.
- Some of the networking concepts that form the core of cloud computing are Virtual Local Area Networks, Virtual Private Networks and the different protocol layers.
- Examples of tools that help in setting up different network topologies and facilitate various network configurations are OpenSSH, OpenVPN etc.

Source: <http://www.slideshare.net/alexamies/networking-concepts-and-tools-for-the-cloud>

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, if we look at another aspect of these clouds is the networking. This networking plays a very what we say important role critical role I should say in case in realization of the things because at the bare metal you are having different hardware resources right. So, they are connected over the network those who are those who have seen these different aspects different say rack server or the specially this blade server etcetera, you will find that that is the backbone network is very high redundant type of things, because these all these servers at a extremely high operating speed work with these networks. So, network is an important role in cloud computing.

There is another aspect of networking that whether I can virtualize this network. Like I say I have a typical network scenario for my organizational need. So, I want to setup a network for organizing a particular event and I have a thing like I can say that these are the these are my routers and these are my servers and so on and so forth. Now, one way of looking at is that I purchase the thing separate things and connect the things right other whether I can virtualizes this whole infrastructure itself, like I have a what I mean to say the as I get a virtual machine whether I can create a overall a virtual network. So, I realize this network for the time being where the events will be there, it may be some examination, it may be some other division making something etcetera and then I give away the network on the things.

So, in other words, I have some sort of a soft way of realizing this overall networking, so that is another aspect of this. So, in cloud computing network resources can be provisioned dynamically as we have seen. Some of the networking concept that can form the cloud core of the cloud computing are virtual local area network right or what we popularly say VLAN, VPN and different protocol layers. So, this allows me to have realize this example of tools that help in there are different tools and techniques like OpenSSH, OpenVPN and some of the popular thing.

(Refer Slide Time: 26:05)

Networking in different Cloud Models

OSI Layer	Example Protocols	IaaS	PaaS	SaaS
7 Application	HTTP, FTP, NFS, SMTP, SSH	Consumer	Consumer	Provider
6 Presentation	SSL, TLS	Consumer	Provider	Provider
5 Session	TCP	Consumer	Provider	Provider
4 Transport	TCP	Consumer	Provider	Provider
3 Network	IP, IPsec	Consumer	Provider	Provider
2 Data Link	Ethernet, Fibre channel	Provider	Provider	Provider
1 Physical	Copper, optic fibre	Provider	Provider	Provider

Source: <http://www.slideshare.net/alexamies/networking-concepts-and-tools-for-the-cloud>

 IIT KHARAGPUR |  NPTEL ONLINE CERTIFICATION COURSES



And if you have networking in different cloud models, so if you see that these are our typical OSI stack. And if we look at that in the IaaS, so in IaaS up to this data link layer is provided by the provider rest of the consumer responsibility. On the other hand, in case of a SaaS, the responsibility of the providers goes up to the whole stack right. So, where in case of a PaaS this what we have the provider has a responsibility up to the presentation or sometimes session layer etcetera that means, it gives a platform. So, if I want to realize these OSI as a service model then I have different type of realization when you look as a whole infrastructure, whole as a platform or as a software as a service. So, what we see that networking plays a important role into the things.

(Refer Slide Time: 27:09)

Network Function Virtualization

- Network Functions Virtualization aims to transform the way that network operators architect networks by evolving standard IT virtualization technology to consolidate many network equipment types onto industry standard high volume servers, switches and storage, which could be located in Datacentres, Network Nodes and in the end user premises.
- It involves the implementation of network functions in software that can run on a range of industry standard server hardware, and that can be moved to, or instantiated in, various locations in the network as required, without the need for installation of new equipment.

Ref: https://portal.eti.org/nfv/nfv_white_paper.pdf

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES



There is another concept what we say network function virtualization will just have a quick introduction. It aims to transform the way the network operator architects the networks by evolving standard it virtualization technology to consolidate many network equipment types into the industry standard switches etcetera could be located in datacenters. What it tries to say it says that it tries to emulate in it work over your over the infrastructure, as we are discussing. It involves implementation of the network functions in software that can run on a range of industry standard server hardware and that can be move instantiated in and various location of the network as required without need of installation of new equipment.

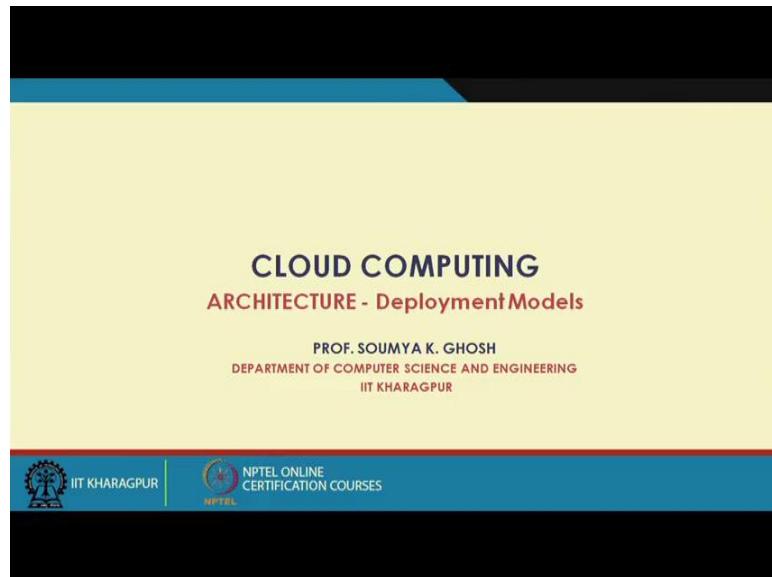
That means as if I will little bit discuss in our subsequent talk. So, as if I have a infrastructure and I realize several network over the things, like I these are the classical approaches and here I have infrastructure then I try to realize different network over this basic infrastructure or I can have different virtual networks over the same infrastructure. So, if it is possible then not only this having this different virtual machines, I can have different virtual networks for the things or I can realize a virtual network on that. So, will stop here for today's lecture and we will continue with our discussion in subsequent talks.

Thank you.

Cloud Computing
Prof. Soumya Kanti Ghosh
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Lecture – 06
Architecture - Deployment Models

(Refer Slide Time: 00:41)



Hello, so welcome to our next lecture on cloud computing. Today, we will continue some of our discussion on cloud architecture and specially see some aspects of a cloud taken is cloud like one of the aspects is virtualization.

(Refer Slide Time: 00:48)

Deployment Models

- Public Cloud
- Private Cloud
- Hybrid Cloud
- Community Cloud

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

46

So, as if we look at remember that earlier lectures we are discussing on different type of a service models, and also we have different type of deployment models in cloud namely public, private, hybrid and community cloud. Different aspects and all sort of services can be hosted in different type of deployment models, right.

(Refer Slide Time: 01:05)

Public Cloud

Cloud infrastructure is provisioned for open use by the general public. It may be owned, managed, and operated by a business, academic, or government organization, or some combination of them. It exists on the premises of the cloud provider.

Enterprise to Cloud

Source: Marcus Hogue, Chris Jacobson, "Security of Cloud Computing"

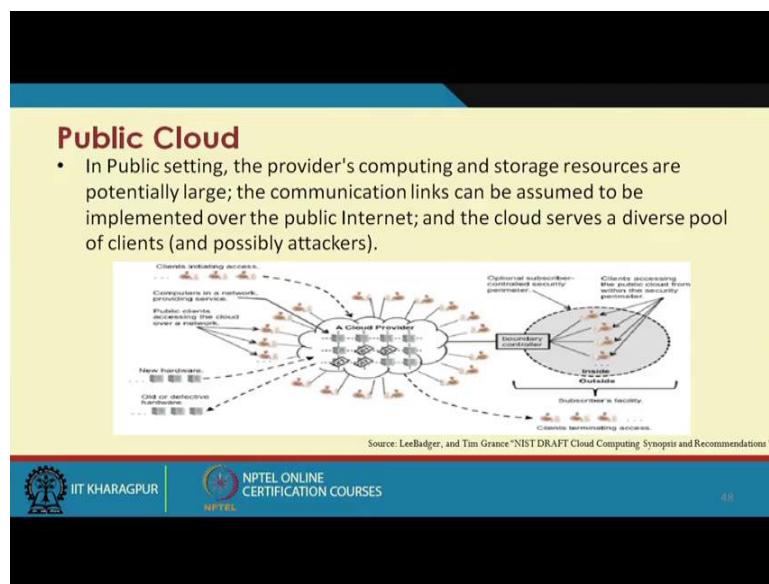
IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, in case of a public cloud the as the name suggests it is available for public at large. So, it is you anyone can purchase that and it is somewhat omnipresent across the internet. Some of the very popular examples are Google app engine, Microsoft Azure, IBM cloud,

Amazon EC2 and many others right many, many others clouds are there. So, what happened that we have this public cloud and enterprise or individual can subscribe this public cloud, subscribe this public cloud over the internet and can have that its services over this. So, the cloud infrastructure is provisioned for open use by general public, organization, enterprises and anyone who can pay and use the things, there are definitely some legal policy issues, we need to be conformed too.

So, it may be owned managed and operated by a business, academics, government organization or some combination of them, right. It exists in the premise of the cloud provider. So, typically the physically the public cloud is at the CSPs premise or premises, so, that means whatever the computing infrastructure storage infrastructure and other type of things are there those are residing in the CSPs or the cloud providers premises not at the private or not at the users premises, so that is one aspects of the thing.

(Refer Slide Time: 02:42)



And in public setting, provider's computing and storage resources are potentially large, right, so it is serving to all. Communication links can be assumed to be implemented over public Internet, services; and the cloud service serves a diverse pool of client and may be out of them do not all faithful clients there can be some attackers, hackers etcetera, etcetera. So, it is open to anybody who can subscribe a typically can have a service provider and you can there can be different type of users at the things.

(Refer Slide Time: 03:19)

Public Cloud

- **Workload locations are hidden from clients (public):**
 - In the public scenario, a provider may migrate a subscriber's workload, whether processing or data, at any time.
 - Workload can be transferred to data centres where cost is low
 - Workloads in a public cloud may be relocated anywhere at any time unless the provider has offered (optional) location restriction policies
- **Risks from multi-tenancy (public):**
 - A single machine may be shared by the workloads of any combination of subscribers (a subscriber's workload may be co-resident with the workloads of competitors or adversaries)
 - Introduces both reliability and security risk

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES



So, what are the typical features workload locations are hidden from the clients, one of the clients that is one of the major thing. Like you do not know where your virtual machine is you do not know where your actually the data is residing, which server which location and with whom it is residing, so it is all are hidden. So, as far as if you are not very stringent on the legal and policy matter about the security and other aspects though it is fine that you do not care, so long your services are. There are risk from multi-tenancy; that means, your logically or it may be theoretically always possible, that your computing your storage where it is residing somebody else's things are there. Now, if it is somebody with some organization or some person who is not very faithful, or we are not very comfortable, so that two things two some two different user can reside can work on the same things.

So, in other sense there is a risk there is this; what we say multi-tenancy and there are risks of multi-tenancy because I do not know that where things are there, where there though whether there is a underlining channel to access my data services and other type of things. So, there is a risk of multi-tenancy. So, single machine can be shared by workloads by any combination of subscriber, subscriber workload maybe co-resident with the workload of the competitor or adversaries, so it introduce both reliability and security risk.

(Refer Slide Time: 04:55)

Public Cloud

- Organizations considering the use of an on-site private cloud should consider:
 - **Network dependency (public):**
 - Subscribers connect to providers via the public Internet.
 - Connection depends on Internet's Infrastructure like
 - Domain Name System (DNS) servers
 - Router infrastructure,
 - Inter-router links

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, organization considering use of onsite public cloud should consider network dependency. So, whenever suppose IIT, Kharagpur things that all is or some of its labs will be running on public cloud, so that our overall maintaining and etcetera reduces cost of maintenance or overall load on maintenance etcetera reduces. Now, there first dependency is the network. So, it will be always available the network connectivity should be always available up to a mark. So, there is one dependency.

(Refer Slide Time: 05:31)

Public Cloud

- **Limited visibility and control over data regarding security (public):**
 - The details of provider system operation are usually considered proprietary information and are not divulged to subscribers.
 - In many cases, the software employed by a provider is usually proprietary and not available for examination by subscribers
 - A subscriber cannot verify that data has been completely deleted from a provider's systems.
- **Elasticity: illusion of unlimited resource availability (public):**
 - Public clouds are generally unrestricted in their location or size.
 - Public clouds potentially have high degree of flexibility in the movement of subscriber workloads to correspond with available resources.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

There are limited visibility and control over the data regarding the security. So, we are mentioning that I have limited visibility of the data. I do not know where the data is and how it is secured only thing what I have is some sort of a SLA or some sort of a MOU between provider and me that this data is secured and so and so forth. So, there is an issue of elasticity or illusion of unlimited resource availability. So, this is when you use public cloud, this is pretty fine because theoretically I have infinite amount of elasticity like if I need more computing power it will be provision if I one more other storage things it will be provisions when I do not require I release it, or de provision it. So, those things are feasible. So, theoretically infinite scaling up, scaling down is possible.

(Refer Slide Time: 06:31)

The slide has a dark blue header bar at the top. Below it is a light yellow main content area. At the bottom is a dark blue footer bar containing logos and text. The title 'Public Cloud' is centered in the yellow area. To its right is a large, semi-transparent black rectangular box. Below the title, there are three bullet points:

- **Low up-front costs to migrate into the cloud (public)**
- **Restrictive default service level agreements (public):**
 - The default service level agreements of public clouds specify limited promises that providers make to subscribers

In the footer bar, from left to right, there are: the IIT Kharagpur logo, the text 'IIT KHARAGPUR', the NPTEL logo, the text 'NPTEL ONLINE CERTIFICATION COURSES', and the number '52'.

Another important thing is the low up-front cost to migrate into the cloud. So, if you want to make a private cloud of your own, then you have to purchase the thing make provision where it will be housed, install software and etcetera; run it test it there are issues of maintenance so and so. Here the up-front, there is no there is very low up-front cost, you pay and use it. Restrictive default service level agreements, so there is a now whenever we purchase something there is a somewhere other we need to confirm to the that standard or what we say quote, unquote restrictive service level agreements between the provider and the consumer. So, most of the cases we need to follow the terms and condition provide like whatever is given by the provider unless you do for a large-scale deployment where you negotiate at special rate with special SLA and type of things. But

normally for a small institution and public at large we need to confirm to the; whatever is being provided.

(Refer Slide Time: 07:38)

Private Cloud

- The cloud infrastructure is provisioned for exclusive use by a single organization comprising multiple consumers (e.g., business units). It may be owned, managed, and operated by the organization, a third party, or some combination of them, and it may exist on or off premises.

- Examples of Private Cloud:
 - Eucalyptus
 - Ubuntu Enterprise Cloud - UEC
 - Amazon VPC (Virtual Private Cloud)
 - VMware Cloud Infrastructure Suite
 - Microsoft ECI data center.

So, totally other means the other aspects form the public is the private. So, you have your own cloud and you have all your resources which can be working on it right. So, the cloud infrastructure is provision for exclusive use of a single organization comprising a multiple consumers or business units like a same organization say IIT KGP private cloud my catering all the things which in the IIT departments, sections etcetera. May be owned managed and operated by the organization or it can be outsourced by a third party managing resources out here, but it is in your premises under your jurisdiction under your network control and type of things. And it may exist on or off premises also.

So, that there are usually on premises or I can say I have a private things or what we say outsource private cloud, where I can at the off premises, but nevertheless it is jurisdiction or my policy stipulated rules or the organization rules which driven. So, there are some of the open source and other public cloud one is that Eucalyptus pretty popular, there are open stack, Ubuntu Enterprise Cloud, Amazon VPC - virtual private cloud, VMware Cloud Infrastructures Suite, Microsoft ECI data centers and so on and so forth. So, there are several things which gives private cloud into the thing.

(Refer Slide Time: 09:04)

Private Cloud

- Contrary to popular belief, private cloud may exist off premises and can be managed by a third party. Thus, two private cloud scenarios exist, as follows:
- On-site Private Cloud
 - Applies to private clouds implemented at a customer's premises.
- Outsourced Private Cloud
 - Applies to private clouds where the server side is outsourced to a hosting company.

 IIT KHARAGPUR |  NPTEL ONLINE CERTIFICATION COURSES



So, contrary to popular belief, private cloud may exist off premises and can be managed by third party. So, not only means take the responsibility or I basically I want to maintain the control over the whole thing, but I basically may off premise or I installed out with the help of a third party to a separate thing also. Thus, two private cloud scenarios one is on-site private cloud which is the de facto or which is immediately come to which comes to our mind when we are talking about anything which is private, applies to private clouds implemented as the customer's premises. Another is the outsource private cloud like I have a private chunk of the things which is outside outsource out of my premises, but never the less it is private to me, right. So, applies to private cloud where the server side is outsourced to the hosting company wherever it is there.

(Refer Slide Time: 10:02)

The diagram illustrates the On-site Private Cloud security perimeter. It features a central cloud icon labeled 'A Private Cloud' with several small icons representing users or resources. This central area is enclosed by a dashed oval labeled 'Subscriber-controlled security perimeter'. The oval is divided into two regions: 'Inside' at the bottom and 'Outside' at the top. An arrow labeled 'Blocked access.' points from the Outside region towards the perimeter. Another arrow labeled 'Clients accessing the private cloud from within the security perimeter.' points from the Inside region towards the perimeter. A 'boundary controller' box is positioned at the bottom right of the perimeter, with a 'Legitimate access path.' arrow pointing from it to the Inside region. The entire diagram is set against a light yellow background.

On-site Private Cloud

- The security perimeter extends around both the subscriber's on-site resources and the private cloud's resources.
- Security perimeter does not guarantee control over the private cloud's resources but subscriber can exercise control over the resources.

Source: Leebadger, and Tim Grance "NIST DRAFT Cloud Computing Synopsis and Recommendations"

IIT Kharagpur | NPTEL ONLINE CERTIFICATION COURSES

So, in case of on-site private cloud the security perimeter extends around both the subscriber's on-site resources and private cloud's resources. So, your security perimeter or your legal control is basically to encompass the private cloud right. Security perimeter does not guarantee control over private cloud resources, but the subscriber can exercise control over the other resources, over the resources like. So, that it is in case of onsite like whatever the private cloud it is there, I can have overall control over the whole resources of the private cloud.

(Refer Slide Time: 10:39)

The diagram illustrates the On-site Private Cloud security perimeter, similar to the one above but with more detailed text. It shows a central cloud icon with user/resource icons, surrounded by a dashed oval perimeter. The Inside region is at the bottom, and the Outside region is at the top. An arrow labeled 'Blocked access.' points from the Outside to the perimeter. Another arrow labeled 'Clients accessing the private cloud from within the security perimeter.' points from the Inside to the perimeter. A 'boundary controller' box is at the bottom right, connected to the Inside region by a 'Legitimate access path.' arrow. The entire diagram is set against a light yellow background.

On-site Private Cloud

- Organizations considering the use of an on-site private cloud should consider:
 - Network dependency (on-site-private):
 - Subscribers still need IT skills (on-site-private):
 - Subscriber organizations will need the traditional IT skills required to manage user devices that access the private cloud, and will require cloud IT skills as well.
 - Workload locations are hidden from clients (on-site-private):
 - To manage a cloud's hardware resources, a private cloud must be able to migrate workloads between machines without inconveniencing clients. With an on-site private cloud, however, a subscriber organization chooses the physical infrastructure, but individual clients still may not know where their workloads physically exist within the subscriber organization's infrastructure

IIT Kharagpur | NPTEL ONLINE CERTIFICATION COURSES

So, there are some issues characteristics pros and cons of the maintaining onsite private cloud one is the network dependency onsite private right. Subscriber, so it is dependency on the on your Internet network should be here. Subscriber still needs IT skill my organization is maintain my own cloud or the organization maintain own cloud. So, there should be some sort of a skill to maintain that. Workload location are hidden from the client, even if my clients are different differ in my subunits that also this is hidden from the client. Even if it is within the premises or on-site private cloud that actually infrastructure is hidden from the cloud, where client is my own organization things or my own clients are on the other side.

(Refer Slide Time: 11:24)

On-site Private Cloud

- **Risks from multi-tenancy (on-site-private):**
 - Workloads of different clients may reside concurrently on the same systems and local networks, separated only by access policies implemented by a cloud provider's software. A flaw in the software or the policies could compromise the security of a subscriber organization by exposing client workloads to one another
- **Data import/export, and performance limitations (on-site-private):**
 - On-demand bulk data import/export is limited by the on-site private cloud's network capacity, and real-time or critical processing may be problematic because of networking limitations.

 IIT KHARAGPUR |  NPTEL ONLINE CERTIFICATION COURSES



Risk from multi-tenancy again the same issues of within things also come into play. Data import export and performance limitation there can be issues of data import export because there are lot of data, which will be going out and going down. So, that on demand bulk data import export is limited on on-site private clouds network capacity or real time critical processing maybe problematic because of network limitation.

(Refer Slide Time: 11:55)

On-site Private Cloud

- **Potentially strong security from external threats (on-site-private):**
 - In an on-site private cloud, a subscriber has the option of implementing an appropriately strong security perimeter to protect private cloud resources against external threats to the same level of security as can be achieved for non-cloud resources.
- **Significant-to-high up-front costs to migrate into the cloud (on-site-private):**
 - An on-site private cloud requires that cloud management software be installed on computer systems within a subscriber organization. If the cloud is intended to support process-intensive or data-intensive workloads, the software will need to be installed on numerous commodity systems or on a more limited number of high-performance systems. Installing cloud software and managing the installations will incur significant up-front costs, even if the cloud software itself is free, and even if much of the hardware already exists within a subscriber organization.

 IIT KHARAGPUR |  NPTEL ONLINE CERTIFICATION COURSES



Potentially strong security from external threats, usually if it is a private within your network boundary all your network other features come into play. Like as I was mentioning in some of my in one of my earlier lecture that IIT, Kharagpur has installed or has developed a private cloud for its research purpose what is Meghamala, but it is within my network premise. So, what whatever the network security parameters or features are there for IIT, Kharagpur is also applied for this infrastructure. So, what happens that it has a potentially strong security features. Significant to high upfront cost to migrate into the cloud, so that is another issue, right. Whenever you have a private cloud. So, there is a significant cost in installing maintaining and there is may be a significant cost in migrating the whole thing into the private cloud.

(Refer Slide Time: 12:50)

On-site Private Cloud

- **Limited resources (on-site-private):**
 - An on-site private cloud, at any specific time, has a fixed computing and storage capacity that has been sized to correspond to anticipated workloads and cost restrictions.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

NPTEL

There is a limited resources all if you have your thing that anything, you want to augment you need to purchase install not only that even to need to properly interoperate with the existing things. So, in doing so, you have a times, you have a limited resources like suddenly I can go up or down on the resources. So, as on-site private cloud any specific time has a fixed computing and storage capacity that can be sized to corresponding, correspond to the anticipated workloads and cost. So, what we do whenever I install a private cloud, so I have a estimate of the things like storage, computing, etcetera and then keeps some provision like of it is a buying that the staff at x amount or I install 1.5 amount. But that is the thing that I am limited to that 1.5x amount of the thing.

(Refer Slide Time: 13:47)

The diagram illustrates the architecture of an Outsourced Private Cloud. It features a central 'Private Cloud' represented by a cloud icon containing a server rack. Two security perimeters are shown: one 'Outsourced private cloud perimeter' surrounding the private cloud, and another 'Cloud provider's perimeter' surrounding the entire system. A 'Protected communication link' connects the two perimeters. An inset diagram shows a detailed view of the communication link, labeled 'External communication channel', with arrows indicating data flow between the two perimeters.

Outsourced Private Cloud

- Outsourced private cloud has two security perimeters, one implemented by a cloud subscriber (on the right) and one implemented by a provider.
- Two security perimeters are joined by a protected communications link.
- The security of data and processing conducted in the outsourced private cloud depends on the strength and availability of both security perimeters and of the protected communication link.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, there is another variant of the things I keep this as private, but I outsource it, so that maintaining installing etcetera, I do not means, organization do not take care, but it is outsource it, so outsource without source. Outsource private cloud has two security perimeters one implemented by the cloud subscriber, whoever is there on the right and the implemented by the provider. So, one this is a perimeter. So, what is happening it has a some sort of a channel which connects this to this private cloud which is outsource in some other premises right or maybe a subset of a cloud service provider.

So, I have a channel which hooks into the things, but the whole stuff at that end is a private to me. So, the security of data and processing conducted on the outsourced private cloud depends on the strength and availability of both security perimeters and of the protected communication. So, what we require that my infrastructure to be secured at the external things, another the channel where by the network channel or the network communication link which I talk over which I communicate or over which my organization communicate with the cloud should be secured in up to a particular level or expected level.

(Refer Slide Time: 15:13)

Outsourced Private Cloud

- Organizations considering the use of an outsourced private cloud should consider:
 - **Network Dependency (outsourced-private):**
 - In the outsourced private scenario, subscribers may have an option to provision unique protected and reliable communication links with the provider.
 - **Workload locations are hidden from clients (outsourced-private):**
 - **Risks from multi-tenancy (outsourced-private):**
 - The implications are the same as those for an on-site private cloud.

 IIT KHARAGPUR |  NPTEL ONLINE CERTIFICATION COURSES
NPTEL

61

So, there are again some consideration pros and cons of using outsource private cloud. One is network dependency that again I am dependent on how things will be connected. Workload location are hidden from the client again those type of issues, it is from multi tenancy, where I am hosting my private cloud other people may be hosting also the private cloud. So, data import, export and performance limitation same thing exist. Potentially strong security from external threat because of you have still have a private things, it is not fully public and not all people are jumping on your cloud, but nevertheless you are maybe sharing some infrastructure may be more thereat at the much more lower level at the highest level and so on. But at the higher level you are not allowing anybody to enter into the things.

(Refer Slide Time: 16:01)

Outsourced Private Cloud

- **Modest-to-significant up-front costs to migrate into the cloud (outsourced-private):**
 - In the outsourced private cloud scenario, the resources are provisioned by the provider
 - Main start-up costs for the subscriber relate to:
 - Negotiating the terms of the service level agreement (SLA)
 - Possibly upgrading the subscriber's network to connect to the outsourced private cloud
 - Switching from traditional applications to cloud-hosted applications,
 - Porting existing non-cloud operations to the cloud
 - Training

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Modest to significant up-front costs to mitigate clouds, so those are same as whatever we are having at the things. And most of the cases you need to negotiate in terms SLA with the provider who is providing your or the third party who is provide you this cloud.

(Refer Slide Time: 16:16)

Outsourced Private Cloud

- **Extensive resources available (outsourced-private):**
 - In the case of the outsourced private cloud, a subscriber can rent resources in any quantity offered by the provider. Provisioning and operating computing equipment at scale is a core competency of providers.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES



Extensive resource availability is maybe an advantage, because this is not limited. I am taking a chance, so I request for increase it may it is very much possible to increase at the other end, the provider is not out of resources usually they have lot of resource at their backbone.

(Refer Slide Time: 16:38)

Community Cloud

- Cloud infrastructure is provisioned for exclusive use by a specific community of consumers from organizations that have shared concerns (e.g., mission, security requirements, policy, and compliance considerations). It may be owned, managed, and operated by one or more of the organizations in the community, a third party, or some combination of them, and it may exist on or off premises.

- Examples of Community Cloud:
 - Google Apps for Government
 - Microsoft Government Community Cloud

Community Cloud

Diagram illustrating the architecture of a Community Cloud:

- The diagram shows two main components: a "Community" cloud and an "Enterprise" cloud.
- Both the Community and Enterprise clouds are connected to a central "Cloud (Public or Private)".
- Arrows indicate the flow of data between the local clouds and the central cloud.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, one side is private one side is public another typical type of cloud is community cloud right. So, it basically tries to as we have discussed it basically tries to serve a particular community per say, it is usually can operate in a public or private both, and it basically cater to a particular community which has a somewhat same domain of operation or same focus of interoperating. So, cloud infrastructure is provision for a exclusive use of a specific community of consumer from organization that are shared concerned that means, they have a likeminded concern that is there can be same missions, security requirement, policy compliance consideration, etcetera, it may be owned managed operated by one or more organization in the community.

So, a third party or some combination of that it may exist on or off premises. So, it can be a on premises off premises, there are several community cloud and which are being provided by different service provider.

(Refer Slide Time: 17:47)

On-site Community Cloud

- Community cloud is made up of a set of participant organizations. Each participant organization may provide cloud services, consume cloud services, or both
- At least one organization must provide cloud services
- Each organization implements a security perimeter

Source: LeeBadger, and Tim Grance 'NIST DRAFT Cloud Computing Synopsis and Recommendations'

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, there are there can be one thing that like there are several A, B, C organization there X, Y, Z organization and there can they can form different set of combinations like ABC, XYZ can be one community; A with X, Y can be another community and so and so forth. So, there is possibility of bringing things together there is also possibility that I a community can be existing as some point of time; at the some other point of time it may not existing, I may be a organization can be more than one community of the things like our day-to-day life. I may be a part of my office group; also I am part of my say residential community.

So, there can be different policies and etcetera things are there, but it is the primary objective is that there are like or same type of concerned or same type of workflow it may so happened that this community making them in a single community will help in productivity.

(Refer Slide Time: 18:58)

On-site Community Cloud

- The participant organizations are connected via links between the boundary controllers that allow access through their security perimeters
- Access policy of a community cloud may be complex
 - Ex. :if there are N community members, a decision must be made, either implicitly or explicitly, on how to share a member's local cloud resources with each of the other members
 - Policy specification techniques like role-based access control (RBAC), attribute-based access control can be used to express sharing policies.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES



There are again lots of characteristics pros and cons, etcetera. The participant organization are connected by a links between the boundary controllers, and allow access through their security parameters like whatever the firewall policies or that type of boundary policies are there. Access policy of a community cloud may be a pretty complex, because you can have number of community. So, at what way access which you do not access whether there is a leakage of information, I get some information someone community pass it to the other community, so these need to be properly restricted. So, policy specification techniques like role based access control, attribute based access control are there; like based on my role I access some data, right. And like other form of deployment models, here also we have network dependency.

(Refer Slide Time: 19:48)

On-site Community Cloud

- **Subscribers still need IT skills (on-site-community).**
 - Organizations in the community that provides cloud resources, requires IT skills similar to those required for the on-site private cloud scenario except that the overall cloud configuration may be more complex and hence require a higher skill level.
 - Identity and access control configurations among the participant organizations may be complex
- **Workload locations are hidden from clients (on-site-community):**
 - Participant Organizations providing cloud services to the community cloud may wish to employ an outsourced private cloud as a part of its implementation strategy.

Subscriber still need some IT skills because, it need to maintain with different community things. Workload locations are hidden from the client again.

(Refer Slide Time: 20:00)

On-site Community Cloud

- **Data import/export, and performance limitations (on-site-community):**
 - The communication links between the various participant organizations in a community cloud can be provisioned to various levels of performance, security and reliability, based on the needs of the participant organizations. The network-based limitations are thus similar to those of the outsourced-private cloud scenario.
- **Potentially strong security from external threats (on-site-community):**
 - The security of a community cloud from external threats depends on the security of all the security perimeters of the participant organizations and the strength of the communications links. These dependencies are essentially similar to those of the outsourced private cloud scenario, but with possibly more links and security perimeters.

Data import export performance limitations there are issues on that like how between the community things etcetera whether the data or within the community, when multiple subscriber come in to play that how things will be there, and number of cases this communities can be loosely couples, so that things becomes more critical to manage potentially strong security from the external thing because still you are in the one

community, so that you have a better resistance to the external threats based on your community policies along with your own policy.

(Refer Slide Time: 20:34)

On-site Community Cloud

- Highly variable up-front costs to migrate into the cloud (on-site-community):
 - The up-front costs of an on-site community cloud for a participant organization depend greatly on whether the organization plans to consume cloud services only or also to provide cloud services. For a participant organization that intends to provide cloud services within the community cloud, the costs appear to be similar to those for the on-site private cloud scenario (i.e., significant-to-high).

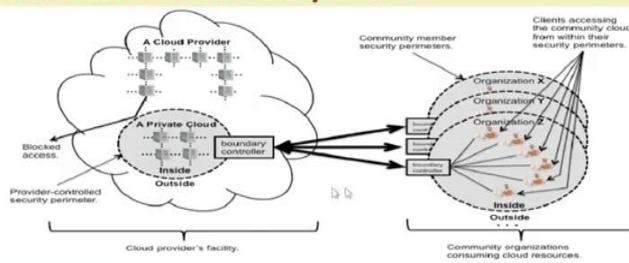
 IIT KHARAGPUR |  NPTEL ONLINE CERTIFICATION COURSES



High variable up-front costs to migrate to the cloud, so there is as we have seen in case of a truly private cloud there is a high variable upfront cost to the migrate of the migration to the cloud because it is not publicly available. So, you need to create the things.

(Refer Slide Time: 20:52)

Outsourced Community Cloud



The diagram illustrates an Outsourced Community Cloud architecture. It features two main components: a "Cloud provider's facility" on the left and "Community organizations consuming cloud resources" on the right. A central "Boundary controller" manages access between them. The "Cloud provider's facility" contains a "Private Cloud" and a "Cloud Provider". The "Private Cloud" has a "Provider-controlled security perimeter" and "Blocked access" to the "Community member security perimeters". The "Boundary controller" is connected to both the "Cloud provider's facility" and the "Community organizations consuming cloud resources". The "Community organizations consuming cloud resources" have their own "Community member security perimeters" and "Clients accessing the community cloud from within their security perimeters".

Source: LeeBadger, and Tim Grace "NIST DRAFT Cloud Computing Synopsis and Recommendations"

 IIT KHARAGPUR |  NPTEL ONLINE CERTIFICATION COURSES



And there can be different sort of things like there are three organizations forming a community with some boundary controller, where with a private subscriber and those issues come into play.

(Refer Slide Time: 21:04)

Outsourced Community Cloud

- Organizations considering the use of an on-site community cloud should consider:
- **Network dependency (outsourced-community):**
 - The network dependency of the outsourced community cloud is similar to that of the outsourced private cloud. The primary difference is that multiple protected communications links are likely from the community members to the provider's facility.
- **Workload locations are hidden from clients (outsourced-community).**
 - Same as the outsourced private cloud

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES



Community cloud can be on the premises or the community cloud can be out of means premises that means, the community cloud can be outsource as we have seen that in case of a private cloud. So, once we outsource the network dependency, workload location are not known or hidden from the clients.

(Refer Slide Time: 21:25)

Outsourced Community Cloud

- **Risks from multi-tenancy (outsourced-community):**
 - Same as the on-site community cloud
- **Data import/export, and performance limitations (outsourced-community):**
 - Same as outsourced private cloud
- **Potentially strong security from external threats (outsourced-community):**
 - Same as the on-site community cloud
- **Modest-to-significant up-front costs to migrate into the cloud (outsourced-community):**
 - Same as outsourced private cloud

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES



Risk from multi-tenancy, data import export and performance limitation issues, this will come into play potentially strong security from external threats as we have mentioned that is still you are on the community. Modest-to-significant up-front cost, if it is outsource there are lot of loads are taken up by the outsource this is in organization. So, there is a chance that the overall loading maybe overall up-front cost will be much less than if you are maintaining in premises. And theoretically if you are outsourcing extensive resource availability are possible.

(Refer Slide Time: 22:07)

Hybrid Cloud

- The cloud infrastructure is a composition of two or more distinct cloud infrastructures (private, community, or public) that remain unique entities, but are bound together by standardized or proprietary technology that enables data and application portability
- Examples of Hybrid Cloud:
 - Windows Azure (capable of Hybrid Cloud)
 - VMware vCloud (Hybrid Cloud Services)

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, apart from this as we can theoretically see or practically see that I can have a cloud which is combination of all those things. So, I can have private, public, community things. Specially with the public private cloud I can have a cloud which is combination of such more than one type of deployment models. The cloud infrastructure is a composition of two more distinct cloud infrastructure private, community or public. So, I have three type of things, and then I want to realize a cloud which has a combination of a these three things.

Why, this is important, first of all it is all depends that what sort of uses pattern, I am having. Like some of the uses pattern what I am having is more critical or more vulnerable to security threats that I want to keep as a more private. I do not want to I have a appropriate network boundary or network perimeter security to be implemented on the things. There are some of the resources which may not be I may not want those to

be so much secure or I do not care about all those security of the all those things and that can be made some of the things public. Like if I say if there are practice sessions for say computing labs for students, so the level of security is much less than when I am keeping say student records or students examination things etcetera, right.

So, though same type of operations may be there, but one I could have gone away with and an outsource this and gone to the public cloud to do it all right. And whereas the other one even it is economical, I want to keep those as my private things. Now, I can have a private combinations, right. As number of cases what sometimes it happens that you do something with the private, and you require some resources to be provisioned due to sudden increase of the things. Then suddenly in a private cloud increasing the resource provisioning or purchasing etcetera is a long process. So, you purchase the thing on a public cloud for a short period of time, so long your process is an place and this goes to the things.

So, the infrastructure community that remains unique entities, but are bounded together in standardize or proprietary technology enables data and application to be portable, this is important. So, portability not only with respect to data, whenever I have this private, public, community all together or a combination of two or more together, then other the issue of intra operative come into play like the data which is working fine here when I take some application from the other whether we still workout the things.

So, the both data and at times the applications suppose your application was running on a private cloud with some resources, now you provision a VM which basically goes to the public domain. Now, the types of applications whether the application wants need to be resized or there are portability issues of the applications need to be looked into. So, there are examples of hybrid cloud some of the popular Windows Azure capable of hybrid cloud, VMware V cloud; there are capability of hybrid cloud and as I have mentioning that there are several other providers which provide this type of things.

(Refer Slide Time: 25:45)

Hybrid Cloud

- A hybrid cloud is composed of two or more private, community, or public clouds.
- They have significant variations in performance, reliability, and security properties depending upon the type of cloud chosen to build hybrid cloud.

Source: LeeBadger, and Tim Grance "NIST DRAFT Cloud Computing Synopsis and Recommendations"

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

NPTEL

So, the hybrid cloud is composed of two or more private public etcetera, they have a significant variation in performance, reliability, security property depending upon the type of cloud chosen to build a hybrid, right. If it is a community cloud or public and etcetera it will there will be difference in the performance different in the security features, etcetera.

(Refer Slide Time: 26:02)

Hybrid Cloud

- A hybrid cloud can be extremely complex
- A hybrid cloud may change over time with constituent clouds joining and leaving.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

NPTEL

So, a hybrid cloud can be extremely complex that is one of the major things. Like suppose your particular application is going on and need to run over a combination of

public and private cloud then the overall underlining architecture maybe very complex, so that your application runs seamlessly over the things. So, at times these private clouds can be extremely complex. A hybrid cloud may change over time with constituent clouds joining or leaving there is another big factor. So, what we are trying that it may so happened that you build a hybrid cloud with your own private cloud and two other public clouds.

So, now it may so change that some of the public cloud may go may wants to disconnect based on your terms and conditions and subscription ends, they do not want to again re subscribe and they have a different pricing model, even some of the thing may go red right there the organization the cloud may not be there. So, in that case, you need to there can be joining, leaving, and joining of new things or sometimes you may require more resources. So, you add some more public or community cloud into the things. All these becomes extremely complex phenomena to handle, right, so that means, over time the constituent clouds may leave or joined and making the whole process pretty complex.

So, now what should I choose, right; what should be my deployment model is another big question right. It totally depends on your requirement. Like if I have a small organization or individuals my public cloud may be a good solution. So, long my business is not going up-front on the something which compiles me to go to a own private cloud. There are other constant like I do business for somebody else, right I have some other subscriber base or client base, now this client may be interned looking for a things like suppose I have a storage provider, data storage provider. Now, I may either I have all this storages on my premises or I outsource this resources or provision this resources from other public clouds.

Now, it may be so happened that the my clients which may be something misson critical clients like may be a financial sector or defense sector they want that no, no, no, it cannot happen, you need to have your own thing. So, it all depends that how what should be my way of looking at it right. I can have a combination as we have talking about hybrid I can have a combination of private public and so forth based on my requirement. Or whether I can classify my application into different things, my data, my applications into different categories, and then I say that this bunch can go to the private, this bunch can go to the public this can go to the community and type of things.

So, managing all those things is another big challenge or for the organization or institution to handle that. So, with this, we will close this lecture; and we will continue our lecture on other aspects of cloud computing in the subsequent talks.

Thank you.

Cloud Computing
Prof. Soumya Kanti Ghosh
Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur

Lecture – 07
Virtualization

Hi; so, welcome to our next lecture on cloud computing. We will quickly go through the virtualization concept; we have already discussed some of these aspects in our previous lectures. But I thought that we will have some few more slides on it, one of the core concepts. So, virtualization is one of the core concepts or one of the core prime mover or having these cloud computing. For which the cloud computing is existing today and if you see this virtualization is not a new concept, it is already there. We; many of our are used to virtualizes. We virtualize a Linux system over a window system over a Linux system. So, I have some sort of a virtual realization of the things.

And if you look at that other end, we are used to different other resources like networking, etcetera. So, I have a; we can have a virtual LAN or VLAN, which is pretty popular which we have VPN virtual private networks and so forth. So, virtualization is there and this cloud computing architecture technology, tries to exploit this feature for giving different services of cloud. So, as we have seen this IaaS or infrastructure as a service. So, from the subscriber point of view, what the subscriber gets. So, access to virtual computers, right, network accessible virtual storage, network infrastructure components like firewall configuration services, etcetera.

(Refer Slide Time: 02:03)

IaaS – Infrastructure as a Service

- What does a subscriber get?
 - Access to virtual computers, network-accessible storage, network infrastructure components such as firewalls, and configuration services.
- How are usage fees calculated?
 - Typically, per CPU hour, data GB stored per hour, network bandwidth consumed, network infrastructure used (e.g., IP addresses) per hour, value-added services used (e.g., monitoring, automatic scaling)

IT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, as if I being a subscriber, I get access to a virtual machine right or set of machines. And then, I may want that particular machine with particular storage and other configurations. So, I do not know where it is, how it is configured. But for me if through this particular interface; It is a type of machines which I am looking for. Even it is possible that, I have a combination of these virtual machines, along with a particular backbone network and realize a network infrastructure for my purpose right. So, this is that what subscriber gets from me and how uses and what I pay for it. Is basically typically, per CPU hour, GB data stored, maybe on a hourly basis. Network bandwidth consumed on a particular rate, network infrastructure used. That how much IP address which are the routers etcetera and value added there can be value added services. Like I monitoring automatic scaling and type of things are the things which are value added services.

(Refer Slide Time: 03:26)

IaaS Provider/Subscriber Interaction Dynamics

- Provider has a number of available virtual machines (VMs) that it can allocate to clients.
 - Client A has access to vm1 and vm2, Client B has access to vm3 and Client C has access to vm4, vm5 and vm6
 - Provider retains only vm7 through vmN

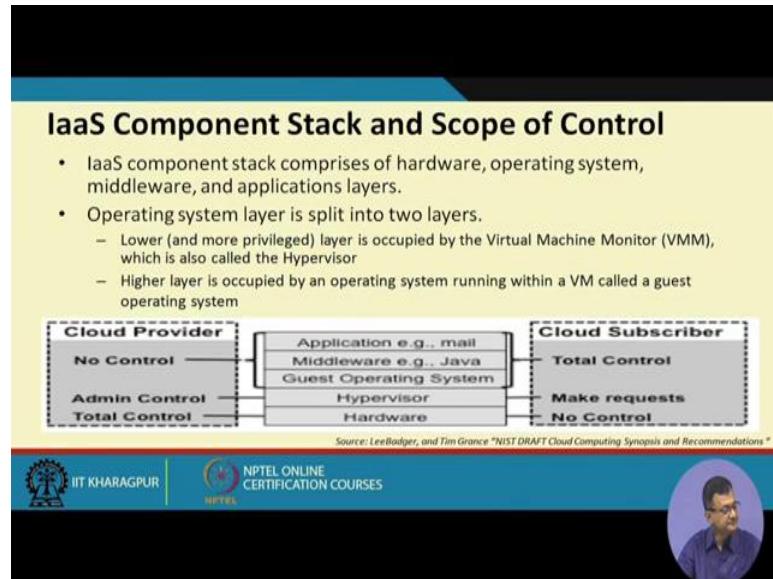
Source: Lee Badger, and Tim Grace "NIST DRAFT Cloud Computing Synopsis and Recommendations"

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, providers theoretically have a large pool of virtual machines; that means, if I am a cloud service provider, I say that these are my flavors of virtual machines which are I am having at my disposal, right. And then I say that if you want a set of virtual machine, then I go on provisioning it right. And usually these are into different categories. Though theoretically, I can configure any of the virtual machine, but in practical sense it is under with specific configuration and etcetera. Like I say that I have VM with very low things, like in case of our 'Megamala', IIT Kharagpur. We have a virtual three type of virtual machines, one is IIT KGP VMs, which has a 4 GB RAM, with 20 GB hard disk space and these are the other processing etcetera. We have a IIT KGP large, which has a 8GB RAM so and so far. Which have a we have a IIT KGP extra large, which has a 60 GB RAM; say these are the three flavor.

Now, based on that based on my backbone resource availability, I have a mix and match of the things not only that, the type of VMs. I will have is also dependent on type of request, I get on this VMs, right. If I have a heavy request on only the smaller type of VM then, I want to provision that smaller type of VM much more than the higher things, So and so far. So, these are the considerations which come into play right. So, like in this case Client A has access to VM1 to VM2, right. Whereas, Client B has access to VM3, where the client C has access to VM4 to VM6. Where the providers retain VM7 through VMN for its other users. So, that can be that can be a modeled and go on doing that and these are typically done like that.

(Refer Slide Time: 05:34)



So, if you look at the IaaS component stack and scope of control. IaaS components stack comprises of hardware operating system, middleware and applications layer. So, these are the typical thing. So, operating system is built into 2 layers, lower the most privilege layer. So, if we look at the more code to the operating system is occupied by the virtual machine monitor or VMM which is also called the hypervisor, right. Higher layer is occupied by the operating system running within a VM called a Guest operating system.

So, the as we have seen that we have a like in this case, In the middle, if you see that I have the bare metal hardware, where the cloud providers have a total control; cloud subscriber has practically low control. Over that, we have a hypervisor right or VMM or virtual machine monitor. Where the administrative control of the cloud provider are cloud subscriber can basically request a request for VM etcetera to this hypervisor. And then we have, other layers like Guest OS, middleware like java etcetera. Application like mail and CRM and other type of things, where the subscriber has the total control in case of a IaaS. Where the provider does not have any control or does not provider does not control it for the thing right. So, this is the way it goes on. So, if we look at it the type of control at the upper layers are moved to the subscriber and type of control at the lower layers are moved to the provider side.

(Refer Slide Time: 07:20)

IaaS Component Stack and Scope of Control

- In IaaS Cloud provider maintains total control over the physical hardware and administrative control over the hypervisor layer
- Subscriber controls the Guest OS, Middleware and Applications layers.
- Subscriber is free (using the provider's utilities) to load any supported operating system software desired into the VM.
- Subscriber typically maintains complete control over the operation of the guest operating system in each VM.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, in IaaS cloud provider maintains total control over the physical hardware and administrative control over the hypervisor layer. Subscriber control the Guest OS, middleware and the application layers. Which are which over the subscriber. Subscriber is free using provider's utility to load any support operating system things. So, if the provider provides the support, the subscriber can load any Guest OS. Subscriber typically maintains complete control over the operation of the Guest operating system in each VM. So, though the subscriber once it loads it has the total control over the Guest OS, like if the cloud provider is providing me some Linux OS and hypervisor and allows me to load Linux or windows, etcetera. So, if I load some flavor of Linux, then I have a total control over that particular thing.

(Refer Slide Time: 08:17)

IaaS Component Stack and Scope of Control

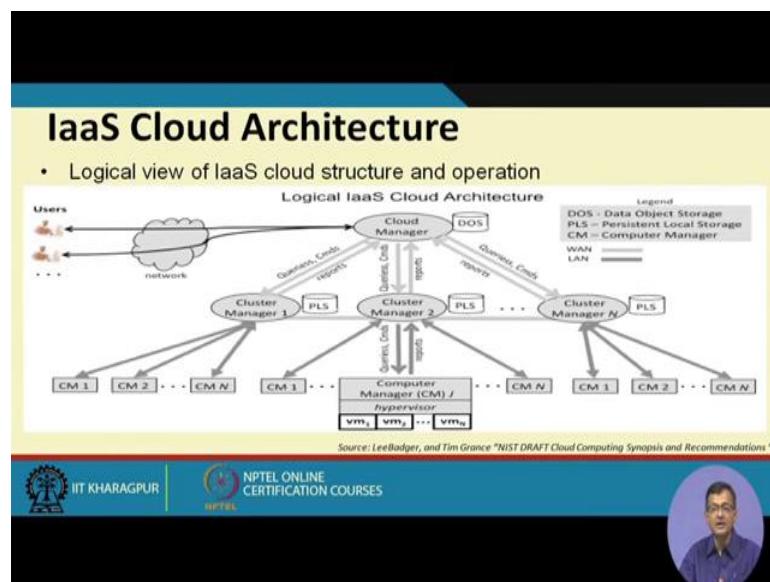
- A hypervisor uses the hardware to synthesize one or more Virtual Machines (VMs); each VM is "an efficient, isolated duplicate of a real machine".
- Subscriber rents access to a VM, the VM appears to the subscriber as actual computer hardware that can be administered (e.g., powered on/off, peripherals configured) via commands sent over a network to the provider.

 IIT KHARAGPUR |  NPTEL ONLINE CERTIFICATION COURSES



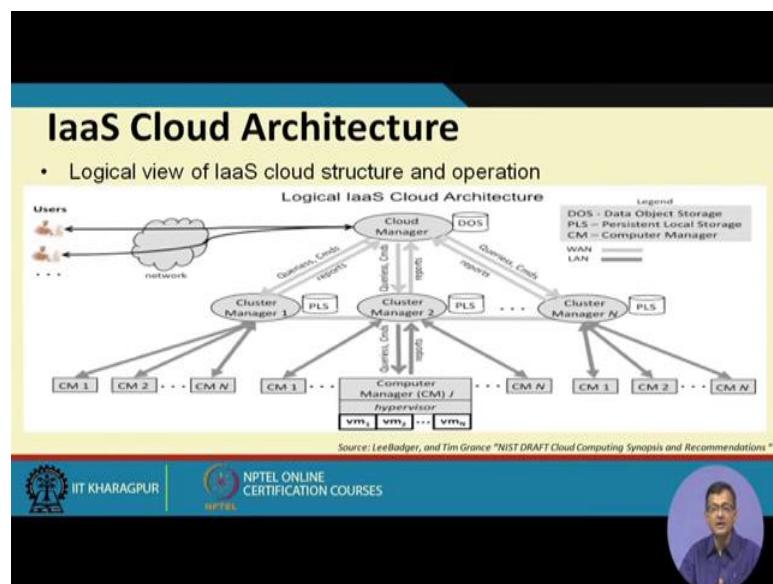
IaaS components stack and other thing, the hypervisor uses hardware to synthesis one or more virtual machines or VMs. Is an efficient isolated duplicate real machine, subscriber rents access to VM, the VM appears to be subscriber as actual computer hardware that can be administered, that is powered on off, peripherals configures via commands send over the network, etcetera. So, that it because what I am having as a subscriber is primarily interface with the provider and I can go on communicating with that.

(Refer Slide Time: 08:49)



So, typically if you look at a typical architecture. So, there are several components over at the top is the cloud manager. There are different cluster managers, which are at the second level of the things and below that is the computer manager or CM, right. And the CM is basically at the; has this hypervisors and different type of virtual machine immolated over the things, right. So, we have at the cloud manager and the data object storage, which is the muster data base of keeping track of the things. Then at the lower level we have a persistent local storage like so; that means, the storage which does not go have a when the cloud provider is not there. When the VMs are not used or shut down and type of things.

(Refer Slide Time: 09:43)



So, if we look at that 3 level hierarchies of the components, have 1 is the top level is responsible for the central control, like the cloud manager. The middle level is responsible for management of possibly large computer cluster and may be geographically distance from one another. So, if you look at this cluster manager, it is managing this large clusters, which maybe geographically spread and the third is the bottom is responsible for running the host computing system on which virtual machines are created. So, these are running this host computing system, where the VMs are created, right. So, this is typically 3 layer of control subscriber quarries and comment generally flow into the system. At the top are forwarded down through the layers which either answer quarries or execute commands.

(Refer Slide Time: 10:30)

IaaS Cloud Architecture

- Cluster Manager can be geographically distributed
- Within a cluster manger computer manger is connected via high speed network.

IT Kharagpur | NPTEL ONLINE CERTIFICATION COURSES



So, IaaS cloud architecture cloud manager can be geographically distributed. Within a cluster manager, computers manager is connected by a high speed network. So, this is this if you look at the cluster manager, these are all connected through a high speed network.

(Refer Slide Time: 10:50)

Operation of the Cloud Manager

- Cloud Manager is the public access point to the cloud where subscribers sign up for accounts, manage the resources they rent from the cloud, and access data stored in the cloud.
- Cloud Manager has mechanism for:
 - Authenticating subscribers
 - Generating or validating access credentials that subscriber uses when communicating with VMs.
 - Top-level resource management.
- For a subscriber's request cloud manager determines if the cloud has enough free resources to satisfy the request

IT Kharagpur | NPTEL ONLINE CERTIFICATION COURSES

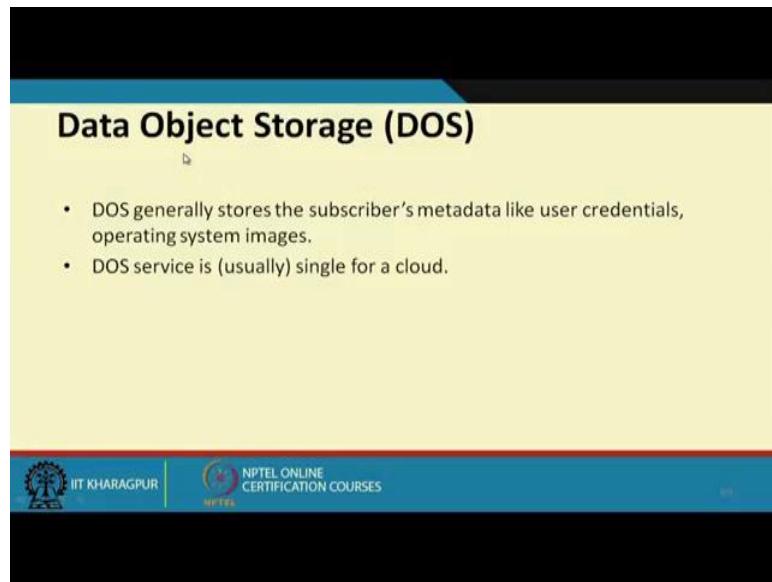


So, now if you look at the quickly that operation of the different things; so, what we have cloud manager at the top, then the cluster manager then the computer manager which manages the individual things right. So, if you look at the cloud manager duty, is the

public access point of the cloud that is the first thing. So, whenever somebody access the cloud in the public, where the subscriber sign in for accounts manage and resources they rent from the cloud etcetera.

Cloud manager has the mechanism for authenticating subscriber. Whether the authentication mechanisms, generating or validating access credential with the subscriber uses when the communication with the VMs like when it is basically the frontend of whole system, top level resource management. So, for subscriber request cloud manager determines, if the cloud has enough free resource or to satisfy the thing. So, subscriber resource on the things. So, cloud manager maintains metadata information. So, if I request for a particular VM or a set of VMs which are not available on the cloud, then it becomes the cloud manager has to take a call. So, it is manages at the subscriber level.

(Refer Slide Time: 12:01)



There is a thing called data object storage or DOS. DOS generally store the subscriber's metadata as we are talking about the user credential operating system etcetera. DOS service usually single for a cloud.

So, what we say that the particular DOS service is for a particular cloud. So, it maintains a; what we say registry or a cataloging or a metadata service for that whole cloud. So, it is the binding block of the whole thing, that what are the services available whether the free resources available etcetera are all updated in those.

(Refer Slide Time: 12:35)

Operation of the Cluster Managers

- Each *Cluster Manager* is responsible for the operation of a collection of computers that are connected via high speed local area networks
- *Cluster Manager* receives resource allocation commands and queries from the *Cloud Manager*, and calculates whether part or all of a command can be satisfied using the resources of the computers in the cluster.
- *Cluster Manager* queries the *Computer Managers* for the computers in the cluster to determine resource availability, and returns messages to the *Cloud Manager*

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, if we look at the cluster manager that this middle level. So, it is responsible for operation of collection of computers that are connected via high speed local area network, right. So, that is the; it manages the lower level computers, cluster manager receive resources allocation commands and queries from the cloud manager at the top and calculate whether the part or all of a command can be satisfied using the resource of computers in the cluster. That means, when the cluster manager gets a request, it checks that whether the; it is available resource within the cluster. Whether it is able to satisfy the things and accordingly it says appropriate signal to the things. Like so, that it can accept or cannot accept and type of things.

Cluster manager queries the computer manager for the computers in the cluster to determine resource availability, returns messages to the cloud manager. That mean, it becomes a middleware or a agent between the cluster manager and the cloud manager. If you look at the operation, there it get the direction for the cloud manager and then instructs the computer managers to perform resource allocation, reconfiguration, de allocation of resources and etcetera. Is cluster managers connected to a persistent local storage; that means, it is the. So, called coat uncoated non volatile storage. When the particular virtual machine is shutdown, deprivation or some issues come up. So, that it has a local storage into the things.

So, this PLS provide persistent disk like storage to the virtual machine. So, it is as if persistent disk like the same things are there. So, the data when you next time log in it will be available with you.

(Refer Slide Time: 14:25)

The slide has a yellow header bar with the title 'Operation of the Computer Managers'. Below the title is a bulleted list of four items. At the bottom of the slide are three logos: IIT Kharagpur, NPTEL Online Certification Courses, and a portrait of a man.

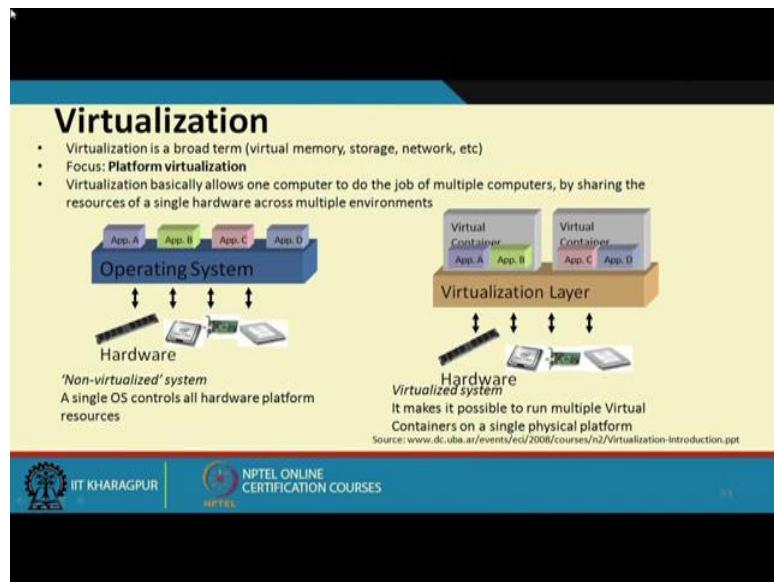
Operation of the Computer Managers

- At the lowest level in the hierarchy computer manager runs on each computer system and uses the concept of virtualization to provide Virtual Machines to subscribers
- Computer Manager maintains status information including how many virtual machines are running and how many can still be started
- Computer Manager uses the command interface of its hypervisor to start, stop, suspend, and reconfigure virtual machines

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | NPTEL

So, look at the; if we look at the operation of the cloud manager. At the lowest level in the hierarchy cloud manager runs on each computer system uses the concept of virtualization to provide virtual machine to the subscriber. So, cloud manager is basically providing the VMs. So, it is the to the subscriber this providing VMs it is responsible for that, computer manager maintains status information including how many virtual machines are running, etcetera. So, it maintains the status information, as of now how many virtual machines are running. Also cloud manager uses command interface to hypervisor to start, suspend and reconfigure virtual machines.

(Refer Slide Time: 15:12)



So, it also same if there is a need to the computer manager. So, all this what we see is, virtualization place important role, right. So, how this virtualization is made and what are the aspects etc: thing we will try to see quickly in the subsequent slides. So, virtualization is a broad term, it can be a virtual memory, it can be virtual network, it can be virtual storage, etcetera. So, anything which can be virtualized is a virtualization aspect. Our primary focus is the platform for virtualization, right. Virtualization basically allows, one computer to do the job of multiple computers, right, by sharing resources of a single hardware across multiple environment.

So, that is important so; that means, I have a bare metal then, I create different machines or virtual machines which can cater to the things. But at the back end, my same bare metal is running. Now this becomes a very tricky. So, because suppose you are running if you are having the your bare metal on particular environment and then you are running one machine on windows, one machine Guest OS on Linux, one machine or some other things etcetera. Then there is a problem of issue of how this say instruction set of this Guest OS will be running on the hardware at the end. So, that there are issues of like that, there are issues of application sizing there are issues when the VM says that, I require more resources or I want to release other things, how it manage So and So forth.

So, if we look at the hardware. So, if it is a no virtualizes system a single OS controls the all hardware platform resources. If it is a virtualization system it takes the possibly, it

makes it possible to run multiple virtual containers on a single physical platforms. So, I can have multiple virtual containers, which can run on or which can be plugged into a single physical infrastructure.

(Refer Slide Time: 17:14)

Virtualization

- Virtualization is way to run **multiple operating systems** and **user applications** on the same hardware
 - E.g., run both Windows and Linux on the same laptop
- How is it different from **dual-boot**?
 - Both OSes run **simultaneously**
 - The OSes are completely **isolated** from each other

So, we are somewhat experienced with this virtualization. So, virtualization is a way to run multiple operating system and user application on the same hardware. So, that is virtualizes. So, I can I have 2 different operating system on the same; that means, I say I virtualized one this OS out of this, right.

So, run both windows Linux on the left off. So, how is it different for dual boot also? We are doing that once windows etcetera. Here in case of virtualization this different OS or all the OSs are running together right or simultaneously. The OSs are completely isolated from each other, right. They are completely isolated from each other in case of a true virtualization.

(Refer Slide Time: 18:02)

Hypervisor or Virtual Machine Monitor

Research Paper Popek and Goldberg, "Formal requirements for virtualizable third generation architectures", CACM 1974 (<http://portal.acm.org/citation.cfm?doid=361011.361073>).

A **hypervisor** or **virtual machine monitor** runs the guest OS directly on the CPU. (This only works if the guest OS uses the same instruction set as the host OS.) Since the guest OS is running in user mode, privileged instructions must be intercepted or replaced. This further imposes restrictions on the instruction set for the CPU, as observed in a now-famous paper by Popek and Goldberg identify three goals for a virtual machine architecture:

- *Equivalence*: The VM should be indistinguishable from the underlying hardware.
- *Resource control*: The VM should be in complete control of any virtualized resources.
- *Efficiency*: Most VM instructions should be executed directly on the underlying CPU without involving the hypervisor.

Source: www.dc.ubo.ar/events/ecu2008/courses/n2/Virtualization-Introduction.ppt

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | NPTEL

Now, another important aspect of things which come into play is the virtual machine monitor or hypervisor. So, a hypervisor or a virtual machine monitor runs the Guest OS directly on the CPU right. so that means, I have a Guest OS and at the as my which has been immolated or installed by the subscriber or running on the or it is application on the of the client. And then the hypervisor is responsible to execute this or run this Guest OS directly on the hardware, at the back and the things.

So, this works if the Guest OS uses the same instruction set of the host OS right. If it is a different instruction set, then there is a instruction translation should come into play right. So, there are several issues, as depicted by Popek and Goldberg, that 3 goals of virtual machine architecture: are there, 1 is that equivalence the VM, should be indistinguishable from the underlying hardware. So, as if the virtual machine is running on the hardware itself. So, what we say it is equivalence or resource control, the VM should be incomplete control of the virtualizes resource.

So, if you give me 4 GB machine, 30 GB or 60 GB hard disk so and so forth that I should have to complete control over the thing right as subscriber. And efficiency, most VM instruction should be executed directly on the underlying CPU without involving the hypervisor. So, that is another thing. So, efficiency will increase if the most of the virtual machine instruction, should able to execute directly on the underlying CPU without intervention of this other involvement the hypervisor. So, these are the aspects which

need to be looked into, when we have virtual machine monitor or hypervisor in place which allows me to emulate VMs.

(Refer Slide Time: 20:04)

Hypervisor or Virtual Machine Monitor

Popek and Goldberg describe (and give a formal proof of) the requirements for the CPU's instruction set to allow these properties. The main idea here is to classify instructions into

- **privileged** instructions, which cause a trap if executed in user mode, and
- **sensitive** instructions, which change the underlying resources (e.g. doing I/O or changing the page tables) or observe information that indicates the current privilege level (thus exposing the fact that the guest OS is not running on the bare hardware).
- The former class of sensitive instructions are called **control sensitive** and the latter **behavior sensitive** in the paper, but the distinction is not particularly important.

What Popek and Goldberg show is that we can only *run a virtual machine with all three desired properties if the sensitive instructions are a subset of the privileged instructions*. If this is the case, then we can run most instructions directly, and any sensitive instructions trap to the hypervisor which can then emulate them (hopefully without much slowdown).

Source: www dc uba ar/events/ecl/2008/courses/n2/Virtualization-Introduction.ppt

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

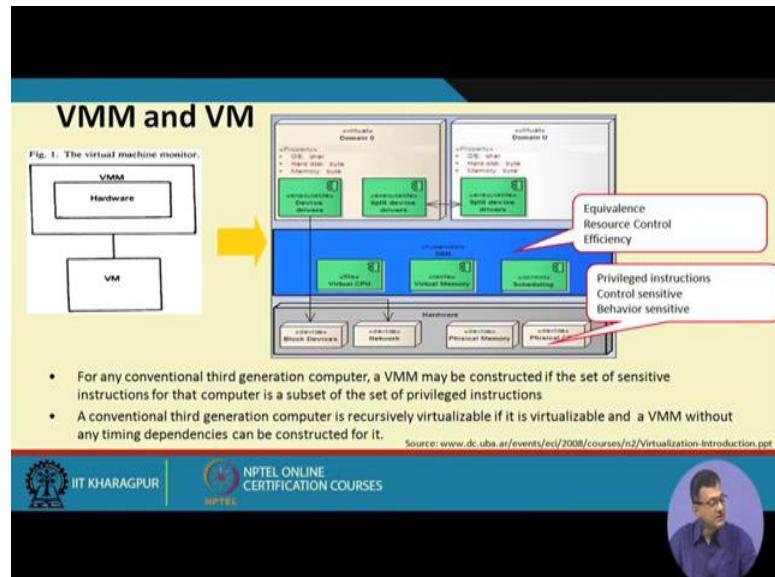
So, again in the same work these Popek and Goldberg describe and also give a provided a formal proof of the requirement of the CPUs instruction set to allow these properties to happen. Like to allow truly virtualization happen at the IaaS level, the main idea is classifying this instruction into 3 different thing. One is privileged instruction, which cause a trap in the executed in the user mode. There is a sensitive instruction, which change the underlining resources, that is doing IO or changing the fact that the Guest OS is running on the bare metal etc; a sorry means first of all changing the IO or the page tables or observe the information indication on the current privilege level.

So, if you remember your basic architecture things. So, I have different level of operations like level 0 level 1 so and so forth. So, more the lower the level, as I go more on the more closer to the bare metal. So, it all depends that your virtualization at which point it operates. More the higher level then more latency will come into play, more translation will come into play. So, we need to look into the things that what level operations I need to do; especially when we do IaaS type of operations.

And the former class of sensitivity instructions also called controls sensitive or latter call are behavior sensitive operations right. So, what this Popek and Goldberg show, that if we can run a virtual machine with all the 3 desired property if the sensitive instruction

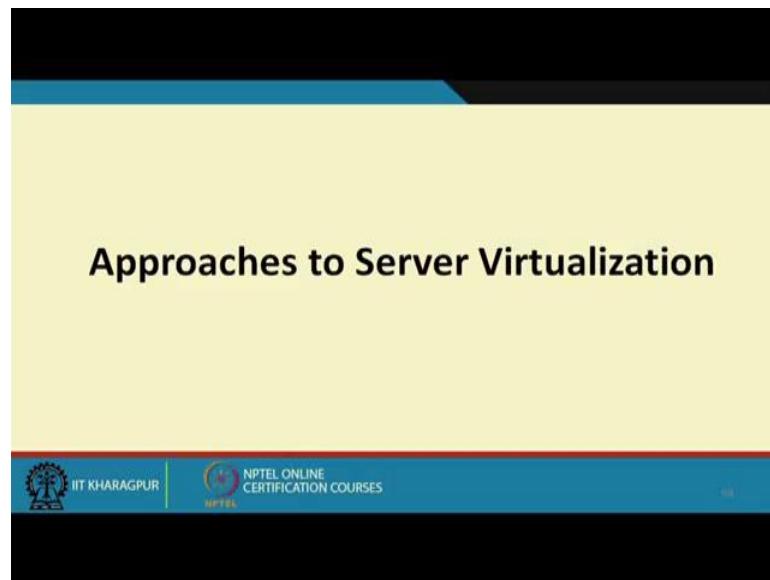
are a subset of a privilege instruction. In their work they have shown that, these sorts of things are true and we can realize a true virtual machine.

(Refer Slide Time: 21:57)



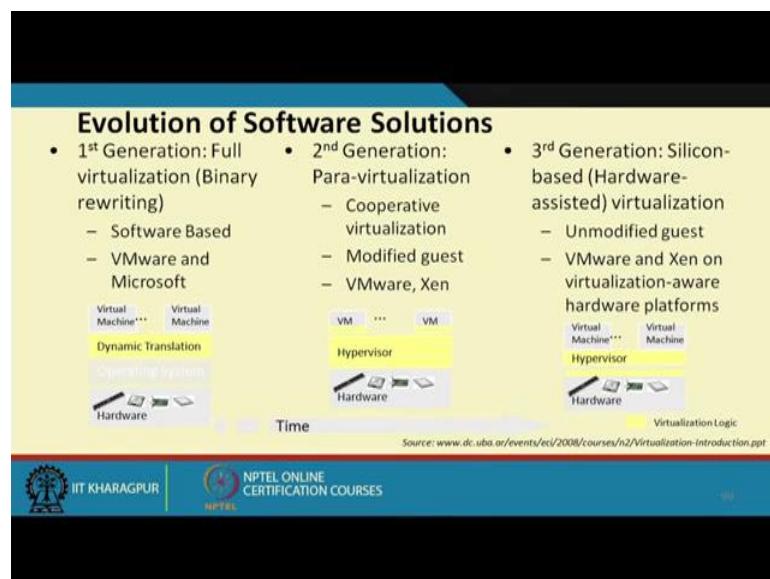
So, the same thing or thus whatever we are discussing. So, we have this VMs and this hardware over this virtual machine monitors is running and it is allows us to emulate this different VMs right. So, it can be yours. So, some where other whatever the Guest OS running on the VM. The instruction set need to be converted which can be understood by this hardware, right. So, that is our bottom line. So, whether it is whether the whole VMM will have a total abstraction or also have some sort of overlapping with the things, that that depends on whatever the implementation.

(Refer Slide Time: 22:39)



So, if we look at the approaches typical approaches to virtualization.

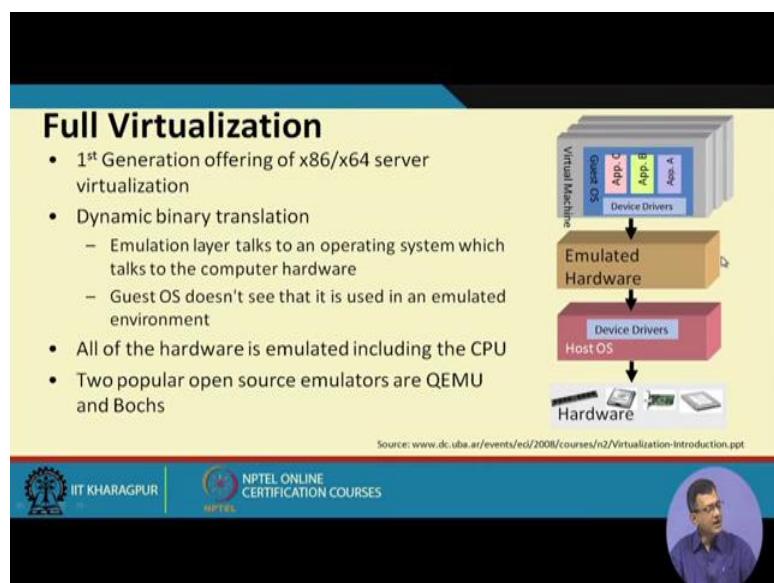
(Refer Slide Time: 22:42)



So, if you see the evolution of this virtualization, one is what we say first generation or full virtualization so, binary rewriting. So, it is software base, VMware and Microsoft supports this. So, whatever the virtual machines are generating a dynamic translator, rewrites to the underlining hardware So and So forth. In case of a second generation: virtualization or Para virtualization cooperative virtualization. So, it is a cooperative virtualization the Guest OS or the modified guest and VMware and xen does it. So, it

basically if you look at here. So, it penetrates little bit into the VM; that means, that is modified on the things. In case of a third generation or what we say silicon based, there is hardware assisted virtualization, unmodified Guest, VMware and xen on the virtualization aware hardware platform. So, now your hardware platform is a virtualization aware. So, it is hardware assisted virtualization. So, in these cases these hardware platforms are not aware of this virtualization. Whatever is done by this your hypervisor?

(Refer Slide Time: 24:00)



So, there are different things in case of full virtualization, first generation offering. So, dynamic binary translation of the source code, as we have seen all the hardware is emulated including the CPU. To popular open source emulator, many of you might have seen QEMU and Bochs. These are the two popular emulator which change that open source.

(Refer Slide Time: 24:27)

Full Virtualization - Advantages

- Emulation layer
 - Isolates VMs from the host OS and from each other
 - Controls individual VM access to system resources, preventing an unstable VM from impacting system performance
- Total VM portability
 - By emulating a consistent set of system hardware, VMs have the ability to transparently move between hosts with dissimilar hardware without any problems
 - It is possible to run an operating system that was developed for another architecture on your own architecture
 - A VM running on a Dell server can be relocated to a Hewlett-Packard server

Source: www.dcsa.ar/events/eci/2008/courses/n2/Virtualization-Introduction.ppt



Full virtualization, there are advantages of emulation layer, like you do not bother about that two things are separated. Total VM portability, like I can put take the VM from one to another. So, put total portability, because only the VMM, I need to understand this. There are drawbacks this hardware emulations comes with a performance price.

(Refer Slide Time: 24:42)

Full Virtualization - Drawbacks

- Hardware emulation comes with a performance price
- In traditional x86 architectures, OS kernels expect to run privileged code in Ring 0
 - However, because Ring 0 is controlled by the host OS, VMs are forced to execute at Ring 1/3, which requires the VMM to trap and emulate instructions
- Due to these performance limitations, para-virtualization and hardware-assisted virtualization were developed



Traditional x86 Architecture

Full Virtualization

Source: www.dcsa.ar/events/eci/2008/courses/n2/Virtualization-Introduction.ppt

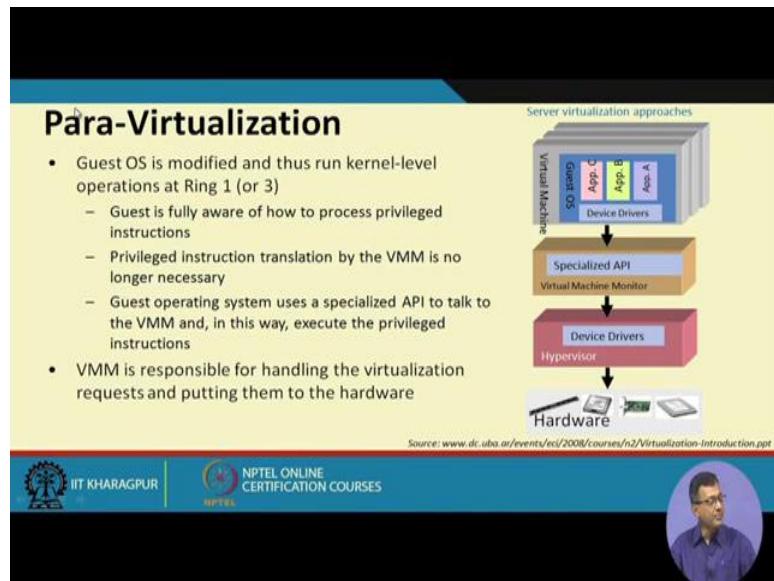


Whenever you want to emulate a hardware performance price in traditional x86 architecture, OS kernels expect to run on the privilege ring 0. And now you want to run

them at a higher level and that need to be, there should be more latency and other things come into play.

So, we need to pay for that, or what we need to there is a performance price for that thing.

(Refer Slide Time: 25:07)



In case of Para virtualization, Guest OS is modified and thus run kernel level operating system at ring 1 or ring 3. That is much higher level Guest is fully aware of how to process privileged instruction. Privilege instruction translated to VMM is no longer necessary. Guest operating systems uses specialized API to talk to the VMM right. So, it is Para; that means, your VMM is penetrating to use some sort of a some into your VMs. So, that now it is not full translation, but I have a Para virtualization the performance increases and this thing is becoming pretty popular.

(Refer Slide Time: 25:50)

Para-Virtualization

- Today, VM guest operating systems are para-virtualized using two different approaches:
- Recompiling the OS kernel**
 - Para-virtualization drivers and APIs must reside in the guest operating system kernel
 - You do need a modified operating system that includes this specific API, requiring a compiling operating systems to be virtualization aware
 - Some vendors (such as Novell) have embraced para-virtualization and have provided para-virtualized OS builds, while other vendors (such as Microsoft) have not
- Installing para-virtualized drivers**
 - In some operating systems it is not possible to use complete para-virtualization, as it requires a specialized version of the operating system
 - To ensure good performance in such environments, para-virtualization can be applied for individual devices
 - For example, the instructions generated by network boards or graphical interface cards can be modified before they leave the virtualized machine by using para-virtualized drivers

Source: www dc ubo ar/events/eci/2008/courses/n2/Virtualization-Introduction.ppt

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, in case of Para virtualization recompiling of the OS kernel is required because, there is penetrating into the things installing of Para virtualized drivers are required. So, that it takes care at your Guest OS level.

(Refer Slide Time: 26:03)

Hardware-assisted virtualization

- Guest OS runs at ring 0
- VMM uses processor extensions (such as Intel® VT or AMD-V) to intercept and emulate privileged operations in the guest
- Hardware-assisted virtualization removes many of the problems that make writing a VMM a challenge
- VMM runs in a more privileged ring than 0, a Virtual-1 ring is created

Diagram illustrating Server virtualization approaches:

```
graph TD; VM[Virtual Machine] --> SA[Specialized API]; VM --> VMM[Virtual Machine Monitor]; VM --> DD[Device Drivers]; SA --> VMM; VMM --> DD; VMM --> Hypervisor[Hypervisor]; Hypervisor --> HW[Hardware]
```

Source: www dc ubo ar/events/eci/2008/courses/n2/Virtualization-Introduction.ppt

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

The third one is a hardware OS running on a ring 0. VMM uses processor extensions like in case of Intel VT or AMD V and type of things to interpret the privileged things. Hardware assisted virtualization removes many problems that take place in writing a VMM into a, rather it is a big challenge. So, that it is hardware assistant so; that means,

you go directly over thing is totally fully complied. In doing so, you lose that portability of the VMs right it becomes a more tricky issue because you need to migrate to something which understand that hardware. Pros it allows run unmodified OS, right you the OSs need to not be modified; cons is the speed and flexibility is issues that when that your flexibility you lose and the speed it depends when the migration, etcetera comes into play.

(Refer Slide Time: 26:44)

The slide has a dark blue header bar with the text "Server virtualization approaches" in white. Below this is a yellow main content area with a black sidebar on the left. The title "Hardware-assisted virtualization" is in bold black font at the top of the yellow area. To its right is a small blue progress bar. Below the title is a bulleted list under two categories: "Pros" and "Cons".

- Pros
 - It allows to run unmodified OSs (so legacy OS can be run without problems)
- Cons
 - Speed and Flexibility
 - An unmodified OS does not know it is running in a virtualized environment and so, it can't take advantage of any of the virtualization features
 - It can be resolved using para-virtualization partially

Source: www dc uba ar/events/eci/2008/courses/n2/Virtualization-Introduction ppt

At the bottom of the slide, there is a footer bar with the IIT Kharagpur logo, the text "IIT KHARAGPUR", the NPTEL logo, and the text "NPTEL ONLINE CERTIFICATION COURSES". On the right side of the footer bar is a circular profile picture of a man.

(Refer Slide Time: 27:00)

The slide has a dark blue header bar with the text "Server virtualization approaches" in white. Below this is a yellow main content area with a black sidebar on the left. The title "Network Virtualization" is in bold black font at the top of the yellow area. To its right is a small blue progress bar. Below the title is a block of italicized text defining network virtualization.

Network virtualization is a networking environment that allows multiple service providers to dynamically compose multiple heterogeneous virtual networks that co-exist together in isolation from each other, and to deploy customized end-to-end services on-the-fly as well as manage them on those virtual networks for the end-users by effectively sharing and utilizing underlying network resources leased from multiple infrastructure providers.

At the bottom of the slide, there is a footer bar with the IIT Kharagpur logo, the text "IIT KHARAGPUR", the NPTEL logo, and the text "NPTEL ONLINE CERTIFICATION COURSES". On the right side of the footer bar is a circular profile picture of a man.

So, with that we will just touch upon, some of the more aspects of the things what we say network level virtualization.

So, that now what we are looking as the hardware, now we are looking at the network level; that means, I need to emulate a network over a virtual network over a given network, right.

(Refer Slide Time: 27:21)

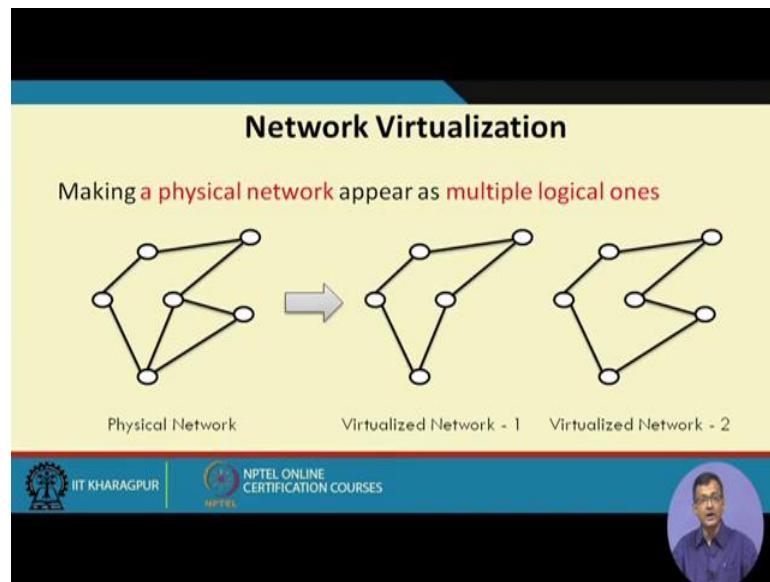
The slide has a yellow background with a black header bar. The title 'Typical Approach' is centered in the yellow area. Below the title is a bulleted list of components:

- Networking technology
 - IP, ATM
- Layer of virtualization
- Architectural domain
 - Network resource management, Spawning networks
- Level of virtualization
 - Node virtualization, Full virtualization

At the bottom of the slide, there is a footer bar with the IIT Kharagpur logo, the text 'IIT KHARAGPUR', the NPTEL logo, and the text 'NPTEL ONLINE CERTIFICATION COURSES'. To the right of the footer is a circular portrait of a man.

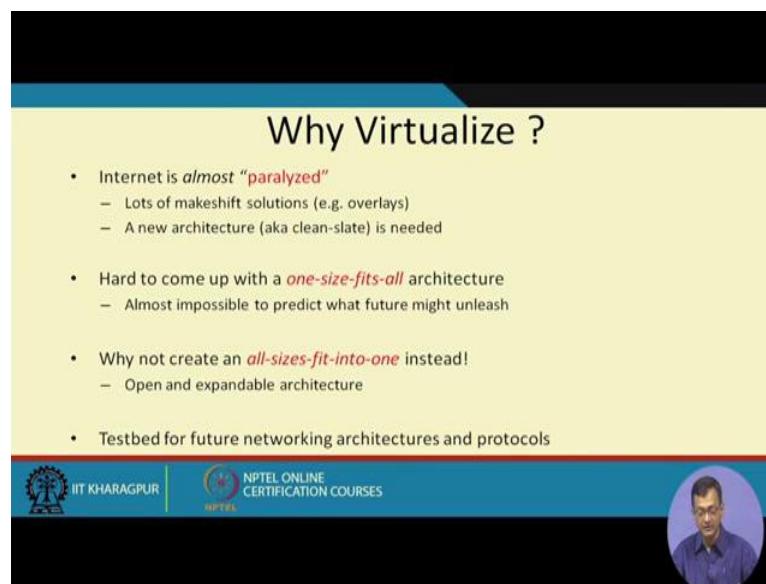
So, typical approaches what we have seen is that in case of a network technology that is whether IP base or ATM base layer. Which layer of virtualization is there are architectural domain, like network resource managements, spawning network and type of things, level of virtualization which is node level or full virtualization of the things like.

(Refer Slide Time: 27:38)



If you look at suppose I have a network like this. Then I can emulate a virtual network like this or I can emulate virtualizes network like this, right. So, this is my base network. So, it I can have 2 type of virtual network on the thing over and above of this thing.

(Refer Slide Time: 27:56)



So, I emulate a virtual network, why to virtualizes now what people say that not only internet in your own networking paradigm is becoming paralyze. Like you do not have much in go space. Hard to come with a 1 size fit all type of architecture in number of cases. Like you have a network configuration and fits everybody. So, why not to create

something that all sizes fit to one instead type of scenario. So, that is the things what we want to do. So, test bed for future and there is that is and any futuristic working on the network, protocols and etcetera. Require a test bed for that putting a whole networking things may be a challenge. So, I can have a sort of virtualizes network.

(Refer Slide Time: 28:43)

Related Concepts

- Virtual Private Networks (VPN)
 - Virtual network connecting distributed sites
 - Not customizable enough
- Active and Programmable Networks
 - Customized network functionalities
 - Programmable interfaces and active codes
- Overlay Networks
 - Application layer virtual networks
 - Not flexible enough

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

And this is not suddenly came into play, we have lot of things which are already in play that is a virtual private network active and programmable networks, overlay networks these are already in place for several years and this is being emulated over that.

(Refer Slide Time: 29:00)

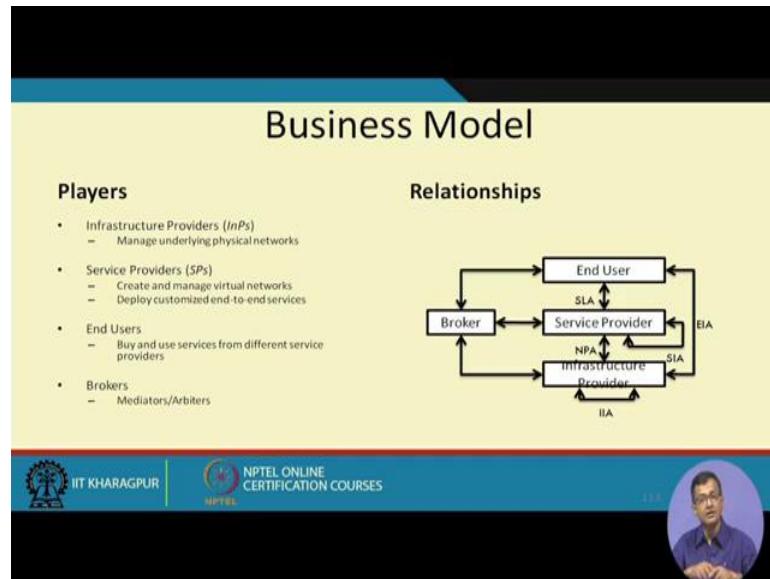
Network Virtualization Model

- Business Model
- Architecture
- Design Principles
- Design Goals

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

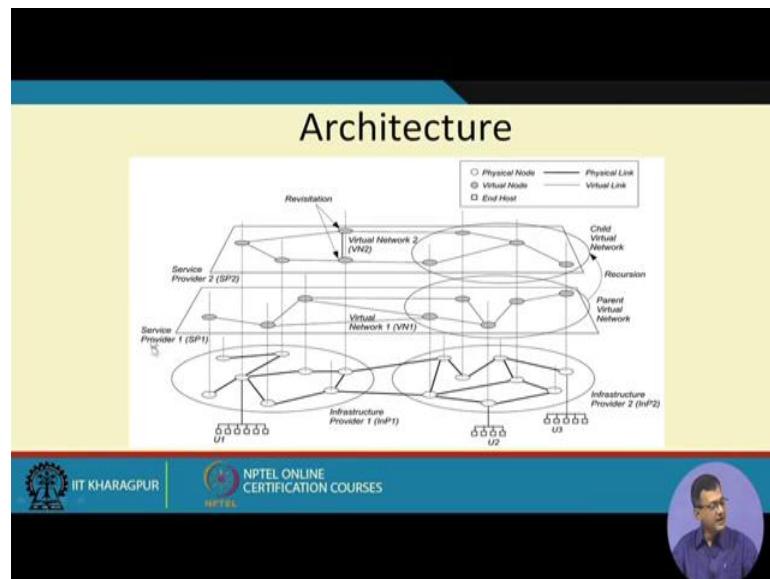
And if you look at there are different type of aspect like business model architecture, design principle, design goal and those are the things.

(Refer Slide Time: 29:09)



And there are different aspects of those, like infrastructure providers, service provider's, end user, brokers. These are the different players and they have a very complex relationship between each other that how things will be there. And this business model is not only true for network virtualization, it is true for other sort of virtualization as well and if we look at that architecture.

(Refer Slide Time: 29:35)



So, if I have that infrastructure provider, then I have a service provider 1 which is have some sort of a virtual network. Emulated from this, like making active some of the networks etcetera and I can have other type of service provider 2, SP2 emulating, other type of networks as well. So, I can emulate different variety of network on a basic underlying network. So, I can realize different networks into the things. So, these are also becoming very popular. So, having say not only having servers like, I want to conduct a particular testing over across the geographical space.

So, not only I want these servers into the place. I also have a particular network infrastructure into play, right. Then those then if you want to have those, type of things then I not only require the virtualization at the IaaS level or the server level virtualization. I also require a network level virtualization into play and making this overall I can have a virtualized IP infrastructure into place right. So, we will continue our discussion in our subsequent lectures we will stop here today.

Thank you.

Cloud Computing
Prof. Soumya Kanti Ghosh
Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur

Lecture – 08
XML Basics

Hello. So, welcome to this course of this cloud computing. We will continue our talk on different aspects of cloud. So, today what we are trying to look at that, what is the underlining technology or underlining protocol which binds together. As we have discussed that there are in cloud, there are various type of services or XaaS, anything, as a service type of model it service model it was. So, we need to have some way or other a protocol which allows that to interoperate between different things.

Now, if you look at whenever a user or a customer or a consumer of cloud, is taking the service of the cloud provider. So, it is independent of where, how this cloud is hosted the provider ends. How the user client is accessing the things. So, it is more of service oriented architecture, what we are basically implement in this cases. So, web service oriented architecture or web services as we have as we know that, allows us to interoperate between loosely coupled heterogeneous services, to achieve some expected output right. So, all this type of service exchanges or service driven architecture the basic one of the basic component or one of the basic building block is XML, right. So, what we will discuss today maybe one or two lectures that the basics of XML.

What I understand that, most of you are used to XML or you know what is XML. But for making it little those who are not accustomed or making that looking at all sorts of viewers and listeners of this particular lecture series we want, we will discuss a very basics of XML and try to see what are its properties and how it allows us to interoperate, right.

(Refer Slide Time: 03:02)

XML ??

- Over time, the acronym "XML" has evolved to imply a growing family of software tools/XML standards/ideas around
 - How XML data can be represented and processed
 - application frameworks (tools, dialects) based on XML
- Most "popular" XML discussion refers to this latter meaning
- We'll talk about both.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, if we look at XML. So, it is acronym XML as evolved from a employee for a growing family of software tool XML standards and ideas. So, what we will try to see that, what are the different properties of XML? How it is, what is how it allows that to interoperate. What are the different types of parser or what are the type of other technologies, which are there in the XML.

(Refer Slide Time: 03:27)

What is XML?

- **A syntax** for "encoding" text-based data (words, phrases, numbers, ...)
- **A text-based syntax.** XML is written using **printable Unicode** characters (no explicit binary data; character encoding issues)
- **Extensible.** XML lets you define your own **elements** (essentially **data types**), within the constraints of the syntax rules
- **Universal format.** The syntax rules ensure that all XML processing software **MUST** identically handle a given piece of XML data.

If you can read and process it, so can anybody else

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, it is extensible markup language right. We are used to HTML, hypertext markup language, which are primarily used in this world of internet, to visualize document, right.

So, XML is a extensible markup language and unlike HTML which is more of a data display or data representation language. Your information representation language XML is more of a data transformation language, right.

So, it allows us, we will see your; it allows us to achieve interoperations, right. So, the syntax for encoding is text based. So, words phrase numbers. So, it is a readable, a text based syntax and XML is when well written using printable Unicode character right, no explicit binary data; character encoding issues etc right. It is extensible; XML let is you define your own elements. Like data types, like unlike html those who are used to HTML or though as have seen the HTML, we have a predefined set of tags right.

Whereas in XML you can defined your own elements or own tags. So, it is within a constant of syntax rule definitely you can define within a context of syntax. So, it is also sometimes referred to a universal format right. A syntax rule ensures that all XML processing software must identically handle given piece of XML data. So, it is universal. So, if you have a XML data, and XML processing software, or say XML parser, it will be able to handle all sort of XML universal right. So, that is the typical basic characteristics of XML, which makes it ubiquitous.

So, if you can read and process it; so, anybody else, this is the means basic; so, if you if at one and if you can read and process it so that everybody else can read and process the same data.

(Refer Slide Time: 05:53)

XML Declaration ("this is XML")

Binary encoding used in file

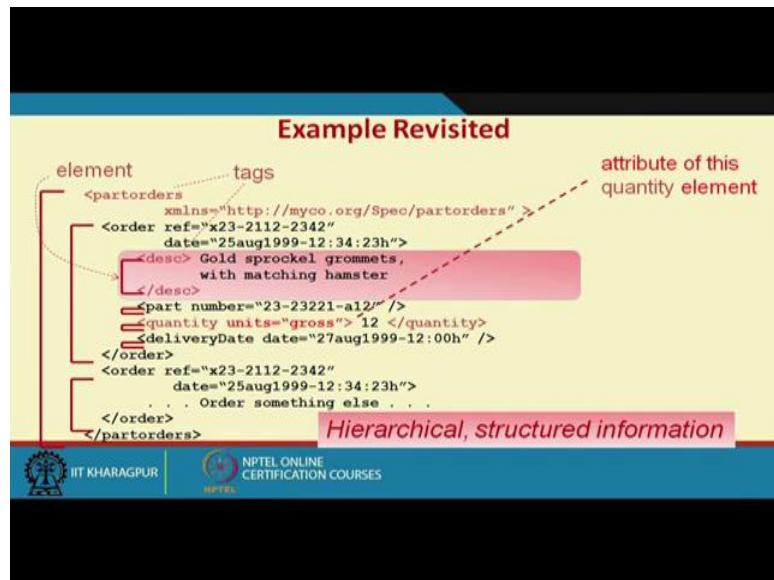
```
<?xml version="1.0" encoding="iso-8859-1"?>
<partorders
  xmlns="http://myco.org/Spec/partorders">
  <order ref="x23-2112-2342"
    date="25aug1999-12:34:23h">
    <desc> Gold sprocket grommets,
    with matching hamster
    </desc>
    <part number="23-23221-a12" />
    <quantity units="gross"> 12 </quantity>
    <deliveryDate date="27aug1999-12:00h" />
  </order>
  <order ref="x23-2112-2342"
    date="25aug1999-12:34:23h">
    ... Order something else ...
  </order>
</partorders>
```

What is XML: A Simple Example

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | NPTEL

So, it is if I look at a simple example like we have taken from resource. So, that there are some of the things like there is a XML declaration. Like we says the version and there is a will come slowly we will come to those things, there is a XML name space and there are definition of different elements and type of things here, right.

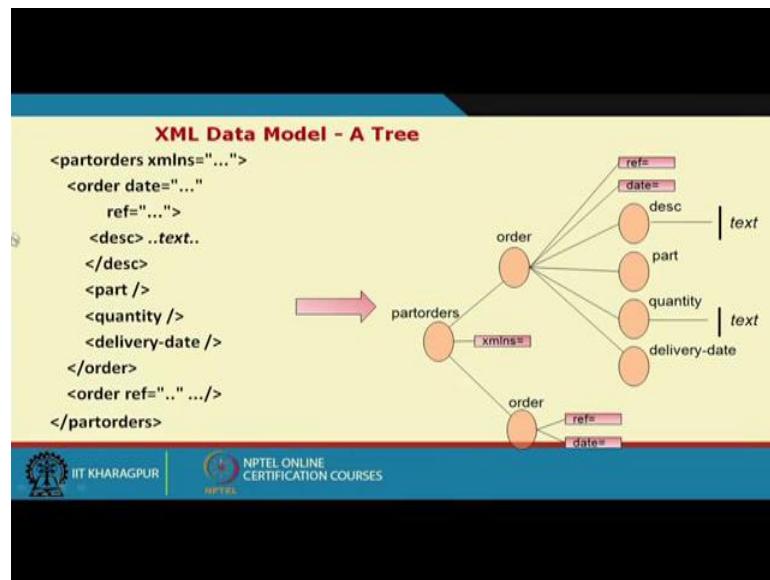
(Refer Slide Time: 06:18)



So, these are different XML tags, if you see this tags are user defined right, tags can be user defined; that means, I can define my own tags, that this part orders or description or part number, these are all user defined tags, right. So, unlike html, where the tags are predefined, you cannot define your own tags here you can defined your own tags. Now if you defined your own tags, the next things which come into play. That there can be you need to say that, how tags how things are defined. Like I say I define a tag called table. Now whether, I am referring to the table, whether this table refers to some furniture classes or group in the category of furniture or the table is basically in the category of word processing or phrase it right.

So, that is important right. I need to know that, if I define a tag or if there are I am using table one from this furniture category and another for the word processing category. Then I have to define, the things or how the other end knows that what I am looking for. So, there is a concept of XML namespace, which primarily allows us to define my own user defined elements and other things. So, if we look at the XML document, it is hierarchical and it is structured information, right.

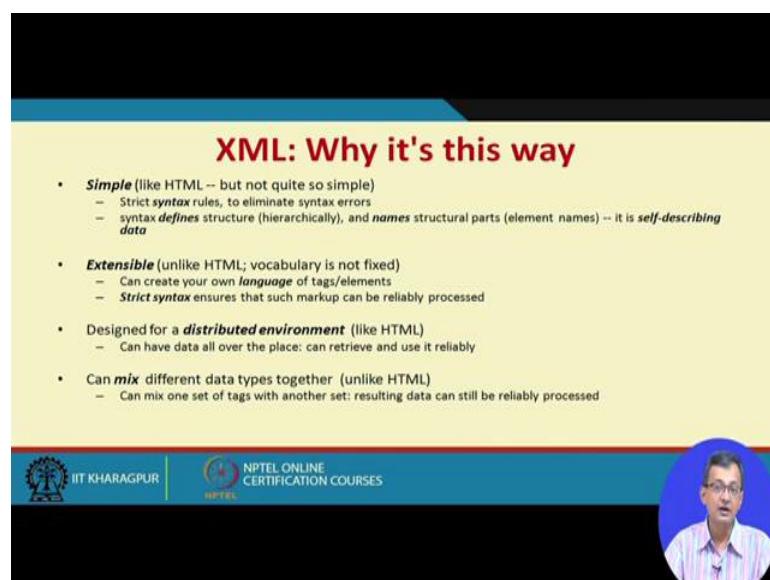
(Refer Slide Time: 07:51)



So, if I look at this document. So, it is a part order all right. It is the top most node which is a XML namespace, that it says that part order is of this namespace and it has different levels or it looks it basically hierarchical and we generate a XML tree, right.

So, XML comes with this property of hierarchical structured information which can be represented by the XML tree.

(Refer Slide Time: 08:32)



So, again if we come back to our basic premise or basic definition, so it is simple like html, but not quite so simple, html is very vanilla type. So, strict syntax rules to eliminate

syntax error. So, XML has a very strict syntax rule. Syntax defines structure hierarchical and name structural element names and it is self describing data. So, it is say the I can have my own data, which is self describing. It is extensible unlike html vocab is not fixed in case of a; it is XML. You can define your own vocab or tags or elements and that means, that you can basically extend this data.

Design for distributed environment right, you are like HTML it is designed for distributed environment. It is say; it is able to talk to or it able to communicate or interface or interoperate with different systems in a or different nodes of a distributed systems. So, can create, your own language of tags and elements sorry it can have data all over the place can retrieve and use it reliably. So that means, it is distributed can mix different data types together unlike HTML. So, html can do it. But in XML, I can have multiple sources and mix them and generate another data out of it.

It is like that, I from the 2 repositories. I take different component of data and finally, generate another data which is a mix of these two things as they HTML, XML allows me to do that. It is extremely helpful in different cases, like you are getting two types of data one form your, say for in academic institution, one data I am getting from academic section, other maybe from all managements centre that regarding students and then I want to club this data and try to find out some things, like I want to find out that who is the best all rounder of this particular semester or particular batch or particular year. So, I need to look at different student activity data, student academic data and other type of data impossible and then generate another data set.

So, this sort of distributed diverse data set on the fly how you can basically interoperate, how you can mix them and generate a thing is this XML is one of the is helpful in this type of cases. So, it is a mix of different data types together.

(Refer Slide Time: 11:44)

The screenshot shows a presentation slide with the title "XML Processing" in red at the top right. The main content is a block of XML code representing bank transfers:

```
<?xml version="1.0" encoding="utf-8" ?>
<transfers>
  <fundsTransfer date="20010923T12:34:34Z">
    <from type="intrabank">
      <amount currency="USD"> 1332.32 </amount>
      <transitID> 3211 </transitID>
      <accountID> 4321332 </accountID>
      <acknowledgeReceipt> yes </acknowledgeReceipt>
    </from>
    <to account="132212412321" />
  </fundsTransfer>
  <fundsTransfer date="20010923T12:35:12Z">
    <from type="internal">
      <amount currency="CDN" >1432.12 </amount>
      <accountID> 543211 </accountID>
      <acknowledgeReceipt> yes </acknowledgeReceipt>
    </from>
    <to account="65123222" />
  </fundsTransfer>
</transfers>
```

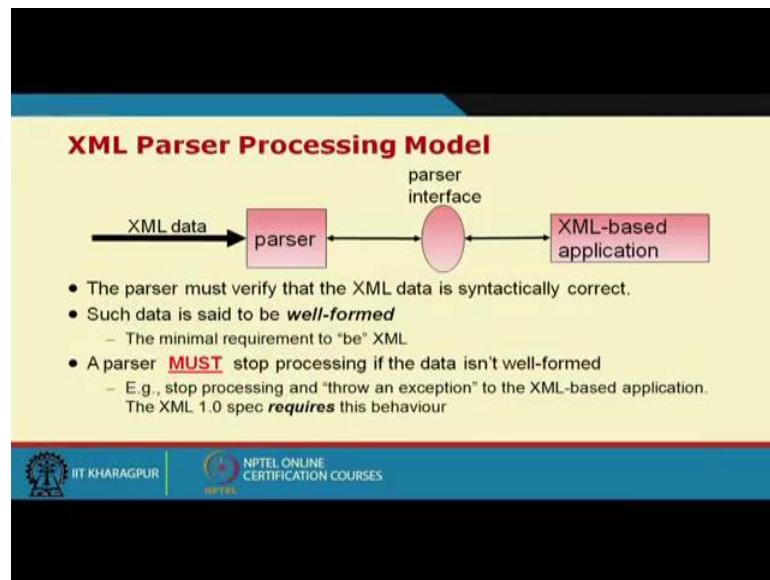
A pink box labeled "xml-simple.xml" is positioned to the right of the XML code. At the bottom of the slide, there is a logo for IIT Kharagpur and NPTEL Online Certification Courses, along with a small circular portrait of a man.

Now, other access is how to process XML data. So, one thing what we have seen, it say tree type of structure. Now in order to now unlike other standard like HTML or where you have no user defined things. So, you know the not only the syntax they are elements etc defined.

But here in this case, I need to first of all find out that, what are the in numbers of element, what are whichever is users defines and I need to find out that what they mean in the first place. So, in order to process this data, suppose I have a data like this, taken from some Internet resources. Like, file transfer date is so much type of things is interbank and so much amount transit ID and type of things, different type of transactions. So, it is like bank transactions, some are internal that is within the bank and some are interbank or between different banks so, then how to process these data.

So, for this I need to first of all extract the information from the XML data all right then my processing will be going on. It is basically enveloped in a XML type of language and then I have do extract these and process it using some application some sort of a tool that I have to extract the data; that what we do is doing through a XML parser.

(Refer Slide Time: 13:17)



So, XML parser processing model, what it does that XML data the parser, it passes through the parser and there is a parser interface than XML based application on the other side. So, XML data then the processer, it extracts the things that XML based applications are triggered, all right.

So, the parser must verify the XML data is syntactically correct. So, what is the role of the parser, first of all the parser should verify the XML data is syntactically correct. So, there is no syntax error in the data. Secondly, such data is said to be well formed right. If it is a syntactically correct, it is a well formed XML. So, the minimum requirement to be a XML is this it should be well formed, so that it is processable, right, I can process these data. The parser must of processing, if the data is not well formed, right. So, that is stop processing and throw an exception to the XML based application.

(Refer Slide Time: 14:40)

```
<?xml version="1.0" encoding="utf-8" ?>
<!DOCTYPE transfers [
    <!-- Here is an internal entity that encodes a bunch of
        markup that we'd otherwise use in a document -->
    <!ENTITY messageHeader
        "⟨header⟩
            <routeID> info generic to message route </routeID>
            <encoding>how message is encoded </encoding>
        </header⟩
    "⟩
]>
<transfers>
    &messageHeader;
    <fundsTransfer date="20010923T12:34:34Z">
        <from type="intrabank">
    </transfers>
```

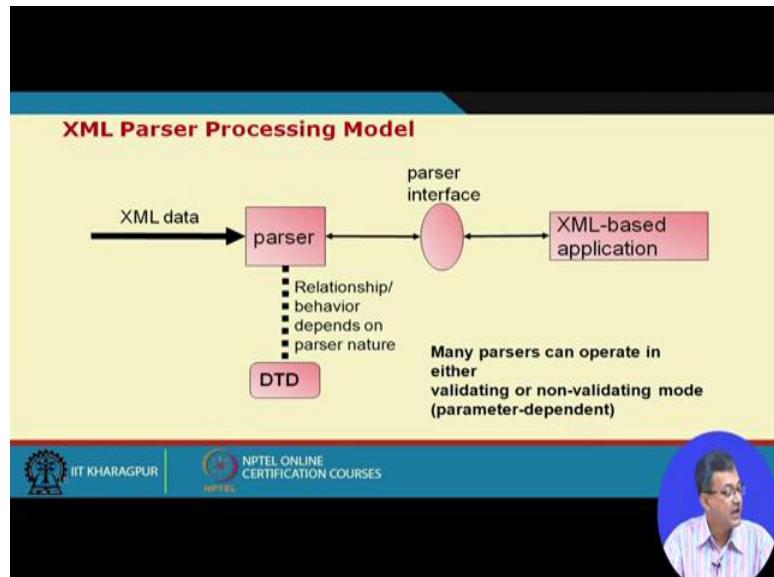
So, that if it is not a well formed; that means, if it is syntactically not a correct data. So, if we look back, the XML processing rule including different parts they so, there is a concept called DTD, that is document type declaration. So, what is important for any XML data, like we have seen in case of a data bases like relational data bases, so, what is what you have to do basically, define a table right or multiple table. So, where the data is different or in form of records or different rows or topples what we say.

Now, this in order to define this table, we need to define the structure of the table or what we say schema of the table all right so that if there are different variable of this table, then which variable is which schema and type of things we need to define, right. So, in case of XML also, as these are user defined structure and it is hierarchical and I do not know that if you define a particular element, then what are the levels of other elements and what are the types of elements are there, and that is why here also, we define we need to define some sort of a schema or in this previously it is used they that stuff called document type declaration.

So, first of all you need to declare the document type. Now based on the document type, if it is a in the then the rest of the data is there. So, one way; one of the work of the parser is to look at whether it is syntactically correct. Like you have syntactically, other than in to look at whether it follows that particular document, definition right or if it is I say that

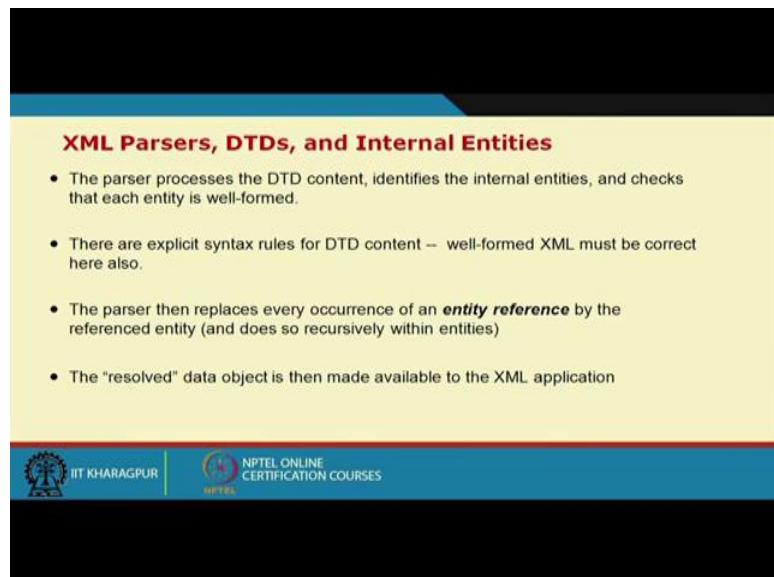
syntactically correct, is a well formed. Then if it is following that schema or document type definition then says it is a valid one XML.

(Refer Slide Time: 16:58)



So, the while getting the XML data, the parser consult this DTD. If it is syntactically correct, to say that the XML is valid for this particular type of operations details, looking for.

(Refer Slide Time: 17:15)



So, we have XML parser, DTD, another internal entities. Like the parser processes, the DTD content identifies the internal entity and checks that the each entity is well formed

correct. So, in term it checks that whether each entity is well formed. There are explicit syntax rule for DTD content, well formed XML must be correct here also right. So, what is there are explicit syntax rules for DTD content so the well form XML must be correct in this respect also.

The parser then replaces every occurrence of entity reference by a referenced entity right. So, does the recursively within the entities. So, it reference entity by the referenced entity it replace so that it is entity reference, the resolved data object is then made available to the XML application. So, when it goes to this XML application, if we just remember this previous picture, then by that time it has done It is well form checking syntactical checking and also the checking with the DTD, that is what we can say some sort of a validating, the particular document before pushing it to the XML application.

(Refer Slide Time: 18:45)

XML Processing Rules: External Entities

Put the entity in another file -- so it can be shared by multiple resources.

```
<?xml version="1.0" encoding="utf-8" ?>
<!DOCTYPE transfers [ . .
<!ENTITY messageHeader
    SYSTEM "http://www.somewhere.org/dir/head.xml"
    >
]>
<transfers>
    &messageHeader;
    <fundsTransfer date="20010923T12:34:34Z">
        <from type="intrabank">
    </transfers>
```

External Entity declaration

Location given via a URL

xml-simple-extEntity.xml

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, in XML processing there can be external entities right. Here it was internal entities, previously what we have seen that is internal entities means that DTD is defined here only or the schema is defined here. So, external entity is referred by a URL. So, it basically referred to external structure or schema to look at the entities. So, the parser processes the DTD content, identifies the external entities and tries to resolve them, all right.

(Refer Slide Time: 19:22)

XML Parsers and External Entities

- The parser processes the DTD content, identifies the external entities, and "tries" to resolve them
- The parser then replaces every occurrence of an **entity reference** by the referenced entity, and does so recursively within all those entities, (like with internal entities)
- But what if the parser can't find the external entity (firewall?)?
- That depends on the application / parser type
 - There are **two types of XML parsers**
 - one that **MUST** retrieve all entities, and one that can ignore them (if it can't find them)

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

The parser then replaces every occurrence of the entity reference, by the referenced entity and does so recursively with all entities like with the internal entities. For the external also it finds out and goes on replacing this. But what if the parser cannot find the external entity, due to may be due to firewall blockage etc. That depends on the application and parser type. There are two types of XML parser one must retrieve all entities; one can ignore them if they can find it.

So, it based on the; what is your basic policy or processing rule for this parsers.

(Refer Slide Time: 20:20)

Two types of XML parsers

- **Validating parser**
 - **Must** retrieve all entities and must process **all** DTD content. Will stop processing and indicate a failure if it cannot
 - There is also the implication that it will test for compatibility with other things in the DTD -- instructions that define syntactic rules for the document (allowed elements, attributes, etc.). We'll talk about these parts in the next section.
- **Non-validating parser**
 - Will try to retrieve all entities defined in the DTD, but will **cease processing the DTD** content at the first entity it can't find. But this is not an error -- the parser simply makes available the XML data (and the names of any unresolved entities) to the application.

Application behavior will depend on **parser type**

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, accordingly we have 2 type of XML parser, one wise say validating parser; that means, must retrieve all elements and must process all DTD content all right. We will stop processing indicate a failure, if it is cannot, right. So, what we are trying to say that in case of a validating parser. It should conform to this schema definition or the DTD content, right. Non validating parser will try to retrieve all elements defined in the entity, but will cease processing, the DTD content in the first entity if it cannot find, right. So, it based on validating parser or non validating parser application behavior will depend on the parser type definitely. So, it the application how it will work is based on that what sort of parser.

So, we add on something more, like here XML data then parser. Then it goes to the parser interface to the XML application, right. So, DTD is linked with the parser, which gives relationship, behavior dependency on parser nature. So, these are the things who is the DTD provides to this parser for making for checking, for validating well formed and validating, right.

(Refer Slide Time: 21:51)

Special Issues: Characters and Charsets

- XML specification defines what characters can be used as whitespace in tags: <element _id = "23.112" />
- You cannot use EBCDIC character 'NEL' as whitespace
 - Must make sure to not do so!
- What if you want to include characters not defined in the encoding charset (e.g., Greek characters in an ISO-Latin-1 document):
- Use **character references**. For example:
 - $\&\#9824;$ -- the spades character (\spadesuit)
 - 9824th character in the Unicode character set
- Also, binary data must be encoded as **printable characters**

So, there are some special issues, characters and character sets right. So, like the XML specification, defines what characters can be used in whitespace tag and like you cannot use EBCDIC character NEL as whitespace tag. What if you want to include character not defined in the encoding character sets right. So, there are different ways out for

handling this sort of character sets all right. So, the XML provides some mechanisms to handle this sort of character sets.

Now, how do I define language dialect, finally, what we are trying to look at is basically how to define an XML document and how it can basically interact with each other right. So, as we know that in service oriented architecture, we are having 3 major components right. Like consumer, provider and a registry or type of services right or some sort of arrays to depository. Now whenever there is a communication between these different service providers, service consumer, service registry what we need to look is that, how this data communication go, will be executed in a ubiquitous way. Like as we you know that, this is a scope messaging maybe one mechanism which is primarily and XML document all right. Like soap WSDL, UDDI, I all are basically XML documents correct.

So, if you look at, two ways to doing that XML document type, declaration part of core XML spec, XML schema, new XML specification, which allows for stronger constraints on XML document. So, what you have looked at is DTD, now 2001 the XML schema or XSD has been defined. So, it is a higher version of on spec it is a new spec of defining the schema of the structure of the XML document. And one basic or fundamental difference is there other than they primarily do the major same type of job; one is that XML schema is written in XML, all right; whereas DTD is actually written in a different way. So, that now my handling of the schema in XML data goes hand in hand.

So, adding dialect specification implies two classes of XML, one we call well formed an XML document that is syntactically correct right. So, what we say, it is a well formed XML, right and where as there is a thing called valid XML, right. So, XML document that is both well formed and consistent with the specific DTD or XSD or schema all right, these we are called valid. So, it should be well formed along with a; that it is basically conforming to the schema; that means, it is a valid scheme.

So, all valid XML document are well formed. So, what DTD and other schema specify allow elements and attribute data, hierarchical nesting rules, element, content and type restrictions. So, this is a schema or thing specified. So, it is not the data. So, the XML schema or the DTD does not content any XML data right. Information is contented in the XML file, but the structure is defined here. So, this makes a unique property, like whenever there is a two organization A, B want to exchange information.

So, what they are looking for is first of all this should agree upon the structure of that exchange protocol right. Like if say for example, IIT Kharagpur having a say data transfer, say payment gateway or some payment transfer, payment interaction with bank or any financial organization. What is important that, first of all the structure of a student data and at the IIT Kharagpur end and the way they stored in the State bank. They are or any banks State bank or Punjab bank or anything, any bank they are should be that the banking organization there should be agreement.

For that date, for that need I do not require any data to be exchanged right? So, that is the basic duty of the thing right. So, for that I do not require any data to be exchange. What we required is more of the structure, to be exchanged and incidentally exchanging structure is not that critical right. So, the data there are privacy issues, there are issues of maintenance costs towards collecting maintenance of this data. If you just like that, say or then it is basically go on out of your hand. So, that is why the data has lots of issues, but if I say that see in IIT Kharagpur, this student structure is like this. So, student schema is like this right. I do not tell that how many students are there what are the properties but the school student's schema is like this.

Then if I am, if it is interacting with some other organization, the schema wise integration is possible. So, that is why allows element attribute, names hierarchical nesting rules element content type restrictions. Schemas are more powerful than DTDs, they are often used for type validation or for relating database schemas to the XML models, right. Schemas are more powerful much more powerful than DTD, right.

They not only keep the structure, but that can be used for different type of things. Like using the schema, I can create a database table and then I can transfer the data ubiquitously among the two things right. So, this gives a extra handling to this schemas. So, these days we are primarily using schemas and schemas plays a important role in handling this sort of inter operations between different parties.

(Refer Slide Time: 29:20)

```
<!DOCTYPE transfers [  
    <!ELEMENT transfers (fundsTransfer)+>  
    <!ELEMENT fundsTransfer (from, to)>  
    <!ATTLIST fundsTransfer  
        date CDATA #REQUIRED>  
    <!ELEMENT from (amount, transID?, accountID,  
        acknowledgeReceipt)>  
    <!ATTLIST from  
        type (intrabank|internal|other) #REQUIRED>  
    <!ELEMENT amount (#PCDATA)>  
    <!ELEMENT to EMPTY>  
    <!ATTLIST to  
        account CDATA #REQUIRED>  
>  
<transfers>  
    <fundsTransfer date="20010923T12:34:34Z">  
        As with previous example . . .
```

So, if you look at the example DTD as a part of documents. So, there are document type, elements transfer, this there is a some other type of elements, right.

Now, if you look at this, this is not exactly following the truly the XML file or XML document, all right. Whereas, in case of a schema, it is represent as the XML document, right.

(Refer Slide Time: 29:50)

```
<!DOCTYPE transfers SYSTEM  
    "http://www.foo.org/hereditis/simple.dtd" >  
  
<transfers>  
    <fundsTransfer date="20010923T12:34:34Z">  
  
    transfers  
  
    • Of course, the DTD file must be there, and accessible.
```

And there can be external DTD like there is a food.org something DTD. So, it refers externally when it needs to referring. So, this allows us you know to referred to some

others DTD and see that what sort of data is coming right suppose I want to look at I am consuming a data from another provider and I want to know that what sort of data is coming up like whether they are ordered in a particular fashion whether there is a typical hierarchy.

So, at the other end I can do external DTD and check it that what sort of data. So, that allows me allows my parser to filter and extract the data I need for my end to process. So, my objective is that I am getting a data from a source, I want to filter it extract the portion of the data I want to do and may need some transformation like I want to change the unit from centigrade to fahrenheit or meter to feed and type of things, even that require some transformation that I can do at the this end and then I put to my applications which work on this. We will continue our discussion on XML some other means some other properties of XML to what to see that how these are useful for this interoperate is interoperation and how these are useful for realization of a cloud especially the software as a service type of models.

Thank you.

Cloud Computing
Prof. Soumya Kanti Ghosh
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Lecture – 09
XML Basics (Contd.)

Hello. So, we will continue our discussion on cloud computing. So, if you remember the last lecture you are talking about XML basics, we are talking about this how XML discussing how XML allows us to interoperate between, between different interacting or cooperating system in a or in a distributed system. And how it can it can become a one of the major backbone for realizing cloud computing specially in SaaS type of cloud software as a service type of cloud, right.

So, we were discussing about DTD and XML schema XSD. So, today we will continue that that discussion and on this basics of XML, alright.

(Refer Slide Time: 01:11)

XML Schemas

- A new specification (2001) for specifying validation rules for XML
- Uses **pure XML** (no special DTD grammar) to do this.
- Schemas are more powerful than DTDs - can specify things like integer types, date strings, real numbers in a given range, etc.
- They are often used for **type validation**, or for relating database schemas to XML models
- They don't, however, let you declare entities -- those can only be done in DTDs.
- The following slide shows the XML schema equivalent to our DTD

Specs: <http://www.w3.org/XML/Schema>
Best-practice: <http://www.xfront.com/bestPracticesHomepage.html>

IIT KHARAGPUR | **NPTEL ONLINE CERTIFICATION COURSES**

So, as if you remember that we are discussing about XML schema which is primarily defines the structure of the XML. So, XML of a used data like say I have a XML of a student data of a particular batch of student of IIT Kharagpur, it will be a huge volume of data now whenever I want this data say our placement section or our account section for

the bank or other things. So, they want to share some of these data with the external agencies, right.

So, one way of looking at it that share a part of the schema look at the schema and say are the part of the things with the with the external agencies, or I want to on the other way in your organization takes the data from the other organization. So, the first of all it there is a need of that schema to be agreed upon that what should be there. So, the XML schema place a vital role in exchange of information helping in interoperability between different repositories.

So, a new specification, schema is a new specification which came up in 2001 to specify validation rule of XML, right. So, what we have seen well formed any XML parser first check whether that it is syntactically correct that is well formed. And secondly, the schema it checks with the schema if available to say that whether it is valid with respect to the schema. So, what we say it is a valid XML document, alright. So, so there are W3 some spec and best practice as you can see. So, it use pure XML to do this XML schema that are written in XML, schemas are more powerful and DTDs can specify things like integer type, date, real thing etcetera and other type of parameters.

They are often use for type validation alright, or for relating data base schemas to XML models all right. So, it is I can if I know the schema than I create the database and then pump this XML data to this database or other way around, alright. They do not however, let you declare entities; those can be done only in DTD.

(Refer Slide Time: 03:46)

XML Schema version of our DTD (Portion)

```
<?xml version="1.0" encoding="UTF-8"?>
<xss:schema xmlns:xss="http://www.w3.org/2001/XMLSchema"
  elementFormDefault="qualified">
  <xss:element name="transitID" type="xs:string"/>
  <xss:element name="acknowledgeReceipt" type="xs:string"/>
  <xss:complexType name="amountType">
    <xss:simpleContent>
      <xss:restriction base="xs:string">
        <xss:attribute name="currency" use="required">
          <xss:simpleType>
            <xss:restriction base="xs:NMTOKEN">
              <xss:enumeration value="USD"/>
              <xss:enumeration value="INR"/>
              (more stuff omitted)
            </xss:restriction>
          </xss:simpleType>
        </xss:attribute>
      </xss:restriction>
    </xss:simpleContent>
  </xss:complexType>
  <xss:complexType name="fromType">
    <xss:sequence>
      <xss:element name="amount" type="amountType"/>
      <xss:element ref="transitID" minOccurs="0"/>
      <xss:element ref="accountID"/>
      <xss:element ref="acknowledgeReceipt"/>
    </xss:sequence>
  </xss:complexType>
</xss:schema>
```

simple.xsd

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, it cannot define this in entities. The followings will see that how schema is equivalent to our DTDs. Like the last DTD we have if have if you remember in the XML basics that first class. So, that same thing is now represented as a schema.

So, what we see it is more of XML type document and in doing. So, I can handle this schema the same way I handle a XML data, alright.

(Refer Slide Time: 04:09)

XML Namespaces

- Mechanism for identifying different "spaces" for XML names
 - That is, *element* or *attribute* names
- This is a way of identifying different *language dialects*, consisting of names that have specific semantic (and processing) meanings.
- Thus `<key/>` in one language (might mean a security key) can be distinguished from `<key/>` in another language (a database key)
- Mechanism uses a special `xmlns` attribute to define the namespace. The namespace is given as a *URL string*
 - But the URL does not reference anything in particular (there may be nothing there)

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES



So, that that is a beauty of the thing; another aspect that the important aspect of XML is XML name space, right; it is a very important aspect as XML allows user defined data. So, you need to be careful on that which type of definition is there and which namespace we are looking at. If you remember last class we are discussing that if I define a entity call table or element call table then I need to specify whether this is a table which type of table. Like I can have a table as we have we have discussed that furniture or the table can be from a say table of a word processor or spread said type of things.

So, in order to distinguish I have to say table of furniture type then it has some definition, table of say word processing has some other definition alright. So that means, where I will define this I will define in the namespace. So, the namespace tell me name space basically specifies that which what element is there what sort of dealing etcetera etcetera there. So, it is a mechanism for identifying different spaces for XML names alright that is element or attributes names. This is a way of identifying different language dialects, consisting of names that have specific semantics and processing meanings all right. The key is one language mean a specific security key, as we are talking about table the distinguished from, the key in another language with maybe a database key, right.

So, key maybe some sort of a security key and in some other thing the key what we are looking for is more of a database related keys. So, this two keys are one from database schema a database name space and another for this security name space need to be defined somewhere. So, it is a may use uses a special XMLNS, XML name stands for XML namespace attribute to define this name space right. The namespace is given as a URL string alright, but the URL does not reference does not reference anything in particular they maybe nothing is there. So, it may not refer to something.

(Refer Slide Time: 06:39)

The slide has a blue header bar with the title "Mixing language dialects together". Below it is a yellow content area containing the following text and code:

Namespaces let you do this relatively easily:

```
<?xml version= "1.0" encoding= "utf-8" ?>
<html xmlns="http://www.w3.org/1999/xhtml"
      xmlns:mt="http://www.w3.org/1998/MathML" >
<head>
  <title> Title of XHTML Document </title>
</head><body>
  <div class="myDiv">
    <h1> Heading of Page </h1>
    <mt:mathml>
      <mt:title> ... MathML markup ...
    </mt:mathml>
    <p> more html stuff goes here </p>
  </div>
</body>
</html>
```

Annotations on the right side of the code explain the namespaces:

- A red box encloses the line "Default 'space' is `xhtml`".
- A red box encloses the line "mt: prefix indicates 'space' mathml (a different language)".

At the bottom of the slide, there is a footer bar with the IIT Kharagpur logo, the text "NPTEL ONLINE CERTIFICATION COURSES", and a circular portrait of a man.

So, namespace so, mixing language dialogues together. So, namespace let you do this relatively easier like I can have something from the something from that like, I want both the keys all right. From the database keys then I use the security key for unlocking the database, and then user database key for accessing the both the keys I am using somewhere someone right. And both are in the XML So, this XML namespace may allow me to do that like I can say that db dot key mean something or somewhere something dot some they that security dot something it means a some other key.

So, I basically specify the type of key is there. So that means, here if you look at these are the name space which are a default space right. Which define by the W3C whereas, this namespace empty is a math related namespace right. That is also a predefined, but I say that the empty is a math related namespace. Now whenever I do anything with this particular math related then I say empty double colon mathml and title and so on and so forth; that means, I say that I am referring these math namespace.

So, empty prefix indicates the space mathml a different language mathml maybe a totally different language and I basically able to use those elements of that particular namespace in my document.

(Refer Slide Time: 08:18)

The slide has a yellow background with a black header bar at the top. The title 'XML Software' is centered in red font. Below the title is a bulleted list of points about XML parsers and software. At the bottom of the slide, there is a footer bar with the IIT Kharagpur logo, the text 'NPTEL ONLINE CERTIFICATION COURSES', and a circular profile picture of a man.

XML Software

- **XML parser** -- Reads in XML data, checks for syntactic (and possibly DTD/Schema) constraints, and makes data available to an application. There are three 'generic' parser APIs
 - SAX Simple API to XML (event-based)
 - DOM Document Object Model (object/tree based)
 - JDOM Java Document Object Model (object/tree based)
- Lots of XML parsers and interface software available (Unix, Windows, OS/390 or Z/OS, etc.)
- SAX-based parsers are fast (often as fast as you can stream data)
- DOM slower, more memory intensive (create in-memory version of entire document)
- And, validating can be **much slower** than non-validating

So, this so, namespace as we sees plays a important role and you should be careful, and the namespace definition are should be there. So, in other words so, whenever we do like we have seen that mathml that what we do is more of looking at that particular namespace and how to use it and type of things.

Another important aspects already we have discussed some of these is XML software, right. So, as XML document is one is representation textual mode whereas, I have a I need to process it and type of thing one a first of all I validate well form or not syntactic recurring this is one part. Secondly, need to process it for different application etcetera. So, this XML software plays a important role.

So, XML parser reads in XML data checks for syntactic possibly with DTD and schema constraint, and makes all data available to an application. There are three generic XML APIs or parser APIs like one is a SAX parser simple API to XML event based it is a very popular widely used parser and available mostly in all platform. Another is DOM parser, document object model or object tree based parser another is JDOM parser, java DOM parser right. So, lots of XML parsers and interface software available in different operating systems. SAX based parsers are fast often as fast as you can stream data and these are very low bellowed parser and this pretty fast though the functionalities less.

Whereas DOM is slower more memory intensive create in memory version of the entire document and type of things all right. And validating can be much slower than non validating parser, because validating parser need to validate whether it is conformed to the schema before doing any processes.

(Refer Slide Time: 10:33)

XML Processing: SAX

A) SAX: Simple API for XML

- <http://www.megginson.com/SAX/index.html>
- An **event-based** interface
- Parser reports events whenever it sees a tag/attribute/text node/unresolved external entity/other
- Programmer attaches “event handlers” to handle the event

- **Advantages**
 - Simple to use
 - Very fast (not doing very much before you get the tags and data)
 - Low memory footprint (doesn’t read an XML document entirely into memory)
- **Disadvantages**
 - Not doing very much for you – you have to do everything yourself
 - Not useful if you have to dynamically modify the document once it’s in memory (since you’ll have to do all the work to put it in memory yourself!)

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES



So, these are some of the quick look at the different parsers like if the SAX parser is the available there, it say event based interface parser reports event whenever it sees a tag attribute text node unresolved external entities and other.

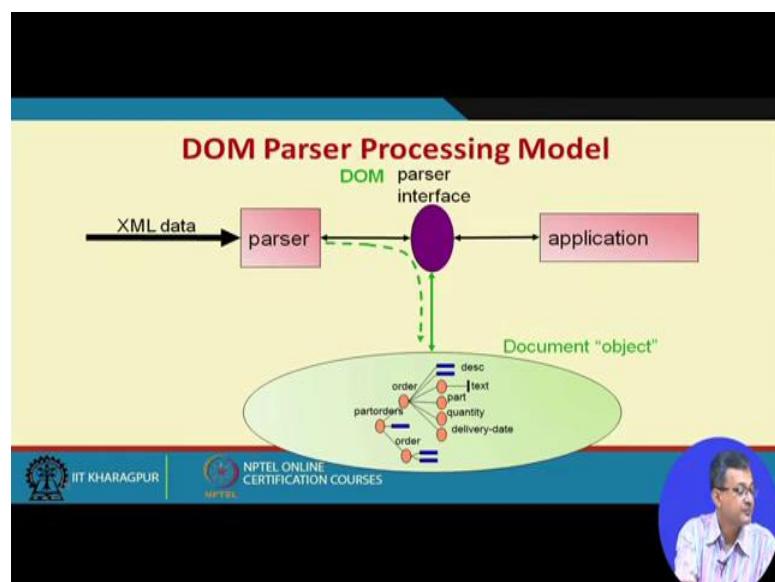
Programmers attach event handlers to handle the events, right. Advantages it is simple to use, very fast, not doing much before you get the tax and data, right. So, it is a it is pretty fast, low memory footprints. So, it is low payload. Disadvantages not doing much of the processing. So, you need to do some of the processing at your end. Not useful if you have dynamically modify document world series in memory. So, it is not useful if you dynamically modify the documents it is memory since you will have to do the all the work and put memory of yourself anyway. So, what it is there it is useful when you have a vanilla type XM processing requirement which process and goes on with low memory footprint and ready fast and overall pretty simple to use.

Whereas the document object model or the DOM parser it is a object based interface parser generates in memory tree, in memory tree or the in memory XML tree

corresponding to the XML document. DOM interface identifies method for accessing and modifying the tree, right. So, that is the advantage very useful for dynamic modification access to the tree. So, if there is a dynamic modification access to the tree dynamically changing then it is extremely useful, useful for query that is looking for data and depends on the tree structure. So, querying because if it is a tree structure which plays a role it is useful. It is the same interface for many programming language C plus plus, JAVA that is as such the interface wise interpret interoperable with different languages; disadvantage can be slow, right.

Needs to produce the tree and may need lots of memory DOM programming interface is little bit not that state forward. So, it needs to be little complex and needs more what we say professional handling of the data, here handling of this programming.

(Refer Slide Time: 13:08)



So, in case of a DOM parser the parser interface it goes to this structure of the tree which is in build maintaining the thing and based on that it grabs the application. The JDOM or the java DOM it is java based object oriented.

(Refer Slide Time: 13:24)

C) JDOM: Java Document Object Model

XML Processing: JDOM

- <http://www.jdom.org>
- A Java-specific **object-oriented** interface
- Parser generates an in-memory tree corresponding to the document
- JDOM interface has methods for accessing and modifying the tree

- **Advantages**
 - Very useful for dynamic modification of the tree
 - Useful for querying (i.e. looking for data) that depends on the tree structure
 - Much nicer Object Oriented programming interface than DOM
- **Disadvantages**
 - Can be slow (make that tree...), and can take up lots of memory
 - New, and not entirely cooked (but close)
 - Only works with Java, and not (yet) part of Core Java standard

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, it is similar to DOM with the; it is from java. Parser generates in memory tree corresponding the document. JDOM interface has methods for accessing and modifying the tree, right. So, where accessing and modifying the tree, advantages very useful dynamic modification of the tree useful for querying that depends on the tree structured. So, if your query is dependent on the tree structure it is pretty first and useful. Much nicer object oriented programming interface than DOM. So, it is much means better programming interface is there.

These advantage can be slow, a new and not entirely cooked not. So, new these days, but there are you need some expertise to work on it and it works on JAVA that is that maybe disadvantage or it may be thought of what we say that they you need to know JAVA to work on it may not be a big issue.

(Refer Slide Time: 14:30)

C) dom4j: XML framework for Java

XML Processing: dom4j

- <http://www.dom4j.org>
- Java framework for reading, writing, navigating and editing XML.
- Provides access to SAX, DOM, JDOM interfaces, and other XML utilities (XSLT, JAXP, ...)
- Can do "mixed" SAX/DOM parsing -- use SAX to one point in a document, then turn rest into a DOM tree.

- **Advantages**
 - Lots of goodies, all rolled into one easy-to-use Java package
 - Can do "mixed" SAX/DOM parsing -- use SAX to one point in a document, then turn rest into a DOM tree
 - Apache open source license means free use (and IBM likes it!)
- **Disadvantages**
 - Java only; may be concerns over open source nature (but IBM uses it, so it can't be that bad!)

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

There are few more like a DOM4J which is java framework for reading writing navigation and editing XML right. Provides access to SAX, DOM, JDOM interface and other XML entities like XSLT and other type of interface, can do mixed SAX, DOM parsing. And it has some of the some advantages like good of all lots of goodies and all rolled into one things in JAVA package can make SAX and DOM parser Apache open source license, means free use and so and so forth.

(Refer Slide Time: 15:17)

Some XML Parsers (OS/390's)

- Xerces (C++; Apache Open Source)
<http://xml.apache.org/xerces-c/index.html>
- XML toolkit (Java and C++; Commercial license)
<http://www-1.ibm.com/servers/eserver/zseries/software/xml/>
I believe the Java version uses XML4J, IBM's Java Parser. The I believe the Java version uses XML4J, IBM's Java Parser. The
Java version is available at:
<http://www.alphaworks.ibm.com>
- XML for C++ (IBM; based on Xerces; Commercial license)
<http://www.alphaworks.ibm.com/tech/xml4c>
- XMLBooster (parsers for COBOL, C++ ... Commercial license; don't know much about it; OS/390? [dunno])
<http://www.xmlbooster.com/>
Has free trial download.; can see if it is any good :-)
- XML4Cobol (don't know much about it, any COBOL85 is fine)
<http://www.xml4cobol.com>
- www.xmlsoftware.com/parsers -- Good generic list of parsers

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

There are disadvantages it is again JAVA only and maybe concerns for open source nature of things like that, some of the things, but nevertheless it is a good one. There are some other XML parser like Xerces and XML toolkit, XML for C plus and etcetera, etcetera. So, rather there are a good number of XML parser to not only well formed and validating also try to do more complex operations in our, in basically in when we are doing interoperating in some distributed system like cloud.

(Refer Slide Time: 15:50)

The slide has a yellow header bar with the title "Some parser benchmarks:". Below the title is a bulleted list comparing SAX and xDOM. The list includes links to developerworks and devsphere websites, their respective dates (Sept 2001 and late-2000), and a section titled "Basically" with further comparisons. The footer features logos for IIT Kharagpur and NPTEL.

Some parser benchmarks:

- <http://www-106.ibm.com/developerworks/xml/library/x-injava/index.html> (Sept 2001)
- <http://www.devsphere.com/xml/benchmark/Index.html> (Java) (late-2000)

• Basically

- SAX faster xDOM slower
- SAX less memory xDOM more memory
- SAX stream processing xDOM object / persistence processing
- nonvalidating is always faster than validating!

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, there are some of the benchmark like how the parser will be there, whether regarding speed memory and stream processing, there are few parser benchmarks.

(Refer Slide Time: 16:08)

XML Processing: XSLT

D) **XSLT eXtensible Stylesheet Language -- Transformations**

- <http://www.w3.org/TR/xslt>
- An XML language for processing XML
- Does tree transformations – takes XML and an XSLT style sheet as input, and produces a new XML document with a different structure

- **Advantages**
 - Very useful for tree transformations -- much easier than DOM or SAX for this purpose
 - Can be used to query a document (XSLT pulls out the part you want)
- **Disadvantages**
 - Can be slow for large documents or stylesheets
 - Can be difficult to debug stylesheets (poor error detection; much better if you use schemas)

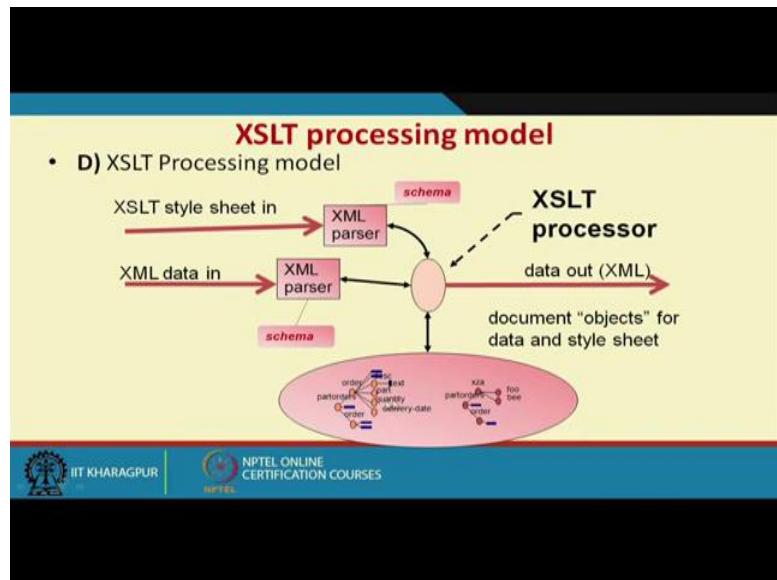
IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

XML processing another important aspect is XSLT right. So, extensible style sheet language, right. So, how XML to be represented is given by this style sheet, right. So, in some other sense I can say that XML along with the styling allows me to represent it or displayed right. So, sometimes XML plus XSLT is something equivalent to HTML type of things like not that all power wise, but it is characteristics wise. So, XSLT extensible style sheet language is primarily used for transformation type of things right. Or it is more of a transformation type of things. XML language for processing XML data, does tree transformation takes XML and XSLT style sheet as input, and produce a new XML document with a different structure and type of things.

So, as it is XSLT. So, I can basically do some sort of a filtering operations on the XSLT right. So, it takes the XML data, does some sort of process it XSLT and generate a another XML data on the other hand and produces a new XML document with different structure and so on and so forth. So, advantages of course, very large very useful for tree transformations; so, if I want to transform the tree or say I have a tree XML tree, and for my application I require want to use a sub tree of it, right. So, how can I do it? So, I can have this filtering using this XSLT and then filter these things to the other part.

So, disadvantage can be slow for large document or style sheets right. It can be disadvantages and can we slow can be difficult to debug style sheet, poor error detection, etcetera.

(Refer Slide Time: 18:16)



So, that another is that not so versatile in error detection. So now, if you look at our big picture. So, XML data is coming right. So, we have now a style sheet with a it requires again another XML parser. And then it goes to this processor, processing unit. And we have this different tree.

So, I have a tree here and I can have a part of the tree out here, right. So, taking these as input the data out is a XML which is which is based on the XSLT has been filtered for the and generated new XML. XML messaging use XML as the format for sending messages between the systems. Advantages common syntax, self describing easier to parse can use common existing transport mechanism to move the XML data like https etcetera.

(Refer Slide Time: 18:59)

XML Messaging

- Use XML as the format for sending messages between systems
- Advantages are:
 - Common syntax; self-describing (easier to parse)
 - Can use common/existing transport mechanisms to "move" the XML data (HTTP, HTTPS, SMTP (email), MQ, IIOP/(CORBA), JMS,)
- Requirements
 - Shared understanding of dialects for transport (required registry [namespace!]) for identifying dialects
 - Shared acceptance of *messaging contract*
- Disadvantages
 - Asynchronous transport; no guarantee of delivery, no guarantee that partner (external) shares acceptance of contract.
 - Messages will be much larger than binary (10x or more) [can compress]

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, XML needs a carrier protocol to transfer the data. So, I can we can use any transport data transport mechanism like transport we should not mix up this transport with our transport layer protocols. So, it is a transport mechanism means underlying the network and by which you can do that http, https, SMTP and so on and so forth. And there are some requirement and disadvantages of the same this type of XML messaging.

(Refer Slide Time: 19:55)

Common messaging model

- XML over HTTP
 - Use HTTP to transport XML messages
 - ```
POST /path/to/interface.pl HTTP/1.1
Referer: http://www.foo.org/myClient.html
User-agent: db-server-olb
Accept-encoding: gzip
Accept-charset: iso-8859-1, utf-8, ucs
Content-type: application/xml; charset=utf-8
Content-length: 13221
.

<?xml version="1.0" encoding="utf-8" ?>
<message> . . . Markup in message . . .
</message>
```

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

And there are common messaging models. So, a using a http to protocol for XML messages, in other since this XML messages are if you see this is a post at said that this portion of the thing is a standard protocol, http protocol mechanisms.

Whereas, it is the XML message or the XML document is embedded into the things. So, that it is a envelope which is the http protocol itself and then it is embedded into the stuff.

(Refer Slide Time: 20:32)

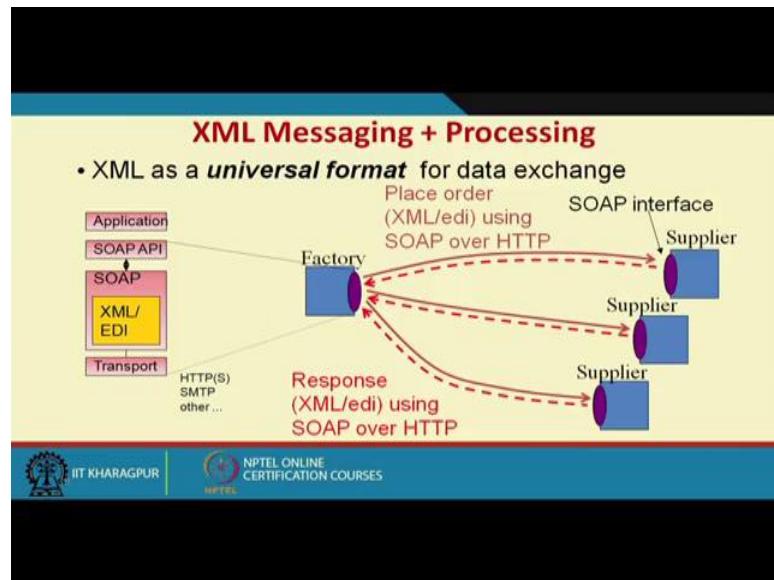
The slide has a yellow background with a black header and footer. The title 'Some standards for message format' is in red. The content is a bulleted list:

- Define dialects designed to "wrap" remote invocation messages
- **XML-RPC** <http://www.xmlrpc.com>
  - Very simple way of encoding function/method call name, and passed parameters, in an XML message.
- **SOAP** (Simple object access protocol) <http://www.soapware.org>
  - More complex wrapper, which lets you specify schemas for interfaces; more complex rules for handling/proxying messages, etc. This is a core component of Microsoft's .NET strategy, and is integrated into more recent versions of Websphere and other commercial packages.

At the bottom, there are logos for IIT Kharagpur and NPTEL, with the text 'NPTEL ONLINE CERTIFICATION COURSES'.

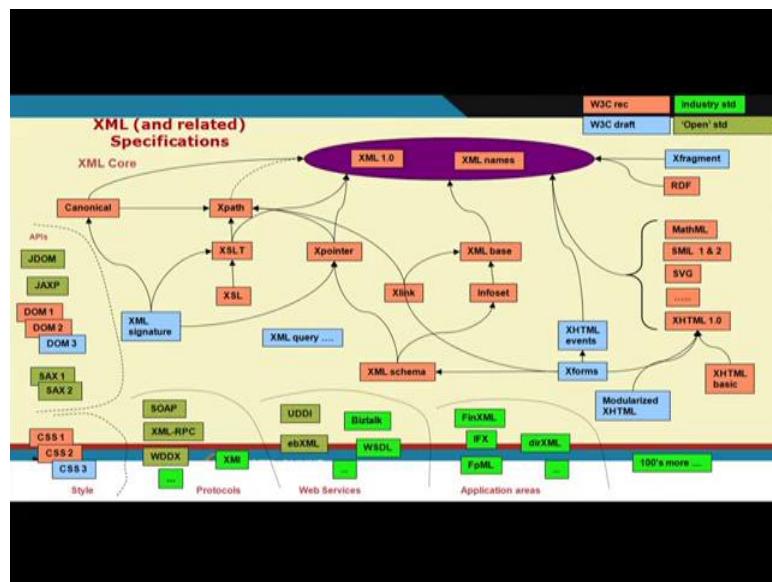
So, there are some of the standards for message format like XML-RPC. So, a very simple way of encoding function method calls call name and passed in the things. So, it is one of the message format is soap simple object access protocol primarily used in web services in a service oriented architecture. So, more complex wrapper which lets you specify schemas for interface more complex rules for handling and proxy messages and so and so forth.

(Refer Slide Time: 21:13)



So, that can be done through this type of soap message format. And if you see again the XML messaging plus processing XML is a universal format for data exchange right. So, I have this application with soap API, and from here I can basically two different type of supplier. So, interact with different type of supplier though, at the backbone the overall it is the same XML applications which interacts with different other applications over http. So, it is a soap message which is which piggyback on http and carried over as a payload of this http and deliver at the things, the other entities extracts and deliver to that particular applications which is running.

(Refer Slide Time: 22:12)



So, if we look at this family of XML all right. So, if you look at this is this middle portion is the cores of the things like we have a XML, XSLT, XSL, X path, X pointer and so forth we have not covered or we have not discussed on all the component it is first of all that amount of time is not there. And Secondly, we want to have a fill that how this allows me or what are the typical properties of XML is there. Those who are interested in a more looking at the XML can follow any standard book for event W3C see a school tutorial and try to do some small projects of the things.

Nevertheless; so, we have this sort of APIs these are different styling, these are different protocols, web services application areas where the XML feeds in. And there can be different other type of things. So, what it tries to show the versatile nature of this XML and it is companion technologies. This XML and this community plays a vital role in realizing this distributed structure of systems right. So, I have distributed systems which are which are talking to each other and doing that is why I have soap messaging, WSDL messaging, WSDL UDTI which are primarily XML based.

And so, and we have different type of other namespaces like mathml and there are SVG for generating graphics and type of things. So, this shows this XML is a versatile in nature in allowing data transformation and if I want to represent this data then I require some styling either through XSLT or some XLD some style sheets which allows me to

represent the data and type of things even; I can generate a HTML file which can be viewed in to the viewed at the other end, right.

So, we basically have what we had is a very basic introductory things or XML try to see that what are the major components of the XML, and how this component help us in realizing interoperability. If you look at the cloud or any distributed systems where the sort of interact interfaces interacting or interoperability needs are there XML is the de facto language to realize this. So, with this we will conclude today, for our this basics of XML.

Thank you.

**Cloud Computing**  
**Prof. Soumya Kanti Ghosh**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**

**Lecture – 10**  
**Web Services, Service Oriented Architecture**

Hi, welcome to this series of lectures on cloud computing, today we will be discussing on one aspect which is sometime considered as one of the major prime mover or building block of cloud computing, that is web services and service oriented architecture. So, we look at that what exactly it means and try to have a overall a overview of the whole thing. So, that it will help us in understanding that; how it build up and how things are going on.

(Refer Slide Time: 00:56)

**What are “Web Services”?**

"Software application identified by a URI, whose interfaces and bindings are capable of being defined, described, and discovered as XML artifacts" – W3C Web Services Architecture Requirements, Oct. 2002

"Programmable application logic accessible using Standard Internet Protocols..." – Microsoft

"An interface that describes a collection of operations that are network accessible through standardized XML messaging ..." – IBM

"Software components that can be spontaneously discovered, combined, and recombined to provide a solution to the user's problem/request ..." - SUN

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Now, what are web services there are several definitions, you will find in the literature and the internet. So, like if you see the W3C web service architecture requirement specification, it says that software application identified by a URI, whose interfaces and bindings are capable of being defined described discovered as XML artifacts. So, this is a very versatile definition.

So, it is, what it says that the software applications, which are identified by URI and whose interfaces are bindings are capable so that you can define, you can describe, you

can discover and all is underlining XML base languages, right. Whereas if you look at the Microsoft define as a programmable application logic accessible using standard internet protocol right. Other definition an interface that describes a collection of operation that the network accessible through standardized accessible messaging and software components that can be spontaneously discovered combined recombined to provide solution to the users problems and request right. So, from this different sort of definitions what we can see one thing is there that is a, it is a XML based phenomena, right.

So, XML as you know it is a more of a data transformation language, it helps in interoperability and helps in application talking to each other right. So, web services are those type of services, which are available on with can be defined in the, can be accessible through a standardize URI and it works on a message exchange type of protocol which is based on XML, right. So, what it gives us it gives us a you can any application can now to application can talk to each other, I am not bothered about what is at the background of the how things are there, I only bothered about where the service is available and how I can talk to that services, right I do not bother about the background processing of the things. This also helps in bringing this legacy application to talk to other applications, right.

So, this is a different way of looking at the things, those who are costumed with client server type of protocol, which are more tightly bound this is more loosely bound and can talk to each other very easily and if you look at the; if I say the genesis of the history of the things.

(Refer Slide Time: 03:34)

**History!**

- Structured programming
- Object-oriented programming
- Distributed computing
- Electronic Data Interchange (EDI)
- World Wide Web
- Web Services

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, there are different aspects which help or which promoted this web services one is definitely structured programming, which evolved to a next suggestion of object oriented programming, distributed system as we have discussed in our previous lectures.

So, it is played a important role, there is another concept of electronic data exchange between two entities primarily to business entities, when they want to interface that data for some particular purpose. Of course, this World Wide Web which allows the whole computing phenomena to connect to each other right with backbone internetworking primarily driven by internet protocol. So, these are this is also a one of the phenomena, these you know. So, this all this component has all these components have; now has what we say facilitated that this development of web services or the evolution of web services.

(Refer Slide Time: 04:46)

## Distributed Computing

- When developers create substantial applications, often it is more efficient, or even necessary, for different task to be performed on different computers, called N-tier applications:
  - A 3-tier application might have a user interface on one computer, business-logic processing on a second and a database on a third – all interacting as the application runs.
- For distributed applications to function correctly, application components, e.g. programming objects, executing on different computers throughout a network must be able to communicate.  
E.g.: DCE, CORBA, DCOM, RMI etc.
- Interoperability:
  - Ability to communicate and share data with software from different vendors and platforms
  - Limited among conventional proprietary distributed computing technologies

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Now, distributed computing we are I am not going any detail, because there is we have talked in detail talked in length that in our previous lectures. So, one of the issues; whenever we are doing distributed computing interoperability right one of the major issues; other than it working another type of things.

So, ability to communicate and share data with software from different vendors and platform very interesting phenomena and which allows this whole world of or whole gamut of application to talk to each other that is it allows you to communicate share data software from different vendors to platforms. Limited among conventional proprietary distributed computing technologies like in case of interoperability and more, limited scopes, we need to expand it.

(Refer Slide Time: 05:36)

**Electronic Data Interchange (EDI)**

- Computer-to-computer exchange of business data and documents between companies using standard formats recognized both nationally and internationally.
- The information used in EDI is organized according to a specified format set by both companies participating in the data exchange.
- Advantages:
  - Lower operating costs
    - Saves time and money
    - Less Errors => More Accuracy
    - No data entry, so less human error
  - Increased Productivity
    - More efficient personnel and faster throughput
  - Faster trading cycle
    - Streamlined processes for improved trading relationships

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES



There is another major development means in this area is the EDI or which is popularly known as electronic data interchange, that computer to computer exchange of business data and documents between companies using standard formats recognized both nationally and internationally; that means, it is a primarily a business data or formats, but the formats are recognized by the both the party.

So, I understand that the other the organization A system understand the; what the what type of message is coming from the organization B and then go was exchanging. The information used in EDI is organized according to a specific format said by both companies participating the exchange. So, it is important that the format is somewhat pre known or pre defined that how or exchange that format or the schema exchange previous to the exchange of the data.

So, other party knows that what sort of data is expecting. There are lot of advantages like one is that low operating cost because you do not bothered about when exchanging data less data more accuracy because this data exchange phenomena loop is not there. So, no data entry less human error. So, directly getting the data from the other things; increase productivity obvious and faster trading cycle if there are multiple companies are working together that is one aspect.

(Refer Slide Time: 07:00)

**Web Services**

- Take advantage of OOP by enabling developers to build applications from existing software components in a modular approach:
  - Transform a network (e.g. the Internet) into one library of programmatic components available to developers to have significant productivity gains.
- Improve distributed computing interoperability by using open (non-proprietary) standards that can enable (theoretically) any two software components to communicate:
  - Also they are easier to debug because they are text-based, rather than binary, communication protocols

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Now all those things have somewhat facilitate this as we are discussing is that the web services. So, it take the advantages, advantages of object oriented programming by enabling developers to build application from existing software component in a modular's approach, right. So, it helps developer to develop applications and from the legacy things or the existing thing, and help them to different component to develop different components in a modular approach.

So, transform a network that is internet into a library of programmatic components available to the developers, to have significant productivity gains. So, if we look at its a as if the library of different programming components are there, now you can go on building things to make your application now, right. So, one if we will take a type of example that this how this services are exchange, is maybe what we can think of is our this online reservation or booking systems right, I book a train or say flight.

So, I have the I gave means when I do when we use a particular address or URI to access that, and then I then those in-turn say any that particular travel portal is not having neither having flights, non having hotels or non having anything right. So, it in-turn talk to the other different airlines services pull the data showed to the user its selects and then the if the booking process goes on, then it goes on a it through a credit card debit card, net banking so other services are called.

So, if you see there is a multiple services are being amalgamated in a proper choreographic way to execute a job. So, my job was to select say a best possible flight based on my budget and my convenience of time, and I want to do it online. So, I go for a some sort of a travel portal or what we say some sort of a broker right which allows me to see different airlines stuffs and then I select my suitable things go on paying through my credit card, debit card etcetera and then the I get the ticket generative right. So, see neither this airline organizations, they are directly; I am not directly hooking to them, they are being connected neither this your credit card or debit card service provider, have any clue that whether you are buying a ticket, etcetera. So, it what it says that if it gets a request in a particular form, it will acknowledge it and then replied it one particular form, right.

So; that means, it is some sort of a XML type of message, exchange going on to the things. So, this allows us to generate different type of mega applications of using some different type of different software components or different type of other application.

So, I can mix and match and doing the things and it goes on a choreographic way list minimal involvement of the different service provider and it goes on like that. So, if you see that improve distributed computing interoperability by open non-proprietary standard; that can enable theoretically any two or more software components to communicate; so, it uses a open standard and facilitated interoperability as you are discussing about XML and type of things.

(Refer Slide Time: 10:58)

**Web Services (contd...)**

- Provide capabilities similar to those of EDI (Electronic Data Interchange), but are simpler and less expensive to implement.
- Configured to work with EDI systems, allowing organisations to use the two technologies together or to phase out EDI while adopting Web services.
- Unlike WWW
  - Separates visual from non-visual components
  - Interactions may be either through the browser or through a desktop client (Java Swing, Python, Windows, etc.)

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, it provides capability similar to those of EDI like exchanging data, but a simpler and less expensive to implement, like I do not have any sort of a predefined agreement on the data format etcetera right. Rather somewhere other I should able to know that where the things have goes on. Configured to work with EDI systems can be configured to work allowing organizations to use two technologies together or to phase out EDI while adopting web services right.

So, unlike www or World Wide Web, separate visual from non visual components, right. So, this is important one is that what I visually see right another is that what goes on the background. So, these two things are separated like if you look at the XML type of things XML is more of a data transformation language. So, it is not a data visualization language; in other since html as we know is more of a data visualization or it displays the things, right. So, XML more work on interoperability type of issues and how did I will be exchange and so and so forth.

So, XML with styling of the XML plus styling of the data will help in visualize or displaying the data. So, unlike our normal http base things which displays the thing it is more of a data representation or data transformation type of stuff.

So, interaction maybe either through browser or through desktop client, like can be java swing, python, windows etcetera that can be there are different type of desktop clients

we can interact with the things or the common interface is the; to the interfacing with the browser.

(Refer Slide Time: 12:47)

**Web Services (contd...)**

- Intended to solve *three* problems:
- **Interoperability:**
  - Lack of interoperability standards in distributed object messaging
  - DCOM apps strictly bound to Windows Operating system
  - RMI bound to Java programming language
- **Firewall traversal:**
  - CORBA and DCOM used non-standard ports
  - Web Services use HTTP; most firewalls allow access through port 80 (HTTP), leading to easier and dynamic collaboration
- **Complexity:**
  - Web Services: developer-friendly service system
  - Use open, text-based standards, which allow components written in different languages and for different platforms to communicate
  - Implemented incrementally, rather than all at once which lessens the cost and reduces the organisational disruption from an abrupt switch in technologies

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES



Now, if we look at it tries to address three major component interoperability, like lack of interoperability standard in distributed object messaging. So, interoperability between two software; between two parties.

Firewall traversal. Now, as much as the web service is piggyback on different Internet or TCP/IP based protocol, primarily http protocol. So, it has a it can basically go over the firewall most of our firewalls are http at port 80 are allowed.

So, web service can still work on the things. So, CORBA and DCOM used nonstandard ports web service is use mostly the http. So, it is not only the not only http, it can use others, but primarily http most firewalls allows port 80 http leading to easier dynamic collaboration that is one major aspects and complexity web service is it is a more of a developer friendly service system correct. So, it is the much easier to develop. So, use open text based standards like one of them is XML which allows components written in different language or different platform to communicate, right.

So, this is important. So, and it can be implemented incrementally right not on the day one everything has to be done are other deployment also can be done in incrementally, rather than all at once which lessens the cost and reduces organizational disruption

from the abrupt switch to the technologies, right. So, these are the different aspects another major aspect is still organization can run that legacy software and tools if you have proper web service interface for the external world.

(Refer Slide Time: 14:47)

The slide has a dark blue header and footer. The main content area is yellow. The title 'Web Service: Definition Revisited' is in bold red font. The list below is in black font with bullet points.

**Web Service: Definition Revisited**

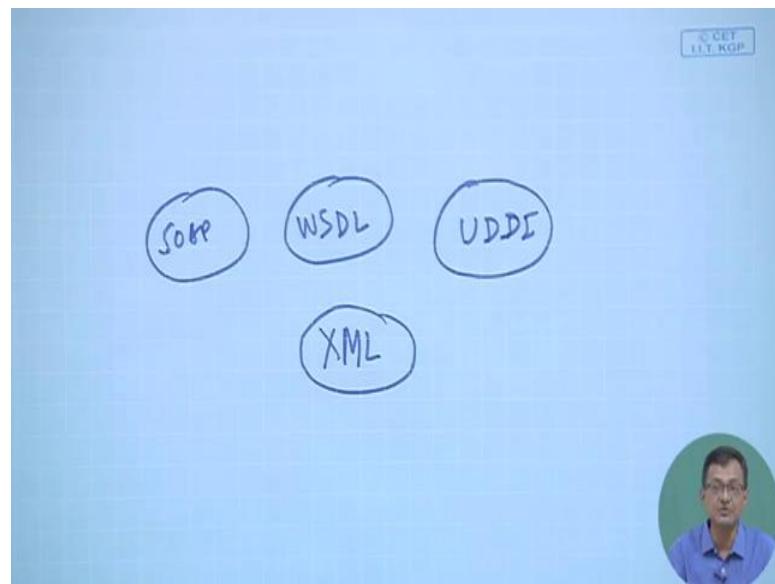
- An application component that:
  - Communicates via open protocols (HTTP, SMTP, etc.)
  - Processes XML messages framed using SOAP
  - Describes its messages using XML Schema
  - Provides an endpoint description using WSDL
  - Can be discovered using UDDI

IT Kharagpur | NPTEL ONLINE CERTIFICATION COURSES

So, if we just try to again sum up that communication via open standard like it can be HTTP it can be SMTP or any type of any other TCP/IP application layer protocol meant be there. Process of XML messages framed using SOAP, we will come to that what is SOAP and. So, it is a it is primarily XML based messaging system and one of the popular thing is the SOAP. Described its messages using XML schema, how the how my data is organized; I can basically describe is using a XML schema. Provides an endpoint description using WSDL will again see what is web service description language.

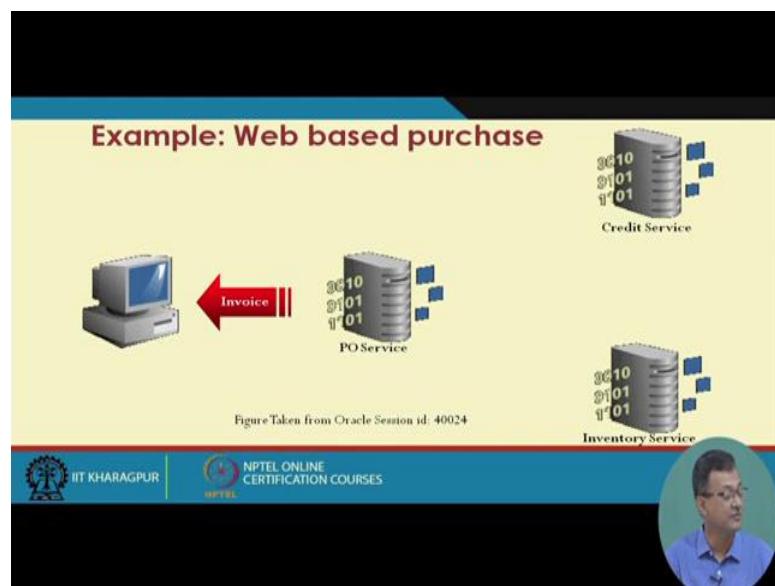
So, where my service is launched how it is configured, I we can do at a WSDL and basically I can publish and discover using a UDDI. So, my web service description discovery and integration it is facilitated by the UDDI. So, one way of implementing is that we have XML as the base.

(Refer Slide Time: 16:00)



And we have three major component SOAP, WSDL, UDDI, right. So, this is the things all are W3C complaint and use the de facto XML based things, will just look into that how things works.

(Refer Slide Time: 16:29)



So, purchase one example. So, as if purchase order goes then credit check, reserve inventory, credit response, inventory response, consolidate result and return to the invoice right. So, there are different parties like credit service, inventory service, PO services and so and so forth and those are being can be choreographed to service the

things. It is taken from a resource, but what it tries to do so that, I can provide different type of services, I can compose it and create a larger application bringing different type of application into the things. This application when choreographs to other type of things or integrated in other fashion; can give some other type of services.

(Refer Slide Time: 17:28)

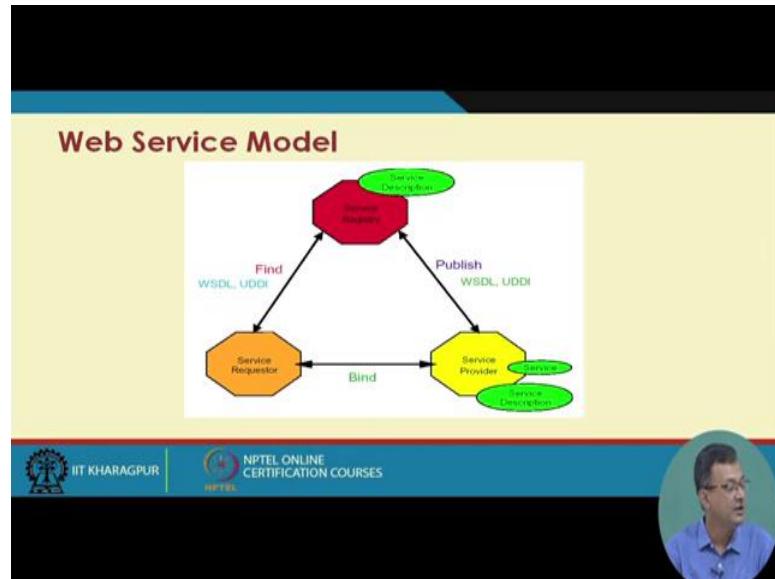
**Service Oriented Architecture (SOA)**

- IBM has created a model to show Web services interactions which is referred to as a **Service-Oriented Architecture (SOA)** consisting of relationships between three entities:
  - A service provider;
  - A service requestor;
  - A service broker
- IBM's SOA is a generic model describing service collaboration, not just specific to Web services.
  - See: <http://www-106.ibm.com/developerworks/webservices/>

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | NPTEL

Now looking at this whole thing, it has evolved to another this overall architecture what is the web service oriented architecture SOA. So, IBM created a model to show web service interaction, which is referred to as service oriented architecture consisting of a relationship between three entities. So, with this basic philosophy we have three entities; one is a service provider, one is a service requestor or service consumer and there is a service broker right or there is something which allows this consumer and provider, requestor or consumer and the broker and to talk to each other to find where the services are there, etcetera.

(Refer Slide Time: 18:13)



So, if we look at this figure a very popular figure, available in you will find in different literature.

So, I have the service requestor of the consumer which wants to get the services to the; from the service provider, but the how the service requestor find that where it services will be there? So, there is a registry where it can find out right that which type of services are there. If I try to look at a analogy those who have seen our, because now a days this is more or less obsolete, our telephone directory right which comes from telephone exchange. So, there is to be things like at the beginning that yellow pages, white pages and so and so forth. So, where you find that where to find what? Rather the directory itself is tell you that by how to search for a particular thing; like suppose you are looking for plumbing or base or something.

So, you go and lift through the thing etcetera. So, it acts as a directory service right or registry service. So, that type of things is here also require which basically have a registry service. So, it is there are service description out here, the service request are finds the this it a required service from this registry service using that QSDL and UDDI and then basically bind with the service provider it goes for it can find more than one service provider and so and so forth.

A service provider once a new service is launched on when say service is updated it is basically it publish the service in the registry, so that prospective buyer or prospective consumer.

Consumer requestor can find that where that these types of services are there. So, you can find that this is a triangle where different component works and if anything is developed based on this type of things what we say that it is a service driven or service oriented architecture.

(Refer Slide Time: 20:24)

**Web Service Model (contd...)**

- Roles in Web Service architecture
  - Service provider
    - Owner of the service
    - Platform that hosts access to the service
  - Service requestor
    - Business that requires certain functions to be satisfied
    - Application looking for and invoking an interaction with a service
  - Service registry
    - Searchable registry of service descriptions where service providers publish their service descriptions

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES



So, if we look at that web service models that the roles of service architecture this service provider owner of the service, platform that hosts access to the things, service requestor business that requires or the consumer which wants consume, and service registry searchable registry of service description where service provider publish their service description.

(Refer Slide Time: 20:48)

**Web Service Model (contd...)**

- Operations in a Web Service Architecture
  - Publish
    - Service descriptions need to be published in order for service requestor to find them
  - Find
    - Service requestor retrieves a service description directly or queries the service registry for the service required
  - Bind
    - Service requestor invokes or initiates an interaction with the service at runtime

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES



So, three major operations in the service web service architecture, one is publish to publish the service provider things, find the consumer, finds the things consumer find the things and bind once is find and it finally, binds with the service consumer with the provider.

(Refer Slide Time: 21:07)

**Web Service Components**

- **XML** – eXtensible Markup Language
  - A uniform data representation and exchange mechanism.
- **SOAP** – Simple Object Access Protocol
  - A standard way for communication.
- **WSDL** – Web Services Description Language
  - A standard meta language to described the services offered.
- **UDDI** – Universal Description, Discovery and Integration specification
  - A mechanism to register and locate WS based application.

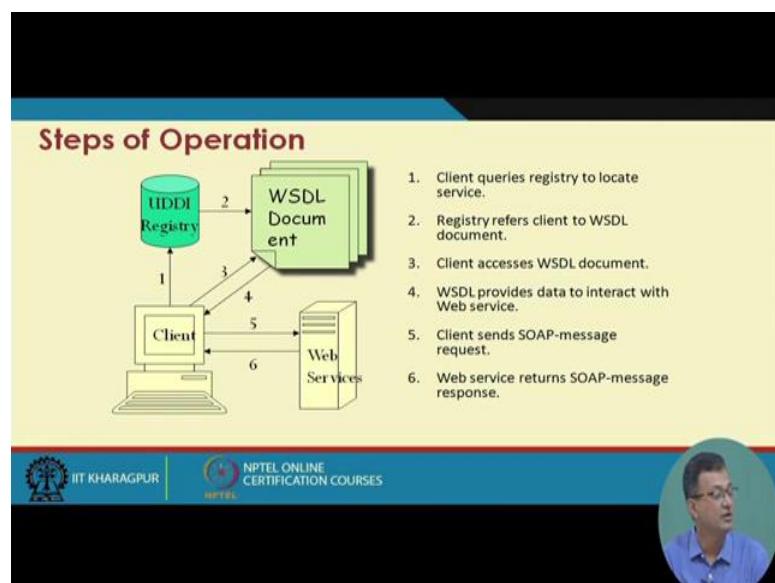
IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES



And one of the popular component is there, there are others also that is one is the which are the component one is the XML, extensible markup language. I believe all of you know if some of if you if not you should go through any standard book or even W3C

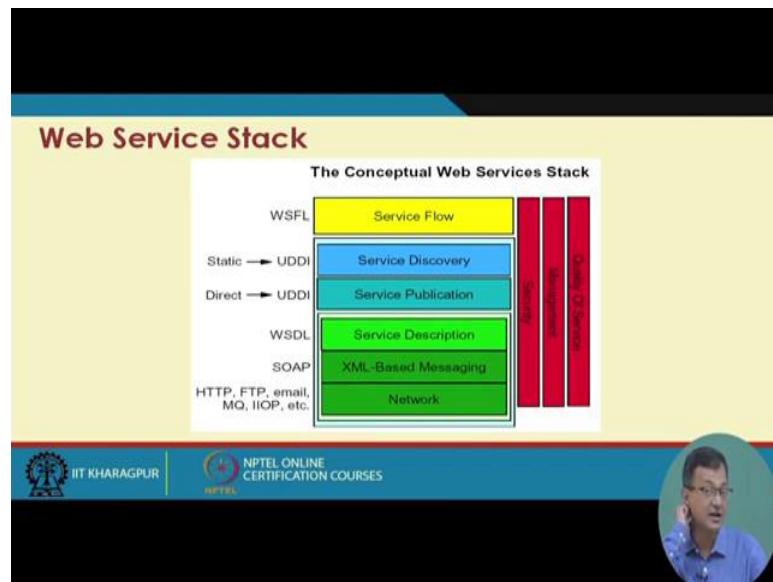
tutorial swearing find a good tutorial on XML. So, its uniform data representation and exchange mechanisms it provides. SOAP, Simple object access protocol, right a standard way for communications over using again XML. WSDL, web services description language, uses XML and it is a standard meta language to describe the services offered. And UDDI, universal description discovery and integration specification, its provide its helps in building of the registry service and a mechanism to register and locate web service applications.

(Refer Slide Time: 22:09)



So, we can look at other way as we have seen that the client goes for the client queries the registry to locate services, registry refers to WSDL document where the description is there, the client access WSDL document provides data to interact with the web services and client send SOAP message to the SOAP message request to the provider, and service returns from the SOAP message from the response.

(Refer Slide Time: 22:42)



So, that is someone binding the things and these are the different components this underlining thing is the it works on the internetworking or the our standard network protocol TCP, IP or OSI. So, this is the backbone over that SOAP messaging the description the UDDI; UDDI can be a static publication or it can be a dynamic publication and then there are other things like WSFL that is flow management and type of things.

So, for other type of aspects and there are three other component, it goes and in an one is the quality of service that what sort of service, they are giving a QS management issues that how whole thing can we manage, and then the security aspects like in doing. So, whether there is a security breach how to things, what will be the security policies, whom to trust and all those components will be there.

(Refer Slide Time: 23:33)

**XML**

- Developed from Standard Generalized Markup Method (SGML)
- Widely supported by W3C
- Essential characteristic is the separation of content from presentation
- Designed to describe **data**
- XML document can optionally reference a *Document Type Definition (DTD)*, also called a *Schema*
  - XML parser checks syntax
  - If an XML document adheres to the structure of the schema it is *valid*

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

XML, I am not dealing in details; it is a standard generalized method of a generalized markup language is evolved from. It is a extensible markup language, primarily used to describe data, and it helps in separation of content from the presentation and XML document can be optimally refer to a by a DTD the more popular is XSD, that XML schema definition language schema definitions so that, whether we can the scheme is defined.

(Refer Slide Time: 24:09)

**XML (contd...)**

- XML tags are not predefined
  - You must **define your own tags**.
- Enables cross-platform data communication in Web Services

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, in case of XML tags are not predefined. So, you can define your own tags unlike html enables cross platform communication in web services. So, this is a typically a XML thing like it describes.

(Refer Slide Time: 24:20)

The slide has a yellow header with the title 'XML vs HTML'. Below the title is the text 'An HTML example:' followed by a code block. The code is:

```
<html>
<body>
 <h2>John Doe</h2>
 <p>2 Backroads Lane

 New York

 045935435

 john.doe@gmail.com

 </p>
</body>
</html>
```

At the bottom of the slide, there is a footer bar with the IIT Kharagpur logo, the NPTEL logo, and the text 'NPTEL ONLINE CERTIFICATION COURSES'. On the right side of the footer, there is a circular video player showing a man speaking.

So, sorry, it is a html thing like it basically shows that how a particular person is address or personal email address etcetera the contact of a person called John does defined.

(Refer Slide Time: 24:40)

The slide has a yellow header with the title 'XML vs HTML (contd...)'. Below the title is a bulleted list: 'This will be displayed as:' followed by a code block. The code is:

```
John Doe
2 Backroads Lane
New York
045935435
John.doe@gmail.com
```

Below the code block is another bulleted list:

- HTML specifies how the document is to be displayed, and not what information is contained in the document.
- Hard for machine to extract the embedded information. Relatively easy for human.

At the bottom of the slide, there is a footer bar with the IIT Kharagpur logo, the NPTEL logo, and the text 'NPTEL ONLINE CERTIFICATION COURSES'. On the right side of the footer, there is a circular video player showing a man speaking.

(Refer Slide Time: 24:48)

The slide also features the IIT Kharagpur logo, the NPTEL Online Certification Courses logo, and a circular profile picture of a man."/>

**XML vs HTML (contd...)**

- Now look at the following:

```
<?xml version="1.0"?>
<contact>
 <name>John Doe</name>
 <address>2 Backroads Lane</address>
 <country>New York</country>
 <phone>045935435</phone>
 <email>john.doe@gmail.com</email>
</contact>
```

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

And if we use these html in standard any browser so, it will show like this, however if I want to look at the same thing as a XML. So, you see this is more of a data description, right. So, it is a name, it is a address, this is the country phone, email and type of things, right.

So, this is a XML type of things this not any presentation. So, for representing we need to do a something.

(Refer Slide Time: 25:17)

**SOAP**

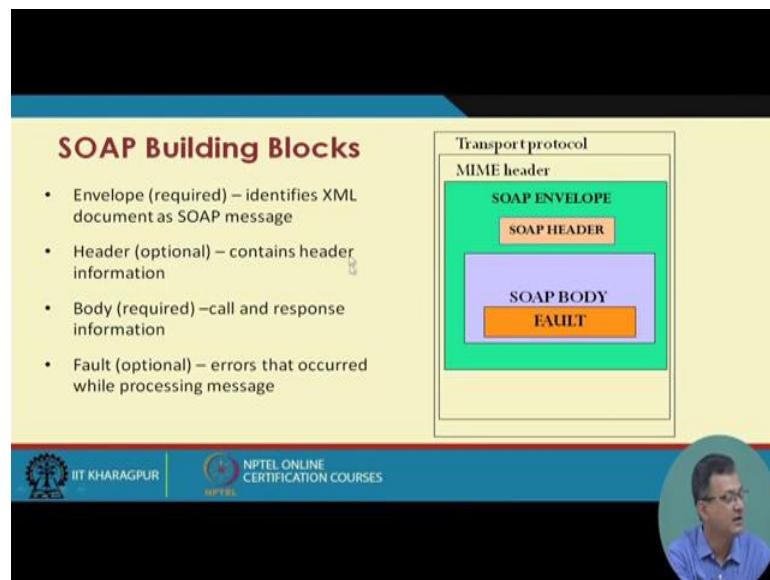
- Simple Object Access Protocol
- Format for sending messages over Internet between programs
- XML-based
- Platform and language independent
- Simple and extensible
- Uses mainly HTTP as a transport protocol
  - HTTP message contains a SOAP message as its payload section
- Stateless, one-way
  - But applications can create more complex interaction patterns

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, as we were discussing three major component one of the thing is SOAP, is a more of a messaging protocol, like simple object access protocol format for sending messages over the internet it is XML based, W3C complaint and stateless and one way uses mainly http as the what we say transport protocol. This this transport protocol should not be mix up with our TCP/IP transport protocol, this is transporting this web services.

So, here the http access the carrier protocol. So, there are different building block I am not going to the nitty-gritty of the things, how a SOAP building blocks are there.

(Refer Slide Time: 25:51)



So, transport protocol is over all envelope, then the MIME header is there then SOAP envelop SOAP header and SOAP body and there are fault and deification scenarios, but all are XML base.

(Refer Slide Time: 26:08)

**SOAP Message Structure**

- Request and Response messages
  - Request invokes a method on a remote object
  - Response returns result of running the method
- SOAP specification defines an "envelop"
  - "envelop" wraps the message itself
  - Message is a different vocabulary
  - Namespace prefix is used to distinguish the two parts

So, message structure it goes on a SOAP envelope and goes through this its piggyback on this transport, basically become a envelope and become a pay load for this transport protocol in this case http, again I am repeating this transport protocols do not be mix up with our TCP/IP or OSI protocols.

(Refer Slide Time: 26:29)

**SOAP Request**

```
POST /InStock HTTP/1.1
Host: www.stock.org
Content-Type: application/soap+xml; charset=utf-8 Content-Length: 150

<?xml version="1.0"?>
<soap:Envelope
 xmlns:soap="http://www.w3.org/2001/12/soap-envelope"
 soap:encodingStyle="http://www.w3.org/2001/12/soap-encoding">
 <soap:Body
 xmlns:m="http://www.stock.org/stock">
 <m:GetStockPrice>
 <m:StockName>IBM</m:StockName>
 </m:GetStockPrice>
 </soap:Body>
 </soap:Envelope>
```

The SOAP request goes like that if you see this is a post message in our http stuff and the SOAP message in this case its goes for a get a particular stock price and stock name particular may be the IBM and so and so forth and the response is again.

(Refer Slide Time: 26:47)

The slide title is "SOAP Response". The content shows an XML response from a web service. The XML code is as follows:

```
<?xml version="1.0"?>
<soap:Envelope xmlns:soap="http://www.w3.org/2001/12/soap-envelope"
soap:encodingStyle="http://www.w3.org/2001/12/soap-encoding">
<soap:Body xmlns:m="http://www.stock.org/stock">
<m:GetStockPriceResponse>
<m:Price>34.5</m:Price>
</m:GetStockPriceResponse>
</soap:Body>
</soap:Envelope>
```

The slide footer includes the IIT Kharagpur logo, the NPTEL logo, and a portrait of a man.

Again if you see that http response, and it response with a again a value. Since this is more of a structured way of exchanging data, it does not say out can be displayed for the display unit to have a html type of a stuff.

(Refer Slide Time: 27:02)

The slide title is "Why SOAP?". The content is a bulleted list comparing SOAP with other distributed technologies:

- Other distributed technologies failed on the Internet
  - Unix RPC – requires binary-compatible Unix implementations at each endpoint
  - CORBA – requires compatible ORBs
  - RMI – requires Java at each endpoint
  - DCOM – requires Windows at each endpoint
- SOAP is the platform-neutral choice
  - Simply an XML wire format
  - Places no restrictions on the endpoint implementation technology choices

The slide footer includes the IIT Kharagpur logo, the NPTEL logo, and a portrait of a man.

So, why soap? It is are there are other technologies which are not could not do this type of application to applications, SOAP is a platform neutral choice simple XML wire format, places no restriction on endpoint implementation of the technology, that you can run your legacy things, etcetera.

(Refer Slide Time: 27:20)

**SOAP Characteristics**

- SOAP has three major characteristics:
  - Extensibility – security and WS-routing are among the extensions under development.
  - Neutrality - SOAP can be used over any transport protocol such as HTTP, SMTP or even TCP.
  - Independent - SOAP allows for any programming model.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES



There are major three characteristics this is extensible, neutral and independent. So, this is exactly suited for our distributed applications talking to each other and type of things.

(Refer Slide Time: 27:33)

**SOAP Usage Models**

- RPC-like message exchange
  - Request message bundles up method name and parameters
  - Response message contains method return values
  - However, it isn't required by SOAP
- SOAP specification allows any kind of body content
  - Can be XML documents of any type
  - Example:
    - Send a purchase order document to the inbox of B2B partner
    - Expect to receive shipping and exceptions report as response

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES



And there are different usage models, one can be RPC like message exchange or SOAP specification allows any kind of body content so and so forth.

(Refer Slide Time: 27:42)

## SOAP Security

- SOAP uses HTTP as a transport protocol and hence can use HTTP security mainly HTTP over SSL.
- But, since SOAP can run over a number of application protocols (such as SMTP) security had to be considered.
- The *WS-Security specification* defines a complete encryption system.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

And there are other security aspects we are not going to that we will talk about these when you talk about our security.

(Refer Slide Time: 27:50)

## WSDL - Web Service Definition Language

- WSDL : XML vocabulary standard for describing Web services and their capabilities
- Contract between the XML Web service and the client
- Specifies what a request message must contain and what the response message will look like in unambiguous notation
- Defines where the service is available and what communications protocol is used to talk to the service.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

The other one is the WSDL, web service description language. So, it allows to it is again XML base W3C compliant allows to describe the language.

(Refer Slide Time: 28:02)

## WSDL Document Structure

- A WSDL document is just a simple XML document.
- It defines a web service using these major elements:
  - **port type** - The operations performed by the web service.
  - **message** - The messages used by the web service.
  - **types** - The data types used by the web service.
  - **binding** - The communication protocols used by the web service.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, it has different structure that what will be port type where the how can be defined that where the particular services will be enabled and a web things will be there; the message types of message and so and so forth. How this binding process will go on between the service provider and the consumer.

(Refer Slide Time: 28:22)

## A Sample WSDL

```
<message name="getTermRequest">
 <part name="term" type="xs:string"/>
</message>

<message name="getTermResponse">
 <part name="value" type="xs:string"/>
</message>

<portType name="glossaryTerms">
 <operation name="getTerm">
 <input message="getTermRequest"/>
 <output message="getTermResponse"/>
 </operation>
</portType>
```

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

(Refer Slide Time: 28:26)

The slide title is "Binding to SOAP". The main content is a block of XML code representing a WSDL message binding:

```
<message name="getTermRequest">
<part name="term" type="xs:string"/>
</message>
<message name="getTermResponse">
<part name="value" type="xs:string"/>
</message>
<portType name="glossaryTerms">
<operation name="getTerm">
<input message="getTermRequest"/>
<output message="getTermResponse"/>
</operation>
</portType>
<binding type="glossaryTerms" name="b1">
<soap:binding style="document"
transport="http://schemas.xmlsoap.org/soap/http" />
<operations>
<operation
soapAction="http://example.com/getTerm">
<input>
<soap:body use="literal"/>
</input>
<output>
<soap:body use="literal"/>
</output>
</operation>
</bindings>
```

The footer includes the IIT Kharagpur logo, NPTEL Online Certification Courses logo, and a portrait of a man.

So, it is a sample WSDL message, it is a binding see that it is a SOAP message with the WSDL the SOAP message is bind with these things; that means, the description over that the messaging of the SOAP message how the data is transferred over that things is bind between the source destination or any like from requestor to the registry to the consumer and so and so forth.

(Refer Slide Time: 28:52)

The slide title is "UDDI - Universal Description, Discovery, and Integration". The content is a bulleted list:

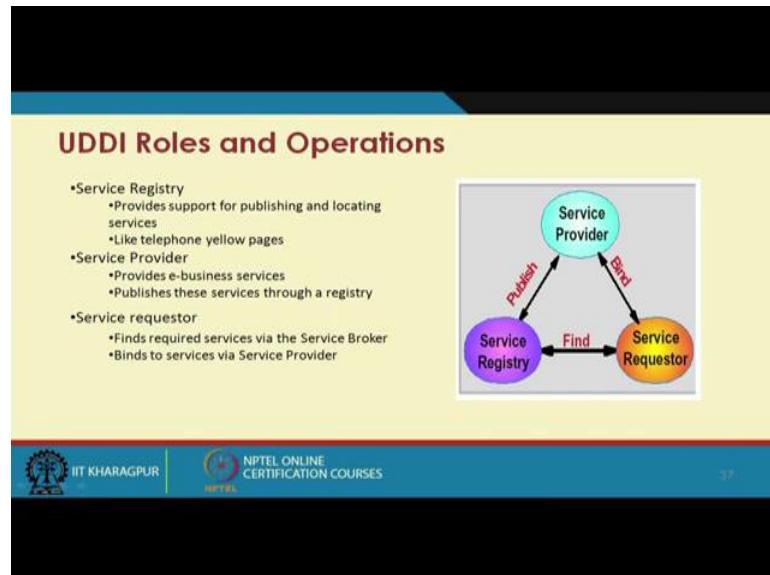
- A framework to define XML-based registries
- Registries are repositories that contain documents that describe business data and also provide search capabilities and programmatic access to remote applications
- Businesses can publish information about themselves and the services they offer

The footer includes the IIT Kharagpur logo, NPTEL Online Certification Courses logo, and a portrait of a man.

Finally we have the UDDI, universal description discovery and integration, it is a registry service right. So, a frame work to define XML based registries. So, that all these

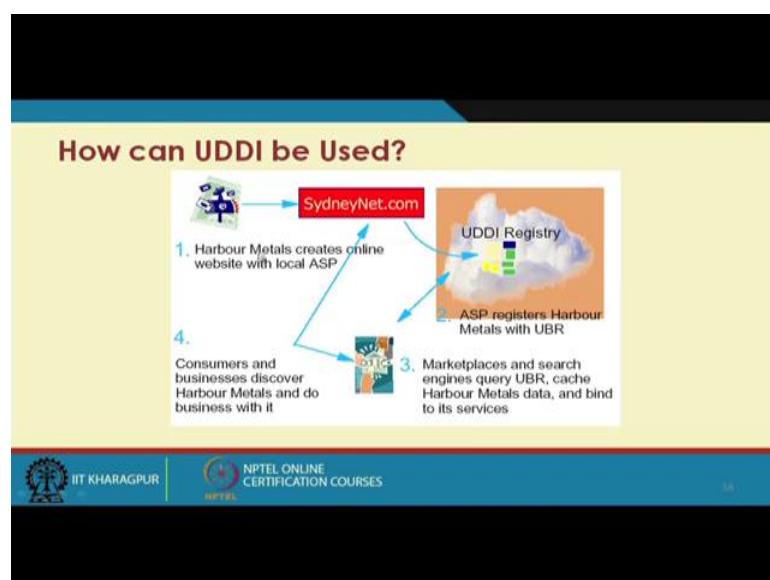
metadata informations are there, where from the this consumer or the requestor look for that particular bases and how to binding with that particular provider and so and so forth.

(Refer Slide Time: 29:13)



So, it is service provider registry and requestor, just if you remember the previous couple of slides before. So, you little bit what we say that orientation is different. So, the registries the service provider publish it, requestor finds it and binds with things. So, it plays extremely important role for keeping the whole things working together.

(Refer Slide Time: 29:41)



So, if you the same thing, it basically publish another thing the consumer search and go on binding with the provider.

(Refer Slide Time: 29:56)

### UDDI Benefits

- Making it possible to discover the right business from the millions currently online
- Defining how to enable commerce once the preferred business is discovered
- Reaching new customers and increasing access to current customers
- Expanding offerings and extending market reach

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

There are several benefits any registry service gives a very benefits of keeping the data in a particular format. So, there it can be search, making it possible discover right business from the millions of currently online.

So, you can finding how to enable say connectivity with the preferred business and so and so forth.

(Refer Slide Time: 30:19)

### UDDI Benefits

- Making it possible to discover the right business from the millions currently online
- Defining how to enable commerce once the preferred business is discovered
- Reaching new customers and increasing access to current customers
- Expanding offerings and extending market reach

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

We will not go to the detail of the things this security will work. So, in security becomes a another important aspect of soap, a aspect of web services it works on a vertical line as we have seen in our previous some of the one of the previous figure, and it has different components right like web service policy, trust, privacy, secure conversation, federation and authorization and different type of components. So, this becomes a integral one there are other things like management and QoS which also works.

So, what we are try to see in this particular lecture is that this web services and service oriented architecture, is a plays a important role in setting up this cloud. The whole cloud process if you see that IaaS, PaaS or SaaS say XaaS or anything as a service. So, it its basic building block is the web services the all this phenomena of web services of publish, bind and find bind and publish find and bind, this also is true for the things whenever we have cloud services. So, I need to basically look at the, who is the service provider and consumer, I need to know that where the service is launched and type of things, right.

So, this is a extended in a appropriate way to realize this cloud services. So, it plays; this development of web services and service oriented architecture, has played a important role in bringing this cloud computing as a viable things. So, we will stop here today and we will continue in our future lecture with other aspects of cloud computing.

Thank you.

**Cloud Computing**  
**Prof. Soumya Kanti Ghosh**  
**Department of computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**

**Lecture – 11**  
**Service Level Agreement (SLA)**

Hello, welcome to this lecture on cloud computing. Today we will be discussing one of the important topic of cloud computing or one of the major aspect of making this cloud computing to realize this is the service level agreement, right. Whenever a consumer and a producer or a service provider and a service consumer want to exchange any services or what we say meter services which involves some costing, there is a there should be some agreement on the things, right. So, the agreement involve the pricing factor, agreement involve the service availability factor, the agreement may involve say for as other quality factors.

Now, this is very tricky issue and because a organization or even individual is switching for his own personal or own proprietary computing paradigm to some cloud taking the service from some service provider, thinking that you will get same type of reliability and other reliability same type of performance same type of security what he was having in his own premises, right. The system said it been in own control.

Now, in order to that he need to have some agreement on the things. So, as such there is a still today till today they we do not have any very standard format of doing that. What you usually face whenever we are purchasing any services or any taking any VM or any cloud provider what we are usually do we basically sign off something right. We say that these are the terms and condition we agree. So, some sort of agreement will be there. So, we want to see that what are the different parameters to be taken into consideration and how this service level agreement typically look like, what are it is different components and at the end some of the popular cloud service provider. What sort of SLA or what sort of SLA parameters they are providing that we will we will see in this particular lecture today.

(Refer Slide Time: 02:42)

**What is Service Level Agreement?**

- A formal contract between a Service Provider (SP) and a Service Consumer (SC)
- SLA: foundation of the consumer's trust in the provider
- Purpose : to define a formal basis for performance and availability the SP guarantees to deliver
- SLA contains Service Level Objectives (SLOs)
  - Objectively measurable conditions for the service
  - SLA & SLO: basis of selection of cloud provider

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, if we look at what is what is SLA or service level agreement a formal contract between the service provider and the service consumer. So, that is the thing SLA is the foundation sometimes known as the foundation or sometimes refer the foundation of the consumer trust in the provider, right. Whenever I purchase some service the first thing come as how much I can trust to the provider. So, this may gives me a formal way of trusting the thing that if the SLA is there. Not only that this may help me this parameters to compare between one provider with the other provider whom to I will take the service.

So, purpose is to define a formal basis for performing and availability of service provider's guarantees to the to delivery, I need to a guarantee to the delivery. If I require a uptime of say 95 percent. So, there should be guarantee of delivery of 95 percent, right. So, I want those service provider which guarantees that I am more than 95 percent uptime it provides, right.

So, SLA now if you see SLA is a broad term right. So, it has to have different objectives, right. So, that what we say service level objectives. So, what this SLOs does? Objectively measurable condition for services right. SLA is the agreement has different components and this these are objectively measureable, I say that performance should be 95 percent. So, it is it is cannot be verbal. So, somewhere objectively I should measure the thing. So, this So, SLA constitute of number of objective basis and these later on we

will see these objectives can be calculate from different parameters right or system level parameters which I can calculate. So, these are these are objectives which are there.

Now, interestingly the objective may vary from the consumer to consumer. Like even a consumer if a service consumer like a academic institution like us, may have different type of objective or may have not totally different I say may have different type of focus on the objectives then a consumer like may be a financial organization or a software company so, they have a different type of strategy level objective. Like I can say that my uptime maybe I can have 95 percent; however, I may require that data persistence to be much higher, right.

Some other organization say that my uptime is typically more than 99 percent it cannot be compromised less than 99 percent; however, my I always take a backup of the data. So, I may not always require a backup facility and objective etcetera right. So, that can be other things I can have my objectives varies or varying over time I say during peak hours my availability should be more than 99 percent; whereas, during of peak hours my availability requirement of availability may be more than 90 percent, right. So, these things are somewhere objectively measurable thing. So, this components should come into play.

(Refer Slide Time: 05:59)

**SLA Contents**

- A set of services which the provider will deliver
- A complete, specific definition of each service
- The responsibilities of the provider and the consumer
- A set of metrics to measure whether the provider is offering the services as guaranteed
- An auditing mechanism to monitor the services
- The remedies available to the consumer and the provider if the terms are not satisfied
- How the SLA will change over time

 IIT KHARAGPUR |  NPTEL ONLINE CERTIFICATION COURSES

So, what are the different SLA typical contents we are looking for? So, it is a set of services which provider will delivery. So, it should specify the set of services the

provider will delivery. I would like to again reiterate. So, these are these SLA's are valid for different type of service provisioning, like it can be IaaS type of services where the SLA he want to have you can have pas type of services you can have SaaS type of services or any type of other type of things like may be data services etcetera. So, this SLA's correspond to that type of with that respect right. So, SLA for particular which is a software as a service type of module and so and so forth.

So, it should be complete specific definition of ease service right. So, it is been complete specific definition, responsibilities of provider and the consumer to respect these SLA. A set of metric to measure whether the provider is offering the services as guaranteed like how do I know that whether the thing. So, some metrics should be there that which we shows that that it is providing the whatever it is guaranteed.

A auditing mechanisms right. Interesting a auditing mechanism to monitor the services. So, I need to have a some auditing mechanism to monitor the services right. So, it may be a third party auditing like I purchase as a service consumer purchased from service provider, but a third party auditor basically sees that where that there is a violation of the services etcetera right. So, these are the things which are to be monitored so, that is auditing is required. The remedy is available to the consumer and the provider these are terms are not satisfied, that is another thing is there, what should be that some sort of a payback mechanism or what should be the remedies if this is not. Like if I say if it I require a uptime of 95 percent and if the provider fails to give me 95 percent for sometime. So, whether he should pay me back or provide some other incentive to do that right. So, that should be some mechanisms where things are violated right. So, there should be.

And if you if we see that for a our day to day activity organizational activity there are this type of agreements MoUs etcetera, where there are things for defining that objectively there are remedies for violating those things what will happen. And so and so forth there are penalty what we say penalty of not providing the guaranteed service. If I I say that I have give a guaranteed service and it is not provided then what I should pay give some penalty or I need to give some other incentive to do to compensate that that should be also specified, right.

What if there is not guaranteed etcetera and how whether the SLA will change over time right. It may change over time during the particular day of the things it may change over time over different days of the time and etcetera, etcetera. So, there is a whether the SLA will change over time that is another. So, whether it is a temporally changing phenomena, right.

(Refer Slide Time: 09:04)

The slide has a blue header bar. The main content area has a yellow background. At the top left of the yellow area, the title 'Web Service SLA' is displayed in red. Below the title is a bulleted list of points. At the bottom of the slide, there is a footer bar with the IIT Kharagpur logo on the left and the NPTEL logo on the right. On the right side of the footer bar, there is a circular profile picture of a man.

- WS-Agreement
  - XML-based language and protocol for negotiating, establishing, and managing service agreements at runtime
  - Specify the nature of agreement template
  - Facilitates in discovering compatible providers
  - Interaction: request-response
  - SLA violation : dynamically managed and verified
- WSLA (Web Service Level Agreement Framework)
  - Formal XML-schema based language to express SLA and a runtime interpreter
  - Measure and monitor QoS parameters and report violations
  - Lack of formal definitions for semantics of metrics

Now, there is a concept of web services, I believe most of you are having a thing service oriented architecture and web services, which are one of the what we say prime over for coming up of this cloud services for benefit of all, I will take one short lecture on later on in subsequent talks one web services and service oriented architecture those who are not a custom, but I believe that most of you are used to it and type of things right. So, there is a in case of web services where service provider and consumer there is a concept of they communicate with each other.

There is also a agreement thing right. Or what we say web service SLA. So, web service SLA there are component like one is that web service agreement. The XML based language and protocol negotiating establishing and managing the service agreement at the run time. So, it is a de-facto XML based things actually the whole web services things and so on basic foundation of XML. Specify, specify the nature of agreement template. So, what should be the agreed upon format of the template of the things.

Facilitates in discovering compatible providers right. So, I there can be more than one provider. So, it should be able to see that where the compatible providers.

Interaction usually request response right. And SLA violation dynamically managed and verified. So, if there is a SLA violation this need to be dynamically managed and verified, and necessary action to be initiated of there is a violation of the SLA. There is a web service level agreement framework there is a concept called WSLA web service, level agreement framework. So, there is a framework for WS services formal xml schema based languages to express SLA and runtime interpreter.

Measure and monitor QoS parameter, quality of service parameters and report violations if any, lack of formal definitions for semantics of the metrics right. So, there are there is what we are talking about is more of a syntactic way of looking at there are few issues of semantics of looking at, but there are less standardization of what the semantics of the whole thing means like whether the whole this parameter still something like I can say that whether this parameter tells that the system is performing well is going down and type of things, or I am expecting a failure something or this type of things may happen when this type of things. So, there are different underlining semantics which is which is did to be formally defined or standardized.

So, if you look at this WSLA, this as the cloud evolve from this web service and service oriented architecture keeping those the framework in mind. So obviously, this SLA's are etcetera are also having a relationship with them with the things, right.

(Refer Slide Time: 12:12)

**Difference between Cloud SLA and Web Service SLA**

- QoS Parameters :
  - Traditional Web Service : response time, SLA violation rate for reliability, availability, cost of service, etc.
  - Cloud computing : QoS related to security, privacy, trust, management, etc.
- Automation :
  - Traditional Web Service : SLA negotiation, provisioning, service delivery, monitoring are not automated.
  - Cloud computing : SLA automation is required for highly dynamic and scalable service consumption
- Resource Allocation :
  - Traditional Web Service : UDDI (Universal Description Discovery and Integration) for advertising and discovering between web services
  - Cloud computing : resources are allocated and distributed globally without any central directory

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, but if you if we look and try to see that little nitty-gritty of the what are the differences between the things mainly divide mainly trying to look at into three major components, one is QoS parameters that quality of service parameters, one is automation and another is resource allocation. So, these are the three thing what we are looking there.

So, in case of QoS parameter in case of a traditional web services. So, what we see response time, SLA violation for SLA violation rate for reliability, availability, cost of services etcetera. So, in traditional things there is a response time plays a vital role and there are SLA violation rates for reliability, availability and cost of services etcetera. So, if there are violation then we with there are what should be that there is rate that what we things will be there.

In case of a cloud, QoS is related to security, privacy, trust management etcetera, have more importance right. So, the basic way of handling may be there, but in case of a cloud we are more concerned about security, more concerned about privacy, trust overall management and so and so forth. If we look at the automation point of view traditional web services, SLA negotiation provisioning service delivery monitoring are not automated right. In case of a cloud SLA automation is required for high dynamic and scalable services. Like I can scale up scale down and thing. So, I dynamic and scalable services may require this monitoring.

Resource allocation traditional web services UDDI, that is universal description discovery and integration that is a UDDI is one of the protocol which is very prominent in web services. It provides a registry services and we do use those type of things in cloud also for advertising and discovering between the web services. So, this UDDI is there in case of a cloud resources are allocated and distributed globally without any central directory perceive right.

(Refer Slide Time: 14:34)

**Types of SLA**

- Present market place features two types of SLAs :
  - Off-the-shelf SLA or non-negotiable SLA or Direct SLA
    - Non-conducive for mission-critical data or applications
    - Provider creates the SLA template and define all criteria viz. contract period, billing, response time, availability, etc.
    - Followed by the present day state-of-the-art clouds.
  - Negotiable SLA
    - Negotiation via external agent
    - Negotiation via multiple external agents

So, ideally it is a distributed things and there is a different mechanism of knowing that who is having what. So, if we look at the different types of SLA's right. What are the types of SLA's, one is the off-the-shelf or non-negotiable SLA or sometimes known as direct SLA right. So, one type of things off-the-shelf. So, non conductive not conducive for mission critical data or applications. So, it may not be if n would be suitable if you have a if you have a very mission critical application you want to do this thing and so and so forth.

Provider creates a SLA template and define all criteria that is contract period billing responsible, response time availability etcetera provides, say template and all whatever the things is there. Followed by the present day state of art clouds. So, the this specially this public cloud follow this type of thing. So, whenever you want to buy some services from the public cloud it will it will go you this. So, with this form with other terms and condition. So, you need to agree on the things to work on the things right.

Whereas negotiable SLA; that means, you negotiate and find out that what things will be there, negotiable via external agent. So, I may have there may be external agents which negotiates between the provider and consumer. And there can be negotiable via multiple external agent, if there you are buying multiple services and then you amalgamate those services to achieve one particular thing then it can be multiple layers.

(Refer Slide Time: 16:11)

**Service Level Objectives (SLOs)**

- Objectively measurable conditions for the service
- Encompasses multiple QoS parameters viz. availability, serviceability, billing, penalties, throughput, response time, or quality
- Example :
  - “**Availability** of a service X is 99.9%”
  - “**Response time** of a database query Q is between 3 to 5 seconds”
  - “**Throughput** of a server S at peak load time is 0.875”

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, it can be. So, it can be either off-the-shelf that is a standard or whatever is provided or negotiable. Usually what happened whenever we are having small requirement want to do a something which is there it in our own means some very small scale operations then we go for this type off-the-shelf staff. Whenever we have a large requirement like he want to have whole organization process into the thing and etcetera then we want to look at the more have a negotiable SLA. So, want a special rate it is as good as you are booking a particular hotel or transport for one or two percent then you go for that whatever is available, but if you are buying the whole booking the whole hotel or the booking a transport bus or something then you have a negotiation to do that that what, we have also negotiate to have some fine tune the agreement process.

So now what you are discussing about. So, we have SLA's right. Service level agreement and this SLOs are contributing to from this SLA's right. So, service level objectives. So, objectively measurable condition for service. So, that is one point, encompasses multiple QoS parameters. So, SLOs can have a different QoS parameter. Like availability,

serviceability, billing, throughput, response time, quality etcetera. So, these are somewhere other need to be measured. So, for example, may be availability of service is 99.9 percent. So, that is a thing I can say that response time of a database query should be between 3 to 5 seconds right. So, that is a response time I am looking for throughput of a server particular server at peak time should be something 0.875, right. So, that may be another point of looking at the thing. So, these are this can be different type of SLOs.

(Refer Slide Time: 18:15)

**Service Level Management**

- Monitoring and measuring performance of services based on SLOs
- Provider perspective :
  - Make decisions based on business objectives and technical realities
- Consumer perspective :
  - Decisions about how to use cloud services

IT Kharagpur | NPTEL ONLINE CERTIFICATION COURSES



So, service level management. So, I have agreement I have SLOs, now I have to manage this whole scenarios. So, monitoring and managing the performance services based on the SLOs. So, SLOs are reporting that what are the different values etcetera. Provider perspective make decision based on the business objective and technical realities right. So, it has a business objective it is has a business goal and technical realities how much things the provider is having at the back end that is also important. From the consumer perspective decision about how to use cloud services, right.

So, it is more that whether it is suitable for my organization where I suitable for my personal dream need, etcetera, then that way I measure.

(Refer Slide Time: 19:00)

### Considerations for SLA

- **Business Level Objectives:** Consumers should know *why* they are using cloud services before they decide *how* to use cloud computing.
- **Responsibilities of the Provider and Consumer:** The balance of responsibilities between providers and consumers will vary according to the type of service.
- **Business Continuity and Disaster Recovery:** Consumers should ensure their cloud providers have adequate protection in case of a disaster.
- **System Redundancy:** Many cloud providers deliver their services via massively redundant systems. Those systems are designed so that even if hard drives or network connections or servers fail, consumers will not experience any outages.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, there are several considerations for SLA's. Few of them are as follows. One is our business level objectives. Responsibilities of the provider and consumer is important; that means, the balance of responsibility between the provider and consumer will vary according to the type of services, right.

So, there are some services where the provider's responsibility is much, much higher, whether there is a service which also requires the consumer responsibility right. Like I put some data and get output of it I may have a things that I can the consumer, your provider can basically accept data at the rate of say  $x$  megabits per second. And then if I send some data as I  $x'$  where  $x'$  is more than the  $x$  then there may be problem of things problem of overflow of data and type of things then, which may internally violate other SLA's of the thing SLA of the provider.

So, there is a responsibility for both the try to check and do and there are auditing to maintain that they all are for you. Business continuity and disaster recovery another important aspect for the consumer should ensure that cloud providers have adequate protection in case of a disaster right. It may be disaster natural, manmade, system failure etcetera. So, it should have been business there are system redundancy, so many cloud provider deliver their services via massively redundant systems right. So, that you can get guarantee.

(Refer Slide Time: 20:34)

**Considerations for SLA (contd...)**

- **Maintenance:** Maintenance of cloud infrastructure affects any kind of cloud offerings (applicable to both software and hardware)
- **Location of Data:** If a cloud service provider promises to enforce data location regulations, the consumer must be able to audit the provider to prove that regulations are being followed.
- **Seizure of Data:** If law enforcement targets the data and applications associated with a particular consumer, the multi-tenant nature of cloud computing makes it likely that other consumers will be affected. Therefore, the consumer should consider using a third-party to keep backups of their data
- **Failure of the Provider:** Consumers should consider the financial health of their provider and make contingency plans. The provider's policies of handling data and applications of a consumer whose account is delinquent or under dispute are to be considered.
- **Jurisdiction:** Consumers should understand the laws that apply to any cloud providers they consider.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, these are some of the things which are interlinked. There are other issues like maintains maintenance of the cloud infrastructure effects any kind of cloud offering applicable to both software and hardware. So, maintenance is a big factor. Location of the data if the cloud provider premises to enforce premises to enforce data location regulation the consumer must be able to audit the provider to prove the regulations are being followed.

Like and say that I can it may be So happen that IIT Kharagpur for example, say that all my data if in cloud should be residing within the jurisdiction of India right. I do not care how So, there should be a where to I verify that the data is not in some other country or some other land. Seizure of data, if low enforcement targets that data and application associated with the particular consumer the multi tenant nature of the cloud computing makes it likely that the other consumer will be effected also.

Like if suppose a consumer c one it is data the law enforcement or department want to seize then if it is residing multi tenant with some other consumer then the it may want to seize the whole at this then other things will be there. So, there is a issue which need to be there failure of the provider if in case of a failure of the provider.

(Refer Slide Time: 21:56)

**SLA Requirements**

- **Security:** Cloud consumer must understand the controls and federation patterns necessary to meet the security requirements. Providers must understand what they should deliver to enable the appropriate controls and federation patterns.
- **Data Encryption:** Details of encryption and access control policies.
- **Privacy:** Isolation of customer data in a multi-tenant environment.
- **Data Retention and Deletion:** Some cloud providers have legal requirements of retaining data even if it has been deleted by the consumer. Hence, they must be able to prove their compliance with these policies.
- **Hardware Erasure and Destruction:** Provider requires to zero out the memory if a consumer powers off the VM or even zero out the platters of a disk, if it is to be disposed or recycled.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES



What should be there and jurisdiction that any type of litigation were where it will be addressed. There are other SLA requirements like one definitely security there is a big issue of data encryption if I encrypt the data how this key management will be there where the encrypted data will be residing where the key will be residing how the key will be communicate and etcetera.

Privacy issue isolation of the consumer data in a multi tenant environment how to isolate a consumer data and services in a multi tenant environment. Data retention and deletion right. Say if there is a if the how long the cloud will retain the data and where it will delete and type of things. Hardware erasure or destruction. So, provider requires to 0 out the memory if consumer powers of the VM or even 0 out the platters of the disk if it is to be disposed or recycled. So, if there is a thing as that the hardware erasure or disposing of the hardware etcetera then what will be the other effect on the things. So, this things need to be there. There are several other requirements.

(Refer Slide Time: 22:57)

**SLA Requirements (Contd...)**

- **Regulatory Compliance:** If regulations are enforced on data and applications, the providers should be able to prove compliance.
- **Transparency:** For critical data and applications, providers must be proactive in notifying consumers when the terms of the SLA are breached.
- **Certification:** The provider should be responsible in proving the certification of any kind of data or applications and keeping its up-to-date.
- **Monitoring:** To eliminate the conflict of interest between the provider and the consumer, a neutral third-party organization is the best solution to monitor performance.
- **Auditability:** As the consumers are liable to any breaches that occur, it is vital that they should be able to audit provider's systems and procedures. An SLA should make it clear how and when those audits take place. Because audits are disruptive and expensive, the provider will most likely place limits and charges on them.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES



One may be one is regulatory compliance, transparency that what I am providing whether you are transparent to the consumer and the third party, certification whether provider has a certification process that it provide something which is which has a particular level of granted services and certification. Monitoring, how to how to monitor the performance of the provider that is provider should be responsible for providing certification of any kind of sorry to eliminate the conflict of interest between the provider and consumer a neutral third party organization or what we say third party monitoring agency or third party auditor is the base solution for monitor performance.

So, these are easy to tell about, but while implementing it is very difficult, because suddenly in the provider things they may not allow this third party to work on their systems etcetera right. Then auditability, as the consumer are liable to breach liable to any breaches that occur it is vital they should be able to audit provider systems and procedure right. As it will affect the consumers own business if with the providers with. And SLA should make it clear how and when those audit takes place. Because audits and disrupts audits are disruptive and expensive providers will most likely place limits on when charges on them right

So, if you want to do frequent auditing then it may basically it is a expensive both on not only on monitory terms it is on resource term also right. There may be downtime there

may be other resource requirement, etcetera. So, provider may not be interested in that very frequently so, that to be need to be properly provision.

(Refer Slide Time: 24:51)

**Key Performance Indicators (KPIs)**

- Low-level resource metrics
- Multiple *KPIs* are composed, aggregated, or converted to for high-level *SLOs*.
- Example :
  - downtime, uptime, inbytes, outbytes, packet size, etc.
- Possible mapping :
  - *Availability (A) = 1 – (downtime/uptime)*

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

There is another factor call what we component called key performance indicator right. So, it is low level resource metric right. So, multiple KPI's are composed aggregated or converted to high level SLOs. Multiple SLOs are integrated to have that SLA. So, we look at I have SLA, then SLO and then KPI, right.

(Refer Slide Time: 25:19)

SLA  
↑  
SLO  
↑  
KPI

© CCL  
IIT KGP

So, a KPIs is the lowest level metric which we have. So that means, here is a multiple KPI composed aggregated and converted to high level SLOs, like KPIs like typically like may be downtime right, uptime in bytes, out bytes, packet size etcetera right.

So, these are the different KPI there can be there. So, possible mapping like if availability is a one of the objective. Then availability is  $1 - (downtime / uptime)$  right. So, these are the components this KPIs which are very low level and directly measure measured from the system parameters. And SLOs are measured based on this KPIs, and this KPI this SLO aggregation of SLOs are these SLA.

(Refer Slide Time: 26:27)

**Industry-defined KPIs**

- Monitoring:
  - Natural questions:
    - "who should monitor the performance of the provider?"
    - "does the consumer meet its responsibilities?"
  - Solution: neutral third-party organization to perform monitoring
  - Eliminates conflicts of interest if:
    - Provider reports outage at its sole discretion
    - Consumer is responsible for an outage
- Auditability:
  - Consumer requirement:
    - Is the provider adhering to legal regulations or industry-standard
    - SLA should make it clear how and when to conduct audits

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, industry defined KPIs. So, monitoring so, natural question who should monitor the performance of the provider, does the consumer meets the responsibility solution neutral third party organization to perform this monitoring, eliminate conflicts of interest then there is issues of auditability as we have discussed.

(Refer Slide Time: 26:48)

**Metrics for Monitoring and Auditing**

- **Throughput** – How quickly the service responds
- **Availability** – Represented as a percentage of uptime for a service in a given observation period.
- **Reliability** – How often the service is available
- **Load balancing** – When elasticity kicks in (new VMs are booted or terminated, for example)
- **Durability** – How likely the data is to be lost
- **Elasticity** – The ability for a given resource to grow infinitely, with limits (the maximum amount of storage or bandwidth, for example) clearly stated
- **Linearity** – How a system performs as the load increases

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, the metric for monitoring and auditing these are the typical widely used metrics. Throughput, availability, reliability, load balancing So, when elasticity kicks in, So new VMs are booted or terminated for example, then the load balancing of the systems is an important factor, right.

Durability, how likely the data to be lost how much the, how much durable this data or services are there elasticity then linearity, how the system performs at the load increases if it is a linear increases like load increases and the system also increases the system provisioning also increases whether it is a linear craft, if it is a linear craft, then it is easy to scale-up like easy to chase this higher demand. If it is a non-linear specially exponential etcetera then it is a difficult process we will see later on.

(Refer Slide Time: 27:44)

**Metrics for Monitoring and Auditing** (Contd...)

- **Agility** – How quickly the provider responds as the consumer's resource load scales up and down
- **Automation** – What percentage of requests to the provider are handled without any human interaction
- **Customer service response times** – How quickly the provider responds to a service request. This refers to the human interactions required when something goes wrong with the on-demand, self-service aspects of the cloud.
- **Service-level violation rate** – Expressed as the mean rate of SLA violation due to infringements of the agreed warranty levels.
- **Transaction time** – Time that has elapsed from when a service is invoked till the completion of the transaction, including the delays.
- **Resolution time** – Time period between detection of a service problem and its resolution

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, there are few more metrics like agility, automation, customer service, response time, service level violation, transaction time, resolution time these are different other components.

(Refer Slide Time: 27:58)

**SLA Requirements w.r.t. Cloud Delivery Models**

Requirement	Platform as a Service	Infrastructure as a Service	Software as a Service
Data Encryption	✓	✓	✓
Privacy	✓	✓	✓
Data Retention and Archival	✓	✓	✓
Hardware Erasure and Destruction	✓	✓	✓
Regulatory Compliance	✓	✓	✓
Transparency	✓	✓	✓
Certification	✓	✓	✓
Terminology for Key Performance Indicators		✓	✓
Metrics	✓	✓	✓
Auditability	✓	✓	✓
Monitoring	✓	✓	✓
Machine-Readable SLAs		✓	

Source: "Cloud Computing User Cases White Paper" Version 4.0

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Now, if we have this chart there are different requirements at different levels right. Like IaaS may requiring something PaaS may requiring something and SaaS may be requiring, something on these different type of some a few of the requirement components.

(Refer Slide Time: 28:16)

Example Cloud SLAs			
Cloud Provider	Service	Type of Delivery Model	Service Level Agreement Guarantees
Amazon	EC2	IaaS	Availability (99.95%) with the following definitions : Service Year : 365 days of the year, Annual Percentage Uptime, Region Unavailability : no external connectivity during a five minute period, Eligible Credit Period, Service Credit
	S3	Storage-as-a-Service	Availability (99.9%) with the following definitions: Error Rate, Monthly Uptime Percentage, Service Credit
	SimpleDB	Database-as-a-Service	No specific SLA is defined and the agreement does not guarantee availability
Salesforce	CRM	PaaS	No SLA guarantees for the service provided
Google	Google App Engine	PaaS	Availability (99.9%) with the following definitions : Error Rate, Error Request, Monthly Uptime Percentage, Scheduled Maintenance, Service Credits, and SLA exclusions

 IIT KHARAGPUR |  NPTEL ONLINE CERTIFICATION COURSES



So, we will quickly look at some of these popular or example cloud providers SLA's or what they type like try to give the like Amazon EC2 if the IaaS service provider it shows that the availabilities 99.5 percent. Like service year 365 days. Annual percentage uptime region and so and so forth.

Whereas S3 which provides Amazon is to provides storage as a service, availability 99 percent with the following definition error rate monthly uptime percentage etcetera. So, those components those parameter index we can find out.

(Refer Slide Time: 28:59)

Example Cloud SLAs (contd...)			
Cloud Provider	Service	Type of Delivery Model	Service Level Agreement Guarantees
Microsoft	Microsoft Azure Compute	IaaS/PaaS	Availability (99.95%) with the following definitions : Monthly Connectivity Uptime Service Level, Monthly Role Instance Uptime Service Level, Service Credits, and SLA exclusions
	Microsoft Azure Storage	Storage-as-a-Service	Availability (99.9%) with the following definitions: Error Rate, Monthly Uptime Percentage, Total Storage Transactions, Failed Storage Transactions, Service Credit, and SLA exclusions
Zoho suite	Zoho mail, Zoho CRM, Zoho books	SaaS	Allows the user to customize the service level agreement guarantees based on : Resolution Time, Business Hours & Support Plans, and Escalation

 IIT KHARAGPUR |  NPTEL ONLINE CERTIFICATION COURSES



Similarly Google, Sales force, Microsoft provides IaaS PaaS availability 99.9 percents. With following definition monthly connectivity uptime service level and etcetera.

(Refer Slide Time: 29:12)

Cloud Provider	Service	Type of Cloud Delivery Model	Service Level Agreement Guarantees
Rackspace	Cloud Server	IaaS	Availability regarding the following: Internal Network (100%), Data Center Infrastructure (100%), Load balancers (99.9%) Performance related to service degradation: Server migration, notified 24 hours in advance, and is completed in 3 hours (maximum) Recovery Time: In case of failure, guarantee of restoration/recovery in 1 hour after the problem is identified.
Terremark	vCloud Express	IaaS	Monthly Uptime Percentage (100%) with the following definitions: Service Credit, Credit Request and Payment Procedure, and SLA exclusions

Similarly, zoho which provides SaaS, Rack space terremark. So, there are some of the things which are listed here.

(Refer Slide Time: 29:16)

Cloud Provider	Service	Type of Cloud Delivery Model	Service Level Agreement Guarantees
Nirvanix	Public, Private, Hybrid Cloud Storage	Storage-as-a-Service	Monthly Availability Percentage (99.9%) with the following definitions: Service Availability, Service Credits, Data Replication Policy, Credit Request Procedure, and SLA Exclusions

So, what we try to discuss in this particular talk is that this service level agreements in plays a vital role when we want to use want to make these cloud computing in our in practice right. I want to when I want to for my personal use, or organizational use, where

when we want to migrate from my conventional traditional system to this cloud computing thing. Then I need to look at around across these different SLA's right. What I need, what are the parameters I need and whether it is measurable by the at the provider ends and how I can guarantee that my work or my work process business process should not be affected adversely by the cloud service provider. So, with these we will stop here.

Thank you.

**Cloud Computing**  
**Prof. Soumya Kanti Ghosh**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology Kharagpur**

**Lecture – 12**  
**Economics**

Hello. So, we will continue our discussion or lectures on cloud computing. Today we will take up a topic, which is we need to look at that what is this economy behind cloud computing. Why people will go for this cloud computing type of things right. What it is not like that, we are getting a something totally new. So, it is only that the different type of application etc; now we are getting as a service. So, it is what makes this or what will make this viable. Is it always going to cloud is beneficial or where how to decide that; whether I need to go to cloud or whether I need to buy something in house and how much how to balance between in house infrastructure and provisioning of the cloud. So, in order to do that, we will try to see some basic phenomena or basic economic point which is makes cloud viable at which type of operations or which type of situation type of things.

So, we try to more try to look out, that when we organization or individual especially organizations, one switch over cloud partially or fully, what are the consideration it should keep in mind. We have seen SLAs they are there are other issues that even during cloud our initial lectures that there are some limitations, there are some issues with the cloud, but keeping all that even all those things are running fine. Whether it is economically to be on cloud always or at times or at what times and type of things and what should be the business consideration. So, we will have a brief discussion on that. So, those of you are interested in working on this line can basically now leverage on this type of work.

(Refer Slide Time: 02:20)

**Cloud Properties: Economic Viewpoint**

- Common Infrastructure
  - pooled, standardized resources, with benefits generated by statistical multiplexing.
- Location-independence
  - ubiquitous availability meeting performance requirements, with benefits deriving from latency reduction and user experience enhancement.
- Online connectivity
  - an enabler of other attributes ensuring service access. Costs and performance impacts of network architectures can be quantified using traditional methods.

**IIT KHARAGPUR** | **NPTEL ONLINE CERTIFICATION COURSES**

So, from the economic point of view, if we see that what are the different cloud properties relook at the properties one is the common infrastructure.

So, pooled standardized resources with benefits generated by statistical multiplexing right? That is important. First of all, it is a pooled standardized resources. So, at the service provider end it is a pools of resources, which are standardized how they can be integrated, etcetera, right. What is the benefit from the thing because of multiplexing tasks, right. So, statistically multiplexing benefits statistical multiplexing; that means, I have some 100 systems, right. So, it is highly underutilized even 15-20 percents are not utilized of my own workload, or from some workload. So, if I have multiple workloads, I may with the same type of system. I can go up into the things, what I am thinking that why I have purchased this some 10 or 100 systems. Because considering most of the cases, we have considering a peak workload, right.

At peak time I will be require all those things right. Say for example, IIT kharagpur in a particular department, a particular lab for a particular year, say MTech first year. The numbers of seats are say 50 for a particular department and what we expect that 50 percent we will do work on 50 individual systems; we purchase a lab we basically set up a lab of 50 systems. Adjoining we same time we have to go for power, at the same time we have to go for AC and we feel that if there is a system goes out and type of things, keep another 5, 10 percent another 5 system on a standby. So, having 55 systems, but it

may so, happen that the recruitment the number of students joined the course may be less than 50. So, I have any surplus power even they joined the courses right. This 50 system are again underutilized, not may not be utilized all the things during the lab hours. May be daily 8 hours it is utilized otherwise things are not there, right. One is looking at the peak load.

So, but whenever I consider the system individually. I have taken this processor memory etcetera thinking, this lab assignment will be leading up to that level so; that means, at the peak load, right. So, this consideration may be or at many times. I do lot of over sizing of the things.

And if I have is it becomes more critical, when I have a server, a single server which is catering to number of users and then I think that all the user will jump into the thing right and then the peak load will be there right. If those who are working in the networking you might have seen that typically 10, 24 port networks switch right. How many people can connect? 24. Even 100 mbps line it is 24 roughly,  $24 \times 100$  if we not go to the integrity of this binary thing.

So, it is approximately 2.4 gigabytes, but the switch uplink maybe 1 gigabyte. So, it is some sort of a blocking architecture, but if I give a uplink of 2.4 or 3 gigabyte then it is over provisioning, right. If it is all the people are coming statistically at the same peak time may not be there. So, it statistically we need to look at that whether it is viable to provision. Such a higher thing when you provision higher, thing it involves lot of costing and other things in to the things maintenance of the thing, there are things of adjoining other accessories, etcetera.

Even the cost of the equipment goes up so and so far. So, that is one thing, another is the location independence that is another property of cloud like ubiquitously available meeting performance requirement with benefits deriving from latency reduction and user experience in enhancement. So, this is a location independence property. So, whether it is economically, how to make use of that online connectivity as an enabler of other attributes ensuring a service cost and performance impact on the network architecture etcetera. So, I should have online always connectivity. So, there is another factor. So, these are the different factors which may force us to look at the different things.

(Refer Slide Time: 06:55)

**Cloud Properties: Economic Viewpoint (contd...)**

- Utility pricing
  - usage-sensitive or pay-per-use pricing, with benefits applying in environments with variable demand levels.
- on-Demand Resources
  - scalable, elastic resources provisioned and de-provisioned without delay or costs associated with change.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

There are other two direct economic factors, one is the utility pricing right. I pay as you go model, I pay per unit things like electricity etcetera. There is another thing called on-demand resources. So, when I demand for the resources are provisioned right. So, it is on demand for the resources are provided. So, scalable, elastics resources, provision and de-provision without delay and cost associated with the change, right. So, there cannot be delay or cost associated anything. There may be cost factor, rather there should not there minimal human on managing managerial involvement. So, that as I go and go forth and things. I may resources provisioning and de-provisioning.

(Refer Slide Time: 07:43)

**Value of Common Infrastructure**

- Economies of scale
  - Reduced overhead costs
  - Buyer power through volume purchasing
- Statistics of Scale
  - For infrastructure built to peak requirements:
    - Multiplexing demand → higher utilization
    - Lower cost per delivered resource than unconsolidated workloads
  - For infrastructure built to less than peak:
    - Multiplexing demand → reduce the unserved demand
    - Lower loss of revenue or a Service-Level Agreement violation payout.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, with this we want to try to find out this valuation of this or whether there is a economic point of view. We try to look at little bit statistically, very little portion of they need to understand this issue. One is economies of scale, reduced overhead cost buyer power through volume purchase right. One is that, I want to one is the economy scale means I want to reduce this overhead cost right. So, what are the different overhead costs like if I buy a system, it has overhead cost of AC, it has overhead cost of power maintenance like ups or it has overhead cost of AMC of the system right? And there are several others of like human resources to maintain the thing etcetera. So, I want to reduce that overhead cost, right.

So, buying the thing is still, but maintaining the thing at times becomes over the years much costlier than the equipment itself right. And there is a major problem of especially for in the computing world that after typically 2 to 3 years. Not more than definitely within 5 years, that whole thing becomes obsolete. The technology becomes obsolete the whole system power etc, is no longer valid becomes viable for installing new set of tools and software etcetera. So, that is a big problem. There is a statistics of scale on the other end. For infrastructure build on peak requirement like infrastructure, build up peak requirement.

I think that, I build an infrastructure that always all my students will be there in the class or all will register for the course and etcetera. So, multiplexing demand may help me higher utilization right. So, when I use things on multiplexing different demand may help me higher iteration. Lower cost for deliver resources then unconsolidated work. So, it is a lower cost per delivered resources then if there is a unconsolidated workload. So, if it is a consolidated and lower cost will be there.

So, for infrastructure build to less than peak. So, it is not peak less than peak, here also multiplexing demand reduce the unnerved demands right. So, multiplexing there may be if it is a blocking architecture, there can be some of the things which are not served. So, multiplexing may reduce this unnerved, this it is not like that it is good that one it is always based to the one. I go on some sort of a scheduling algorithm and go on serving. Lower loss of revenue or a service level agreement violation. Because, SLA violation means you need to pay out something SLA violation payout. So, it may reduce lower loss of revenue and etcetera. So, both for peak infrastructure build on peak and non peak we have this sort of things.

(Refer Slide Time: 10:30)

**Coefficient of Variation -  $C_v$**

- A statistical measure of the dispersion of data points in a data series around the mean.
- The **coefficient of variation** represents the ratio of the standard deviation to the mean, and it is a useful statistic for comparing the degree of variation from one data series to another, even if the means are drastically different from each other
- In the *investing world*, the *coefficient of variation* allows you to determine *how much volatility (risk) you are assuming in comparison to the amount of return you can expect from your investment*.
- In simple language, the *lower the ratio of standard deviation to mean return, the better your risk-return tradeoff*.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Another term which comes not only for cloud, for any type of these things where this key is involved and you have to give services what you say coefficient of variance or commonly CV.

So, it is not exactly covariance we are talking about, it is coefficient of variance. So, a statistical measure of the dispersion of data in a data series around the mean, right; like a coefficient of variance is represented by the ratio, of the standard deviation of the mean to the mean. So, it is standard deviation or those who understand sigma to mu sigma by mu right. Here standard deviation by mean and it is useful statistics for comparing the degree of variation from one data series to another right. So, I can say that, this whether the coefficient of variation of this data series, is more than less than equal to other data series is this is a good measure to look at it, right, rather even if the means are drastically different then also we can compare to data series.

So, how these two data series behavioral things are there the CV gives me idea to do that right. So, it is widely used or widely looked into in the investing world. So, in investing world coefficient of variation allows, you to determine how much volatile or risk, you are assuming in comparison to the amount of return you can expect from your investment.

So, this CV gives you idea that, how much risk you are assuming in comparison to the amount of return you can expect from the investment. So, how much risk you are

involving. So, this is important if you see, it is also important in our sort of scenarios also like we are basically involving some risk by leverage, by putting my organizational from means infrastructure or on from the on premise infrastructure. So, cloud infrastructure. So, there is a risk of that if the infrastructure is not available so and so far. So, it is not only infrastructure I want to mean to say that, all type of services on the cloud thing. So, if the service is not available at it. So, how much risk I need to take on those things. So, in simple language lower the ratio of standard deviation to mean the better is your risk return trade off right.

So, lower the ratio to standard deviation to mean return. So, the better is the risk. So, I can have more smoothness into the curve that is important.

(Refer Slide Time: 13:05)

**Measure of "Smoothness"**

- Coefficient of variation  $C_V$ 
  - ≠ the variance  $\sigma^2$  nor the correlation coefficient
- Ratio of the standard deviation  $\sigma$  to the absolute value of the mean  $|\mu|$
- "Smoothen" curves:
  - large mean for a given standard deviation
  - or smaller standard deviation for a given mean
- Importance of *smoothness*:
  - a facility with fixed assets servicing highly variable demand will achieve lower utilization than a similar one servicing relatively smooth demand.
- Multiplexing demand from multiple sources may reduce the coefficient of variation  $C_V$

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, it is some sort of a measure of smoothness. So, if it is a very variable load or lot of peak and non peak things, then I am in much bigger trouble in measuring that how things will be there. If it is a smoothing out load then, I am much in a better position to do that. So, coefficient of variation CV as you says that, it is not similar as variance, nor correlation coefficient as we are mentioning ratio of standard deviation to sigma, to absolute value of the mean mu or mu mod.

So, smoother curves large mean for a given standard deviation. So, as we are telling the sigma by mu, if it is a large mean then given standard deviation things will be not varying that much or a smaller standard deviation or of a given mean that also things will

be much under control. So, importance of smoothness, a facility which fixed asset servicing highly variable demand will achieve lower utilization than a similar one servicing relatively smooth demand right. I have I am a service provider, I have some 100 systems are my backbone and I know that the demand will be something smooth right.

Then I can basically have a better management of the things or I have  $n$  number of systems. So, what will be the value of  $n$  is much easier, but if it is a very much varying suddenly demand goes up 100 then 10, etcetera, then I have a problem, right. Similarly for any, if you look if we look at our day to day life any shop shopkeeper, the amount of provisioning you will do that you will keep this in it is store, this depends on the demand thing, if the demand is something smooth, it may vary that Monday demand is different from Tuesday demand, different from week weekdays to weekends demand etcetera. Fine, but he has a idea, but it is totally random or you done cannot do anything, then it is a very difficult to provisional things. So, that is the same thing holds there.

So, multiplexing demand from multiple sources may reduce the coefficient of variation right that is the thing. So, now, I cannot predict that who what are the different sources how things will be there, but if I could have multiplexed the different demand from multiple sources. So, it may happen that this multiplex thing may have a give me a better CV right. Which may have a little smooth CV, where the overall demand things are there it is not like that all are going peak at the time, all are coming down at the lower things, but I have a multiplex type of things. Like if we look at say  $X_1, X_2, \dots, X_n$  are independent random variables of demand identical standard, say for our argument sake they are having though they are independent and then, but they have something identical sigma and mu.

So, aggregate demand in case of mean, some of the means  $n.\mu$  aggregate variance is  $n.\sigma^2$ . So, if we calculate that coefficient of variance, it is  $(1/\sqrt{n})C_v$  right. So, if  $n$  increases I multiplex more than the  $1/\sqrt{n}$  decreases. So, I have more smoothing out, this coefficient of variation or moves smooth type of curve. So, adding  $n$  independent demand reduces  $C_v$  by  $1/\sqrt{n}$ . So, it is it becomes more smoothing out. So, penalty of insufficient excess resources go smaller, right.

So, what is happening if it is smoothing out then, I do not have to plenty I have to keep more resources at my back end right, it reduces right. Otherwise I do not know whether it is a 10 demand of 10 units or 100 units and etcetera and then I have to keep a track of 100 units at the back end. Like aggregating 100 workload bring the penalty down by 10 percent, 1 by root over root over 100 is 10. So, it may bring down the whole thing by 10 percent. So, aggregating multiple resources may allow me to have a reduced loading.

(Refer Slide Time: 17:23)

**But What about Workloads?**

- Negative correlation demands
  - X and 1-X Sum is random variable 1
  - Appropriate selection of customer segments
- Perfectly correlated demands
  - Aggregated demand :  $n.X$ , varianceofsum: $n^2\sigma^2(X)$
  - Mean:  $n.\mu$ , standard deviation:  $n.\sigma(X)$
  - Coefficient of Variance remains constant
- Simultaneous peaks

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

But, what about the different workloads; non negative correlation demands like if  $x$  and  $1-x$  sum of a random variable is 1 appropriate selection of the customer segments right is another important thing right, I say that I have a computing infrastructure. I select those customers, some of them are active on day time, some of them processing are active on the night time, right.

So, it is compensating, right, I if the both are working at the same time then the peak will go much higher, but selecting that negative demands or the time in the timescale. I could have managed those thing with the same infrastructure. So, that is selection of the customer base is the important things, what sort of things will be there when I go on selecting as customers. Perfectly correlated demand, if some of the things are perfectly coordinated the aggregate demand will be  $n.x$ . Variations will be  $n^2\sigma^2(x)$  and mean will be  $n.\sigma$  and standard deviation  $n.\sigma(x)$  and so far, coefficient of variation remains constant. So, it is perfectly correlated demand that this thing will be as a constant.

There are third issue if all demands are coming at the peak at the same time all at the peak at the same time then, I have a serious problem right. Then I have congestion at the things all are demanding at the same time. Like if I say all classes are breaking at the hourly basis. So, at the hourly basis, we have a huge demand for this root network, right. Lot of vehicles like said studying for cycles to it said they are on the demand because all classes are things. So, it is a peak coming to the thing at the same time like this then we have a problem.

(Refer Slide Time: 19:13)

**Common Infrastructure in Real World**

- Correlated demands:
  - Private, mid-size and large-size providers can experience similar statistics of scale
- Independent demands:
  - Midsize providers can achieve similar statistical economies to an infinitely large provider
- Available data on economy of scale for large providers is mixed
  - use the same COTS computers and components
  - Locating near cheap power supplies
  - Early entrant automation tools → 3<sup>rd</sup> parties take care of it

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES



So, common infrastructure in our real world, correlated demands, private, mid sides, large sized providers can experience similar statistic scales. So, it is a more or less correlated demand. Independent demand, midsize provider can achieve similar statistical economy to an infinitely large provider right. Available data on economy of scale for large provider is mixed right because use of same COTS type of things that is commercial of the self systems and components locating near cheap power supplies that is one thing like if as the power supply is a major thing for the data centers, they want to build a data center where the power supply will be much cheaper and nearby. Early entrain automation tool third party parties takes care of it. So, there can be early entrant automation tools which the third party can be taking.

(Refer Slide Time: 20:12)

**Value of Location Independence**

- We used to go to the computers, but applications, services and contents now come to us!
  - Through networks: Wired, wireless, satellite, etc.
- But what about latency?
  - Human response latency: 10s to 100s milliseconds
  - Latency is correlated with:
    - Distance (Strongly)
    - Routing algorithms of routers and switches (second order effects)
  - Speed of light in fiber: only 124 miles per millisecond
  - If the Google word suggestion took 2 seconds ☺
  - VOIP with latency of 200ms or more ☺

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Say value of there is a value of corresponding to the location or value of location independence. We used to go to the computers, but the application services now come to us right. So, there is a paradigm shift like say yesterdays or we used to go to the system to work on it. Now the system power etcetera coming to my own desktop like; so, I have a large pool of huge resources on my very thin system, it can be a simple desktop, laptop and type of things all are provisions here, through networks wired wireless satellite etc, but what about latency?

So, latency also a big thing, right. So, human response is 10 to milliseconds, 10 to 100 milliseconds. So, latency is correlated strongly with the distance right more that distance, more the network latency another type of latency more on the hopes, more on the failure rates and etcetera.

So, though it also depends on what sort of routing algorithms etcetera also coming into play. So, speed of anyway we know that speed of light in fiber. So, some particular 124 miles per milliseconds. So, if suppose I am searching something and it takes more than couple of seconds, then we are not happy with that even a VOIP thing, if it is something delayed more than something, 200 nano second; a 200 millisecond second, then it is very difficult to communicate over this voice over IP.

(Refer Slide Time: 21:49)

**Value of Location Independence (contd...)**

- Supporting a global user base requires a dispersed service architecture
  - Coordination, consistency, availability, partition-tolerance
  - Investment implications

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, there is a supporting of a global user base requires a dispersed service architecture. So, the architecture if I want to support, if my provider is to support all the global user base then, I have to have an appropriate distributed and dispersed architecture to do that and similarly the protocol. So, coordinate and coordination consistency availability partition tolerance these are issues. So, and it has a direct implication on investment that, what sort of investment we want to do. We need to we like to look at another quickly another aspects of the thing.

(Refer Slide Time: 22:20)

**Value of Utility Pricing**

- As mentioned before, economy of scale might not be very effective
- But cloud services don't need to be cheaper to be economical!
- Consider a car
  - Buy or lease for INR 500/- per day
  - Rent a car for INR 5,000/- per day
  - If you need a car for couple of days in a trip, buying would be much more costly than renting
    - It depends on the demand

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

What we say value of utility pricing right. That how things will be priced, how things will be there, how I can provision system and it may also help me to look at that, whether it is useful to be go to cloud or having something at your own premises.

So, economy of scale might not be very effective always means all taking consideration right, but cloud service do not need to be cheaper to be economical right. So, it is economic cop thinks, may be based on that my requirement and etcetera what sort of demand I am going. Like, if we the popular example buildings to give is that a consider a car. So, buy or leasing a car may cost me something 500 per day right. Whereas, renting a car may be say for example, 500 per day. So, when it is economical? Suppose, I by looking at it is always the buying the car may be economical.

But, I if I am commuting say large distance or say one scene; that means, one once in a month or couple of delayed days in a month. Then buying of car may be more costly than renting a car, but if I require that car on a daily basis, that going to my workplace traveling a large distance etcetera then the buying a car, will maybe economical than renting a car; all right.

So, it all depends that, what sort of demand demands you are having.

(Refer Slide Time: 24:11)

### Utility Pricing in Detail

$D(t)$	demand for resources $0 < t < T$
$P$	$\max(D(t))$ : Peak Demand
$A$	Avg $(D(t))$ : Average Demand
$B$	Baseline (owned) unit cost $[B_T : Total Baseline Cost]$
$C$	Cloud unit cost $[C_T : Total Cloud Cost]$
$U (=C/B)$	Utility Premium $[For rental car example, U=4.5]$

$$C_T = \int_0^T U \times B \times D(t) dt = A \times U \times B \times T$$

$$B_T = P \times B \times T$$

- Because the baseline should handle peak demand

When is cloud cheaper than owning?

$$C_T < B_T \Rightarrow A \times U \times B \times T < P \times B \times T$$

$$\Rightarrow U < \frac{P}{A}$$

- When utility premium is less than ratio of peak demand to Average demand


IIT KHARAGPUR

NPTEL ONLINE  
CERTIFICATION COURSES

So, you just to do some simple some mathematic expression little bit simplified so that try to have. Like suppose, I have a demand  $D(t)$ . So, demand varies over time 0 to capital

$T$  0 to time. So, demand for resources  $D(t)$ ,  $P$  is the max demand or peak demand,  $A$  is the average demand. So, average over the time scale,  $B$  is the baseline or own unit cost. So, if I only unit cost, what is the baseline cost  $C$  is the cloud unit cost. So, what is the cloud unit cost if I purchase the thing and  $U$  utility premium rent, a car for that there is a some type. So, there is a rent that car maybe something, it will come whatever in our case is 5000 by 500. So, something 10 is the utility.

So, utility of the premium is that taking a cloud service divided by the baseline service right. That is the utility premium, we are having the things right. That is the utility premium now so, I have a variable demand  $D(t)$ , I have a peak demand  $P$ , we have a average demand, there is a baseline owned unit cost  $dc$ , cloud unit cost  $C$ . And I want to find out the utility premium that is  $C/B$ . Now, if I want to see the overall cloud cost. So,

$\int_0^T U \times B \times D(t) dt$  is the overall cloud cost which is somewhat  $u$  about is the overall costing

of my cloud.

If I want to calculate the overall baseline over time  $T$ ; so,  $P$  that is the peak demand into because whenever I am going for based my baseline. I have to go for the peak demand thing to calculate the thing  $B$  is the baseline own unit cost and  $T$  is the overall time scale.

Now, when cloud is cheaper, when the  $C_T < B_T$ , if the cost of cloud is less than the cost of baseline then it is cheaper. So, or in other sense if you look at it very simplified form. So, when  $P/A$  is greater than the utility premium right. Peak cost by average, cost a peak demand by average demand is greater than the utility premium. So, when the utility premium is less than the ratio of peak demand, to average demand then your cloud may be cheaper then owning the infrastructure or owning the services.

(Refer Slide Time: 27:05)

**Utility Pricing in Real World**

- In practice demands are often highly spiky
  - News stories, marketing promotions, product launches, Internet flash floods, Tax season, Christmas shopping, etc.
- Often a hybrid model is the best
  - You own a car for daily commute, and rent a car when traveling or when you need a van to move
  - Key factor is again the ratio of peak to average demand
  - But we should also consider other costs
    - Network cost (both fixed costs and usage costs)
    - Interoperability overhead
    - Consider Reliability, accessibility

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, utility pricing in real world, in praxis demands are often offense pipe spiky like new stories suddenly, it came into things. Market promotions, product launches, internet flash flood something goes to the internet etcetera. Some seasonal things like Christmas, Tax. So, those time things goes up. Often hybrid model is based. Like you own a car for daily commute, but rent a car when traveling or when you need a larger move to a larger distance to move. Key factor is again the ratio of peak to average demand. So, key factor is there again that, what is the ratio between the peak to average demand, but we should also consider other cost.

Like which had not considered in the previous calculation, then our network cost both fixed and usage costs interoperability overhead Like different one information, a one data is talking to other there in overhead. Consider reliability accessibility so and so far right. So, these are the different factors.

(Refer Slide Time: 28:17)

**Value of on-Demand Services**

- Simple Problem: When owning your resources, you will pay a penalty whenever your resources do not match the instantaneous demand
  - Either pay for unused resources, or suffer the penalty of missing service delivery

D(t) – Instantaneous Demand at time t  
R(t) – Resources at time t

Penalty Cost  $\alpha \int |D(t) - R(t)| dt$

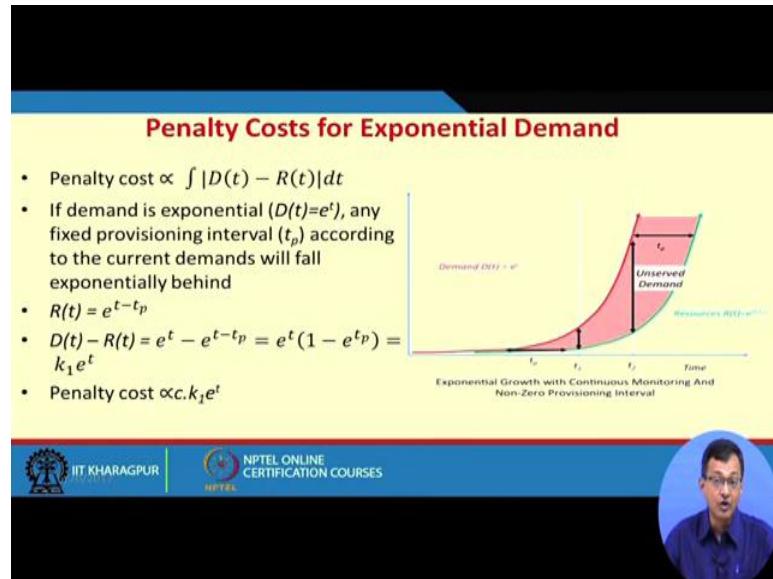
- If demand is flat, penalty = 0
- If demand is linear periodic provisioning is acceptable

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Another aspect, we will just quickly see the value of on demand services. So, simple problem when owning your own resources, you pay the penalty whenever the resources do not match with the instantaneous demand. Suppose, I have 100 resources and I pay the penalty. When it is underutilized, suppose it is utilized with 80 percent, 70 percent then I pay the penalty for the rest of the 10 to a 20, 30 things, right. Either pay for unused resources or suffer penalty of missing service delivery things are there right. So, if it is higher things then I do not able to service.

So, penalty is how do I calculate? Penalty is proportional to  $\int |D(t) - R(t)| dt$ , all right. If demand is flat penalty is 0 rights. If demand is linear periodic provisioning is acceptable, right if it is a linear than periodic provisioning the acceptable.

(Refer Slide Time: 29:17)



If the demand is non-linear, then periodic provisioning in cloud is a big question right. Suppose if the demand is exponential like in this case  $D(t)=e^t$  right, any fixed provisioning time interval  $t_p$  according to the current demand we will feel you will fall exponentially behind. Suppose, I require time  $t_p$  to provision the things right by the time I provision it has gone up. So, it goes on, this at  $t$  equal to  $e$  of  $t$  minus  $t_p$ . This time  $D_t$  that time to provisioning  $t_p$ , will create a havoc right. So,  $D(t) - R(t) = k_1 e^t$  if you say.

So, penalty of cost is  $c.k_1 e^t$ . In other sense this penalty grows exponentially. It is extremely difficult to match this, unless you over provision and try to look at that is also at times difficult because it grows in exponential things. So, if you need to be careful that, when what type of demands we are expecting need to study and based on that the provisioning should be there.

(Refer Slide Time: 30:27)

### Assignment 1

Consider the peak computing demand for an organization is 120 units. The demand as a function of time can be expressed as:

$$D(t) = \begin{cases} 50 \sin(t), & 0 \leq t < \pi/2 \\ 20 \sin(t), & \pi/2 \leq t < \pi \end{cases}$$

The resource provisioned by the cloud to satisfy current demand at time  $t$  is given as:

$$R(t) = D(t) + \delta \cdot \left( \frac{dD(t)}{dt} \right)$$

Where,  $\delta$  is the delay in provisioning the extra computing recourse on demand

  IIT KHARAGPUR |  NPTEL ONLINE CERTIFICATION COURSES

So, based on these I have kept a small assignment, for you to work at your own time and we will discuss this assignment in one of these classes during some free time. So, it says that consider a peak computing demand of an organization 120 units and demand of the functions time express. At this like these, are the demand of with respect  $D(t)$  is that demand over  $t$ , the resource provisioning of the cloud to satisfied the current demand  $t$  is at equal to so and so far where  $\delta t$  is the delay in provisioning extra computing resource on demand. So,  $\delta$  this tilde is the delay, right.

(Refer Slide Time: 31:11)

### Assignment 1 (contd...)

The cost to provision unit cloud resource for unit time is 0.9 units.  
Calculate the penalty and draw inference.

*[Assume the delay in provisioning is  $\pi/12$  time units and minimum demand is 0]*

(Penalty: Either pay for unused resource or missing service delivery)

  IIT KHARAGPUR |  NPTEL ONLINE CERTIFICATION COURSES

So, what we say that, the cost to provision unit cloud resource for unit time is 0.9 units. So, we want to calculate the penalty, if any and draw the inference like why this type of penalty how things are there it is so and so far. So, we are assuming some of the factors like, what sort of the; what should be the delay in provisioning, penalty it may be either pay for on his users or missing service delivery and type of things. So, this I encourage you to look at this problem, with this thing that while things will be there will in subsequent class one of this class, we will discuss this problem. So, that is all for today. So, we looked at out the economy of cloud and where it is beneficial and what are the different aspect to which drives the economy of cloud and type of things.

Thank you.

**Cloud Computing**  
**Prof. Soumya Kanti Ghosh**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**

**Lecture - 13**  
**Managing Data**

Hello. So, we will continue our discussion on cloud computing. Today, we will discuss about some aspects of managing data in cloud, right. So, as we understand that in cloud; as we have discussed in our earlier lectures that in cloud, one of the major aspect is the data because at the end of the day, your data and even processing applications are in somebody else's domain, right. So, they are being executed at somewhere else which is beyond your direct control. So, it is virtually host in some virtual data; a virtual machine somewhere in the cloud. So, it becomes tricky to on the security point of view that we have discussed; not only that if you look at from the other point of view; so, from the clouds provider point of view, managing huge volume of data keeping their replicas and making them queriable and these becomes a again a major issue.

So, all our conventional relational or object oriented model may not directly fit into the thing, right. So, long you are doing on a small instances experimental some database application or some small experimentations, then it is fine, but when you have a large scale thing where huge amount of read write going on or the volume of data is much much higher than the normal operations, then it is; we need to look in a different way. These are the things which come into not only for the cloud, it was there a little earlier also; like how this parallel database accesses; parallel database execution; read-write execute operations can be done. So, those things become more prominent or a de facto mechanisms; when we talk about in context of cloud. So, what we will try to do is more of a overview of how data can be managed in cloud or what are the difference strategies or schemes people or this ISPs follows and it is not exactly the security point of view; it is more of a management data management point of view, right.

(Refer Slide Time: 03:02)

The slide has a dark blue header and footer. The main content area is yellow. The title 'Introduction' is in red at the top of the yellow section. Below it is a bulleted list:

- Relational database
  - Default data storage and retrieval mechanism since 80s
  - Efficient in: transaction processing
  - Example: System R, Ingres, etc.
  - Replaced hierarchical and network databases
- For scalable web search service:
  - Google File System (GFS)
    - Massively parallel and fault tolerant distributed file system
  - BigTable
    - Organizes data
    - Similar to column-oriented databases (e.g. Vertica)
  - MapReduce
    - Parallel programming paradigm

At the bottom, there are two logos: IIT Kharagpur on the left and NPTEL ONLINE CERTIFICATION COURSES on the right.

So, we will talk about a little bit of relational database already known to you then what you know to do that scalable data bases or data services like one of the couple of things are important one is Google file system big table and there is a Mapreduce parallel programming paradigm; those are the things which comes in back to back, when we are doing to the things. So, what we want to do when we were we are managing anything on a cloud platform; whether it is application or data we want to make it scalable in the sense the it suites scale as the requirement goes up. So, scale-up scale-down in a ubiquitous way or minimum interference from the; or minimum human or management interference. So, that type of infrastructure; we want to come up with, right, it is true for data also.

(Refer Slide Time: 04:09)

**Introduction** Contd...

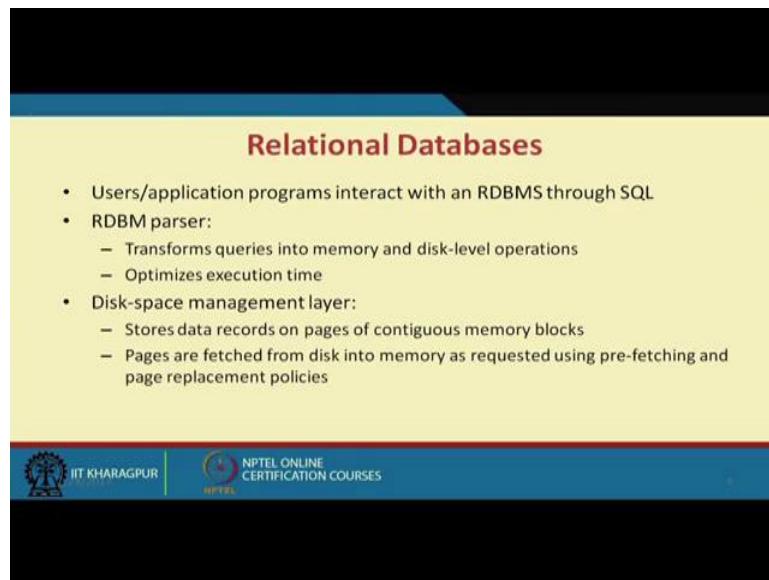
- Suitable for:
  - Large volume massively parallel text processing
  - Enterprise analytics
- Similar to BigTable data model are:
  - Google App Engine's **Datastore**
  - Amazon's **SimpleDB**

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, these are primarily suitable for large volume of massively parallel text processing, right that is one of the major thing or it is suitable for environment say enterprise analytics, right, I want to have a; if we want to do analytics on a distributed data stores, right, it may be a chain of a shopping or commercial staff or it may be a banking organization or financial any financial organization, even it is something to do with large volume of other type of data like it metrological data, it maybe climatological data something which need to be chant or has a distributed things, I need to do some parallel processing down the line where the actual effect comes into play. If you have a simple database with a simple instant, then you may not have gone to cloud for that; right. So, it may be a simple system or you buy a very a VM and work on it then the actual effect of cloud things are actual advantages of cloud you are not taking out.

So, we will see that similar to big table models there are Google app engines datastore, Amazon simple DB which are which different provides provide in different flavor, but the basic philosophy are same.

(Refer Slide Time: 05:46)



The slide has a dark blue header and footer. The main content area is light yellow. The title 'Relational Databases' is in red at the top of the yellow section. Below it is a bulleted list of four items. At the bottom of the slide, there are two logos: IIT Kharagpur on the left and NPTEL ONLINE CERTIFICATION COURSES on the right.

- Users/application programs interact with an RDBMS through SQL
- RDBM parser:
  - Transforms queries into memory and disk-level operations
  - Optimizes execution time
- Disk-space management layer:
  - Stores data records on pages of contiguous memory blocks
  - Pages are fetched from disk into memory as requested using pre-fetching and page replacement policies

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

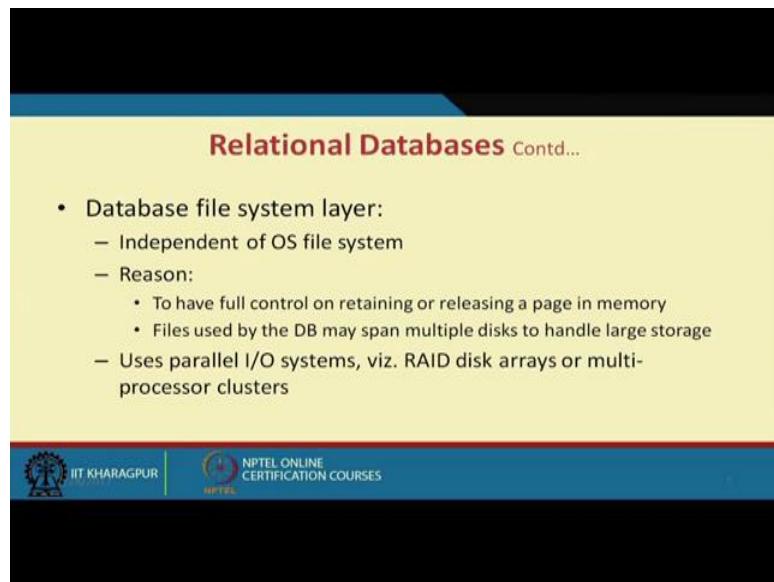
So, if we look quickly look at the relational data base which is known to all of you or most of you users application programs interact with RDBMs through SQL, right. So, it is the structured query language or SQL by which we interact with the user programs, etcetera.

So, there is a relational database management parser which transforms queries into memory and disk label operations and optimize the execution time. So, in any query, we need to optimize the execution time of the query, right. So, if it is a large data base like you, whether you do project before select join before or after select that makes a lot of difference; though the query may be same the query output will be same, but the execution time may vary to a great extent, right, like I have a huge 2 data bases like R1 say relational databases R1, R2 and I do some projection or selection of some of the things, right I select A1, A2 and then do a; then do the; join whether I do the join before or after makes the things like suppose; if I do the select on R1; the number of tuples come down from 1 million to say few 1000s. Similarly for R2, if I do a select on that; right. So, then joining is much less costlier. So, whether you do the join first or it said that becomes a thing that is a database optimization problem nothing to do specifically for cloud, but relational database allows you to optimize those things.

Disk space management layer, this another property that stores data records on pages of contiguous memory block. So, that the disk movement is minimized pages are fetched

from the disk into memory as requested state using pre fetching and page replacement policies. So, this is another aspects of the things like one is looking at that property making it more efficient in the query processing, other aspect it make it more efficient in storage terms of things like nearby things if the query requires the some 5 tables if they are nearby store then the access rate is high. So, database file system layer.

(Refer Slide Time: 08:15)



The slide has a yellow header bar with the text "Relational Databases Contd..." in red. Below this is a black content area containing a bulleted list about the Database file system layer. At the bottom, there is a blue footer bar with the IIT Kharagpur logo and the text "NPTEL ONLINE CERTIFICATION COURSES".

- Database file system layer:
  - Independent of OS file system
  - Reason:
    - To have full control on retaining or releasing a page in memory
    - Files used by the DB may span multiple disks to handle large storage
  - Uses parallel I/O systems, viz. RAID disk arrays or multi-processor clusters

So, previously we have seen that RDBM parser then disk space management layer then database file system layer. So, it is independent of OS file system, it is a separate file system. So, it is in order to have full control on retaining or realizing the page in the memory, files used by the DB or database may span multiple disk to handle large storages, right.

So, in other sense like if I dependent on the operating system for phase all those things then it is fine when your again database load is less if it is pretty large then the number of hope you take it text it becomes costly. So, what you need to do we need to do directly interact at the at the much lower level with the with the hardware or the available resources and that exactly this database file system layer tries to immolate uses parallel IO like we have heard about Raid disk Raid1, Raid2, Raid5, Raid 6eN type of things arrays or multiple clusters. So, which keeps a redundant redundancy into the thing. So, the your this failure down time is much less so; that means, is it is basically full failure proof implementation of the database.

(Refer Slide Time: 09:42)

The slide has a yellow header bar with the title 'Data Storage Techniques' in red. Below it is a white content area with two main bullet points: 'Row-oriented storage' and 'Column-oriented storage'. Each point has several sub-points describing its characteristics.

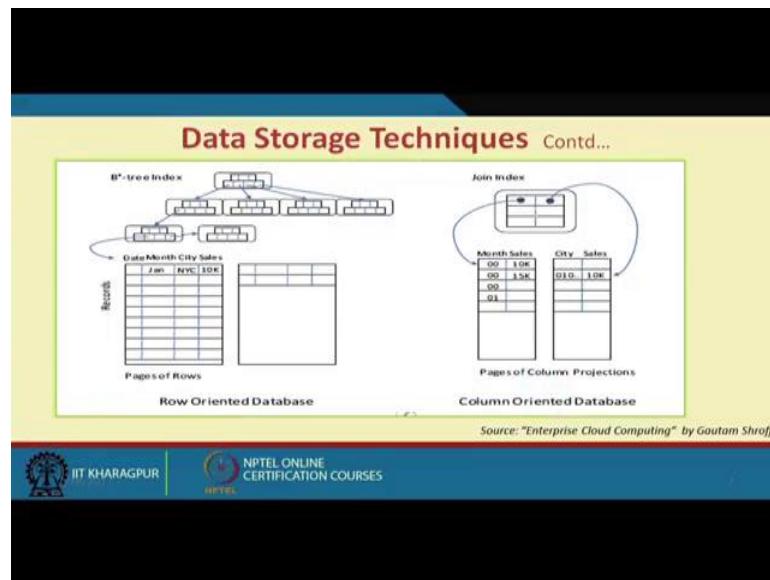
- Row-oriented storage
  - Optimal for write-oriented operations viz. transaction processing applications
  - Relational records: stored on contiguous disk pages
  - Accessed through indexes (primary index) on specified columns
  - Example: B+ tree like storage
- Column-oriented storage
  - Efficient for data-warehouse workloads
    - Aggregation of **measure** columns need to be performed based on values from **dimension** columns
    - Projection of a table is stored as sorted by dimension values
    - Require multiple "join indexes"
      - If different projections are to be indexed in sorted order

At the bottom, there are logos for IIT Kharagpur and NPTEL, along with the text 'NPTEL ONLINE CERTIFICATION COURSES'.

So, usually the databases storage as row oriented that is we had tuples and its a set of row of the same schema optimal for write oriented operation the transaction processing applications, relational records stored in contiguous disk pages access through indexes primary key on specific columns, B plus tree is one of the favorite storage mechanisms for this sort of thing. Column oriented efficient for data warehouse workloads right. So, those who have gone through data warehouses. So, it is a high dimensional data huge volume of data and being collected and populated by different things. So, it is more of a warehouse, rather than a simple database. So, this is this column oriented storage are more suitable for data warehouse type of loads aggregate of measures where rather than individual data it is more of the analysis on analytics come into play. So, it is aggregation of measure columns need to be performed based on the values of the dimension columns. So, we are not going to the data warehouse. So, it has a different dimension tables and type of things and we need to the operations are more aggregate operations, right, we want to do some sort of analysis and type of things.

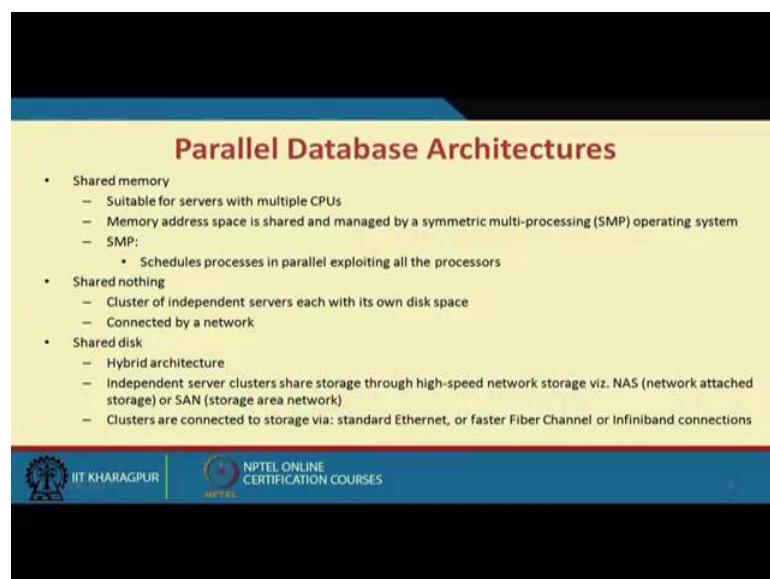
So, projection of a table is stored on as a stored on a dimension table dimension values in case of a column oriented require multiple join indexes if different projection are to be indexed in a sorted order right. So, it is; if it is a different-different thing because the organization may have different views for different type of data and need to be stored in that fashion.

(Refer Slide Time: 11:31)



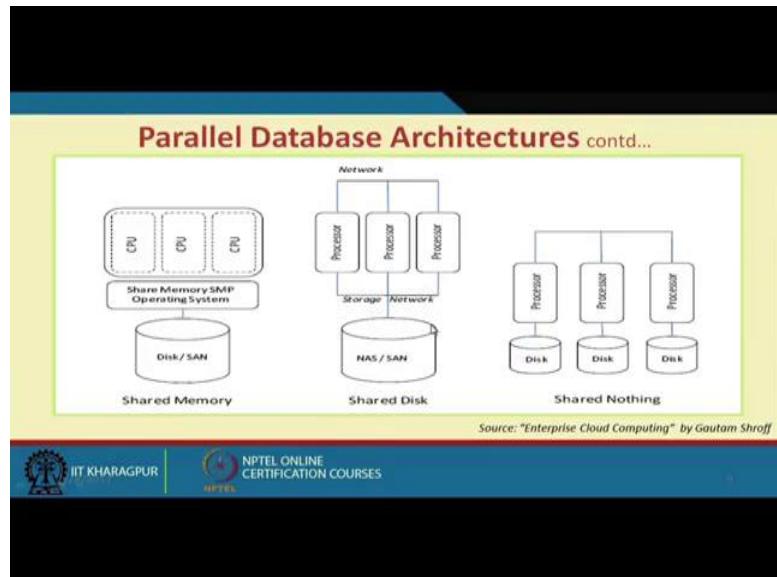
So, data storage techniques as we have seen; it is B plus tree or join indexes. So, one is row oriented, other one is column oriented. So, this is row oriented data and this is column oriented data and we need to have a join index which allows this data to be linked to one another. So, these all these we will get in any standard database book or in standard literature; primarily as we are following that Gautam Shroff's Enterprise cloud computing book for this particular thing. So, that is why we have mentioned, but this is a very standard operation and you can get in any standard books.

(Refer Slide Time: 12:16)



So, if we look at the parallel database architectures. So, it is broadly divided into 3 aspects one is shared memory one is shared nothing another is shared disk, right.

(Refer Slide Time: 12:30)



So, I just see the picture fast then come back. So, this is a typical structure of the shared memory, right. So, these processors different processors shared the memory, here it is a shared disk. So, different processors shared the disk, here we have shared nothing. So, individual processor has individual disk; so, in case of a shared memory suitable for servers with multiple CPUs. So, if there are multiple CPUs. So, if there are multiple CPUs memory address space is shared and managed by SMP operating systems like the memory address. This is shared among these SMPs and schedule processors in parallel exploiting the processors. So, it schedules small things so; that means, I have a shared memory space and I basically do a execution in a parallel mode.

So on the extreme other end is shared nothing. So, cluster independent servers with each of its having own disk space and connected by a network. So, at the with a back bone high speed network if any server shared its own disk space and then do the rest of the execution and if we look at that in between the thing is the shared disk like it is a hybrid architecture. So, to say independent server cluster storage through high speed network that can be NAS or SAN and clusters are connected to storage data via standard Ethernet fiber, etcetera what we have shown here. So, it is a shared storage and these different

processor access this. So, based on your application type of parallelisms you need we can go for any of this structure.

So, here we see that it is more this more efficient if the memory things are more compact where in the other end we if the processors are individually working on separate data sets and there are machine to say then this could have been a advantage.

(Refer Slide Time: 14:32)

**Advantages of Parallel DB over Relational DB**

- Efficient execution of SQL queries by exploiting multiple processors
- For **shared nothing** architecture:
  - Tables are partitioned and distributed across multiple processing nodes
  - SQL optimizer handles distributed joins
- Distributed **two-phase commit** locking for transaction isolation between processors
- Fault tolerant
  - System failures handled by transferring control to "stand-by" system [for transaction processing]
  - Restoring computations [for data warehousing applications]

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, if we look at the advantages of parallel DB of relational database, if you do not want to put that; what are the features of relational parallel database structures which is more advantages for parallel this sort of operations, then the relational database efficient execution of SQL query by exploiting multiple processors, for shared nothing architecture tables partition and distributed across possessing table, right. So, happened that I can partition the table and every the data accountant in the table can be executed parallelly they can be distributed in the different days and the processor can work that totally depends on your; what is your working mechanisms out there.

So, SQL optimizer handles this distributed joint. So, whenever we need to do some join then we need to fall on the; distribute your SQL optimizer. So, distributed 2 phase commit locking for transaction isolation between the processors. So, these are the some of the features, fault tolerant like system failures handled by transferring control to standby system. So, I can have different standby system or some with some protocol or some policy and then if there is a failure, then I can shift that particular execution to

some of the standby system. So, that is possible in this sight of things and restoring computation for data though these are the things which are more required for data warehouse type of applications.

(Refer Slide Time: 16:15)

**Advantages of Parallel DB over Relational DB**

- Examples of databases capable of handling parallel processing:
  - Traditional transaction processing databases: **Oracle, DB2, SQL Server**
  - Data warehousing databases: **Netezza, Vertica, Teradata**

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, there are examples of databases capable of handling parallel processing traditional transaction processing things are oracle, DB2, SQL server data warehouse application are some of the Vertica, Teradata, Netezza; these are the some of the things which are more of a data warehouse type of database. Now with these background or with these things in our in our store what we say we look at that cloud file system.

(Refer Slide Time: 16:50)

The slide has a yellow header bar with the title "Cloud File Systems". Below it is a white content area containing two bulleted lists. The first list is for "Google File System (GFS)" and the second is for "Hadoop Distributed File System (HDFS)". Both lists include sub-points. At the bottom of the slide, there is a blue footer bar with the IIT Kharagpur logo on the left and "NPTEL ONLINE CERTIFICATION COURSES" on the right.

- Google File System (GFS)
  - Designed to manage relatively large files using a very large distributed cluster of commodity servers connected by a high-speed network
  - Handles:
    - Failures even during reading or writing of individual files
    - Fault tolerant: a necessity
      - $p(\text{system failure}) = 1 - (1 - p(\text{component failure}))^N \rightarrow 1$  (for large  $N$ )
    - Support parallel reads, writes and appends by multiple simultaneous client programs
  - Hadoop Distributed File System (HDFS)
    - Open source implementation of GFS architecture
    - Available on Amazon EC2 cloud platform

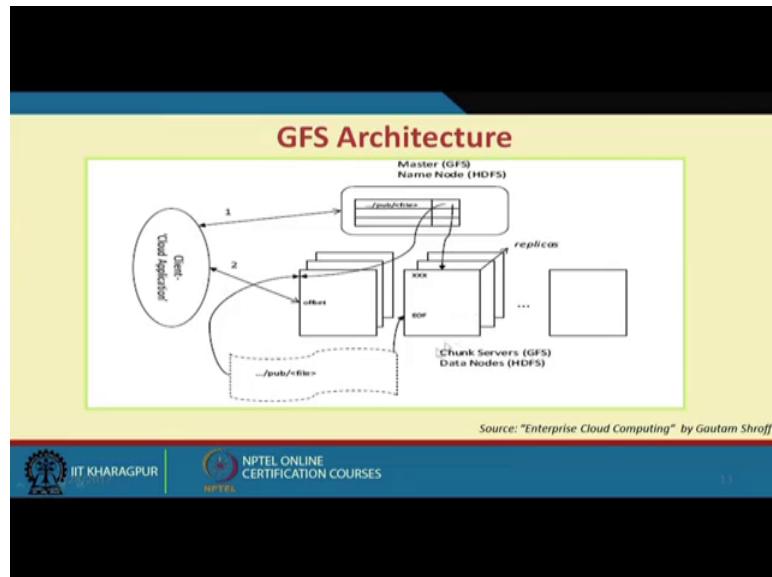
Now, as we understand it will not go something become totally we cannot through the whole thing out of the thing and start doing something new because this database has grown; they are fault tolerant, they are efficient we have raids and type of things we need to exploit some of the things and put some more philosophy of which behind the cloud.

So, one of the predominant thing is cloud file Google file system was GFS and back to back; we have a open source stuff called HDFS; Hadoop distributed file system. So, which is what we say someone to one mechanism set Google file system. So, Google file system, design to manage relatively large files using a very large distributed clusters of commodity servers connected by high speed things. So, it is whether GFS or HDFS, they are enable to work on very large data files which are distributed over this commodity servers; typically some of the things are Linux servers which are interconnected through a very high speed line.

So, they can handle failure even during read write of individual files, right, during the read-write operation if there is a failure it can handled. Fault tolerant it is definitely a necessity. So, if we have any that is any simple system term that  $P(\text{system failure})$  probability of system failure is  $1 - (1 - P(\text{component failure}))^N$ . So, for if the  $N$  is pretty large, then you can say that we can go for that is the risk of this failure is minimum. So, supports parallel reads writes appends multiple simultaneous client program. So, it is parallel read parallel write and update by the client program and we have HDFS that is

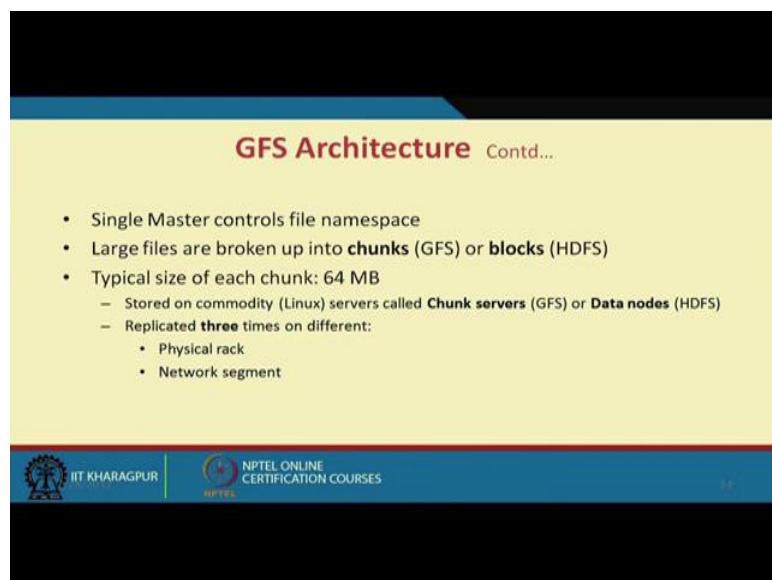
Hadoop distributed file system which is open source implementation of GFS architecture available on Amazon EC2 cloud platform from. So, we have HDFS which is there.

(Refer Slide Time: 19:18)



So, if we have a big picture. So, that how a typical GFS are there. So, there are some of the components are there is master or the name nodes master node in GFS or name node is HDFS and there are client applications and we have different chunk server in case of GFS and data nodes in the case of HDFS in a typical cloud environment. So, single master controls the namespace.

(Refer Slide Time: 19:47)



So, logically a single master is there which control the namespace. So, namespace is important because it gives us that how there are stored; how data can be referred; it is more of a; it may modes of a meta-data sort of information which is controlled by the master large files are broken into chunks, in case of a GFS and block; what we called in case of a HDFS stored on commodity server, typically Linux servers called chunk servers in GFS and data nodes in HDFS, so replicated 3 times on different physical rack network segment. So, this chunk; so, what we have? We have the GFS or HDFS in the things below that we are having a chunk servers which are basically Linux servers chunk server or data nodes in the things which are the main custodian of the data and they are the every data  $D_i$  is replicated on different 3 times at least 3 time on different physical rack and network segments.

(Refer Slide Time: 21:11)

**Read Operation in GFS**

- Client program sends the full path and offset of a file to the **Master** (GFS) or **Name Node** (HDFS)
- Master replies with meta-data for **one** of replicas of the chunk where this data is found.
- Client caches the meta-data for faster access
- It reads data from the designated chunk server

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, if you look at the read operation in GFS, client program sends the full path offset of a file to the master, right where it wants to read or name node in case of HDFS. So, we will refer the GFS master node and which is back to back when we it is refer to the name node in HDFS master replies on meta-data for one of the replicas of the chunk where these data is found, right, client caches the meta-data for faster access. It reads the data from the designated chunk server. So, master from the master; it gets that and gets the mirror this meta-data and from there it basically access this chunk server.

(Refer Slide Time: 22:01)

**Write/Append Operation in GFS**

- Client program sends the full path of a file to the **Master (GFS)** or **Name Node (HDFS)**
- Master replies with meta-data for **all** of replicas of the chunk where this data is found.
- Client send data to be appended to all chunk servers
- Chunk server acknowledge the receipt of this data
- Master designates one of these chunk servers as **primary**
- Primary chunk server appends its copy of data into the chunk by choosing an offset
  - Appending can also be done beyond **EOF** to account for multiple simultaneous writers
- Sends the offset to each replica
- If all replicas do not succeed in writing at the designated offset, the primary retries

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, for read operation any of these chunk server or replicated chunk server will do where write append operation in GFS is little tricky, client program sends a full path of file to the master GFS or name node HDFS right, the master replies on the meta-data for all replicas of the chunks where the data is found. The client send data to be appended into the all chunk servers; chunk server acknowledges the receipt of the data, master designate one of the chunk server as primary, the primary chunks server appends its copy of the data into the chunk by offset choosing an offset, right. So, that it do it appending; appending can also be done beyond end of file to account for the multiple simultaneous, right.

So, this is a pretty interesting thing that even if you can have append end of EOF beyond EOF because there are simultaneous writers which are writing and it basically consolidated at later stage. Sends offset to the replica, if all replica do not success in writing in the designated offset, the client retries, right. So, the all offset; so, idea is that whenever I am looking for a data, I need to know that for all the 3 replicas, it should be at the same offset ideally. So, that I the read processed as there is no delay in that things because once its calculates it is directly access the other chunks on that offset, right.

(Refer Slide Time: 23:42)

The slide has a yellow header bar with the title "Fault Tolerance in GFS". Below the title is a bulleted list of points:

- Master maintains regular communication with chunk servers
  - Heartbeat messages
- In case of failures:
  - Chunk server's meta-data is updated to reflect failure
  - For failure of primary chunk server, the master assigns a new primary
  - Clients occasionally will try to this failed chunk server
    - Update their meta-data from master and retry

At the bottom of the slide, there are two logos: IIT Kharagpur and NPTEL Online Certification Courses.

So, fault tolerant in Google file system; the master maintains regular communication with the chunk server what we say heart beat messages sort of a are you alive type of thing and in case of a failure chunk server meta-data is updated to reflect failure for failure of primary chunk server the master assigns a new primary clients occasionally we will try to this failed we will try to this failed chunk server, update their meta-data from the master and retry. So, in case of a failure the chunk server meta-data after reflect the failure. So, the chunk server meta-data says that there is a failure. So, the next time you do not allocate or like that and for failure of the primary server itself, the master assigns a new primary. So, it assigns a new primary to work on the thing.

(Refer Slide Time: 24:47)

## BigTable

- Distributed structured storage system built on GFS
- Sparse, persistent, multi-dimensional sorted map (**key-value pairs**)
- Data is accessed by:
  - Row key
  - Column key
  - Timestamp

The diagram illustrates a row in a BigTable. The row key is "com.cnn.www". It contains three column families: "contents", "anchor|cnn.com", and "anchor|my.look.ca". The "contents" family has two versions, t<sub>1</sub> and t<sub>2</sub>, with values "HTML" and "text/html" respectively. The "anchor|cnn.com" family has one version, t<sub>3</sub>, with value "CNN". The "anchor|my.look.ca" family has one version, t<sub>4</sub>, with value "CNN.com".

Source: "Enterprise Cloud Computing" by Gautam Shroff

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

And update the clients; occasionally we will try to this failed chunk server because it will be flagged, right. Now another related stuff is big data or related concept of big data, distributed structure storage 5 system build on GFS, right. So, it is build; it is a structure distributed structure storage file system it is build on GFS, right. So, data is accessed by row key, column key, timestamp. So, if you look at. So, it is a multiple instances are stored. So, there is a time key column key and of course, say row key which says that where the data is there.

(Refer Slide Time: 25:26)

## BigTable Contd...

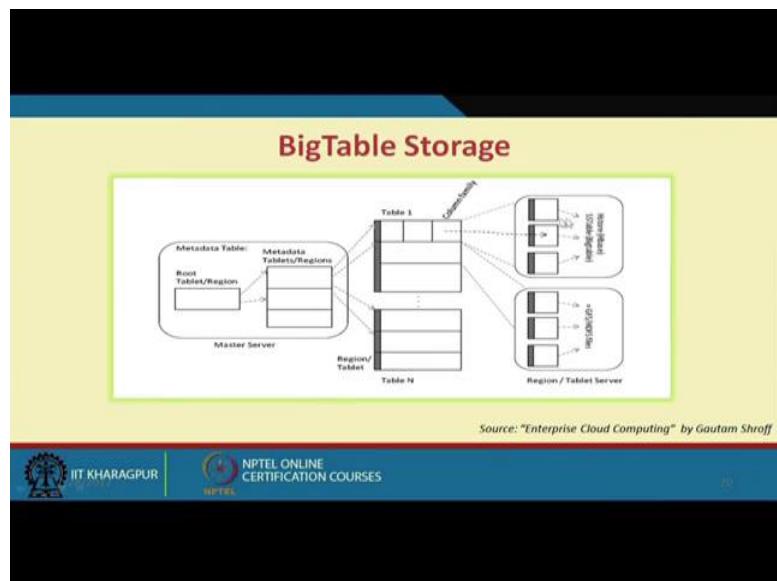
- Each column can store arbitrary **name-value** pairs in the form: **column-family : label**
- Set of possible column-families for a table is fixed when it is created
- Labels within a column family can be created dynamically and at any time
- Each BigTable cell (row, column) can store multiple versions of the data in decreasing order of timestamp
  - As data in each column is stored together, they can be accessed efficiently

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, in big table each column can store arbitrary name value pair in the form of column family and label right. So, here if you can see that these are column families and it is labeled and they store a name value pair. Set of possible column family is of a table is fixed when it is created. So, which are the different column families will be there. So, that is somewhat fix. Labels within a column family can be created dynamically and at any time. So, I can recreate or create the table each big table cell row and column can store multiple versus of the data in decreasing order of the time stamp.

So; that means, it is the chronology is meant it in that fashion. So, it is multiple persons are stored in a decreasing time stamp.

(Refer Slide Time: 26:20)



So, again we see these things. So, there are different tables there are different tablets which are referred to this table and it is a hierarchical structure and we have a master server it is primarily a registry or a meta-data repository. So, each table in big data is split into rangers called tablets, each table is manage by tablet server. So, its stores each column family for a given row range in a separate distributed file called SS table. So, this type of management goes into play. So, that my access rate end of the day the access rate or will be pretty high.

(Refer Slide Time: 27:03)

**BigTable Storage** Contd...

- Each table is split into different row ranges, called **tablets**
- Each tablet is managed by a **tablet server**:
  - Stores each column family for a given row range in a separate distributed file, called **SSTable**
- A single meta-data table is managed by a **Meta-data server**
  - Locates the tablets of any user table in response to a read/write request
- The meta-data itself can be very large:
  - Meta-data table can be similarly split into multiple tablets
  - A **root tablet** points to other meta-data tablets
- Supports large parallel reads and inserts even simultaneously on the same table
- Insertions done in sorted fashion, and requires more work than simple append

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, a single meta-data table is maintained by the many meta-data servers. If the meta-data itself can be very large, it is again broken down into different tablets. A root tablet points to the other meta-data tablets.

So, if the meta-data are repository a pretty large, it is again broken down into different tablets and there is a root tablet which coordinates with your meta-data; this tablets and want to real a want to emulate or realize that meta-data services. Supports large parallel reads and inserts even simultaneously on the same table, insertion done in sorted fashion, requires more work than the simple append, right. There is true for all other databases also because once you insert it is basically you need to push the data aside and create a insertion point where as in case of a append you are putting data at the end of the end of that storage or data or the tables.

(Refer Slide Time: 28:22)

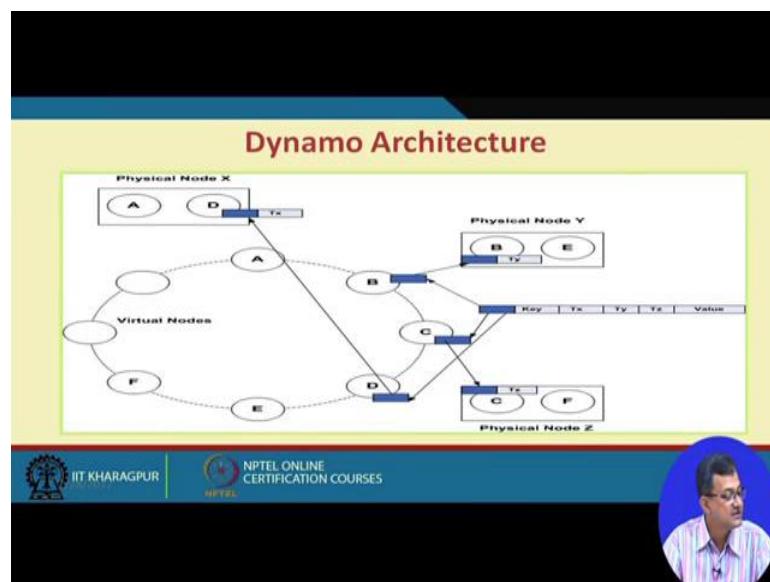
**Dynamo**

- Developed by Amazon
- Supports large volume of concurrent updates, each of which could be small in size
  - Different from BigTable: supports bulk reads and writes
- Data model for Dynamo:
  - Simple <key, value> pair
  - Well-suited for Web-based e-commerce applications
  - Not dependent on any underlying distributed file system (for e.g. GFS/HDFS) for:
    - Failure handling
      - Data replication
      - Forwarding write requests to other replicas if the intended one is down
    - Conflict resolution

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, Dynamo; it is developed by Amazon that supports large volume or concurrent updates each of which can be small in size different from big table supports bulk read and writes right end is. So, data model for Dynamo; it is a simple key value pair well suited for web based e-commerce type of applications and not dependent underlining distributed file systems, right for failure handling conflict resolution, etcetera, they do it their self.

(Refer Slide Time: 28:59)



So, this is typical architecture of the Dynamo where there are several virtual nodes and different physical nodes and they are logical connectivity are zone.

(Refer Slide Time: 29:13)

**Dynamo Architecture** Contd...

- Objects: <Key, Value> pairs with arbitrary arrays of bytes
- MD5: generates a 128-bit hash value
- Range of this hash function is mapped to a **set of virtual nodes** arranged in a ring
  - Each key gets mapped to one virtual node
- The object is replicated at a **primary** virtual node as well as  $(N - 1)$  additional virtual nodes
  - N: number of physical nodes
- Each physical node (server) manages a number of virtual nodes at distributed positions on the ring

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, if you look at the Dynamo architecture. So, it is a key value pair with arbitrary value key value pair with arbitrary arrays of bytes like it uses MD 5 generates a one twenty eight bit1hash table hash value.

So, it basically try to map that were virtual node will be mapping to by using this has function. Range of this has function is mapped as we are discussing that set of virtual nodes arrange in a ring type of thing. The object is replicated as a primary virtual node as well  $N-1$  additional virtual nodes, the N is the number of physical nodes. So, that any the objectives replicated into the things. Each physical nodes are managed is a number of virtual node at a distributed position on the ring. So, if you look at that this physical node server they are basically linked with this virtual node server.

(Refer Slide Time: 30:12)

**Dynamo Architecture** Contd...

- Load balancing for:
  - Transient failures
  - Network partition
- Write request on an object:
  - Executed at one of its virtual nodes
  - Forwards the request to **all** nodes which have the replicas of the object
  - **Quorum protocol:** maintains eventual consistency of the replicas when a large number of concurrent reads & writes take place

 IIT KHARAGPUR |  NPTEL ONLINE CERTIFICATION COURSES

Dynamo architecture, load balancing for transient failure network partition this can handle write request on object that executed at one of its virtual nodes, right.

Forward all the request to all other nodes; it is executed one of the virtual node and say in all other all other nodes which have a replicas of the object so; that means, if I am a object; if it is replicated into another  $N-1$  node. So, one is updated rest are being communicate. So, there is a quorum protocol that maintains eventual consistency of the replicas when a large number of concurrent reads and writes going on. So, this quorum tries to find out that which are the minimum level of replica will be there to handle this large read write of person.

(Refer Slide Time: 31:02)

**Dynamo Architecture** Contd...

- Distributed object versioning
  - Write creates a new version of an object with its local timestamp incremented
  - Timestamp:
    - Captures history of updates
    - Versions that are superseded by later versions (having larger vector timestamp) are discarded
    - If multiple write operations on same object occurs at the same time, all versions will be maintained and returned to read requests
    - If conflict occurs:
      - Resolution done by application-independent logic

So, in next, we are having this dynamo distributed object version right creates a new version of the objects in his local time stamp created. There are algo for column consistency.

(Refer Slide Time: 31:12)

**Dynamo Architecture** Contd...

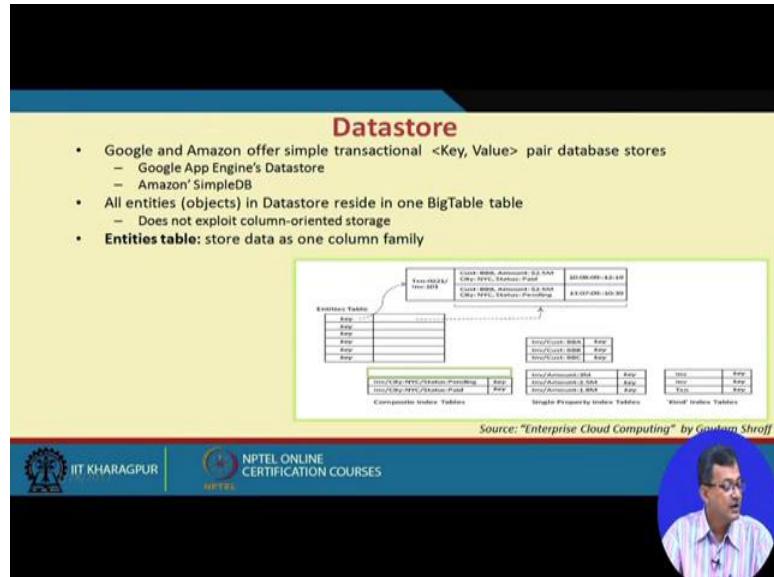
- **Quorum consistent:**
  - Read operation accesses  $R$  replicas
  - Write operation access  $W$  replicas
    - If  $(R + W) > N$  : system is said to be **quorum consistent**
  - Overheads:
    - For efficient write: larger number of replicas to be read
    - For efficient read: larger number of replicas to be written into
- **Dynamo:**
  - Implemented by different storage engines at node level: Berkley DB (used by Amazon), MySQL, etc.



So, read operation R; write operation E. So, read plus write operation should be greater than any of the system is quorum consistent there are overheads which will be coming there is a efficient write large number of replicas are to be read and if it is for a, b, c and read large number of large number of replicas need to be written. So, these are the 2

things which are they are; so, it is implemented by different storage engines at node level Berkley DB used by Amazon and can be implemented to using MySQL and etcetera.

(Refer Slide Time: 31:51)



Another; the final concept what we are having is the data store. Google and Amazon of a simple traditional key value pair database stores, right, Google app engines data store in case of Amazon what we say simple DB; all entities objects in the data store reside on in one big table, right.

Data store exploit column oriented storage right, data store as I mean store data as a column families. So, unlike our rational traditional thing is a more of a row family or tuple based it is called column family.

(Refer Slide Time: 32:30)

**Datastore contd...**

- Multiple index tables are used to support efficient queries
- BigTable:
  - Horizontally partitioned (also called **sharded**) across disks
  - Sorted lexicographically by the key values
- Beside lexicographic sorting Datastore enables:
  - Efficient execution of **prefix** and **range** queries on key values
- Entities are 'grouped' for transaction purpose
  - Keys are lexicographic by group ancestry
    - Entities in the same group; stored close together on disk
- Index tables: support a variety of queries
  - Uses values of entity attributes as keys

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, there are several advantages or several features or characteristics like multiple index tables are used to support efficient query. Big table horizontally partitioned call sharded and across the disk whereas, stored lexicographically in the key values other thing. Beside lexicographic sorting of the data enables there is a execution of prefix and range queries on key values entities are grouped for transactional purpose because if there is if when we are having transaction. So, that is a set of entities which are accessed in a more frequent way and index table to support varied varieties of queries.

So, we can have different indexes or different type of queries. So, it is not we should understand is not a simple a low a database it is a large database. So, in order to do that; I cannot churn the whole database. So, need to slice them appropriately. So, that based on the different variety different queries it can be executed more efficiently.

(Refer Slide Time: 33:37)

**Datastore** Contd...

- Automatically created indexes:
  - Single-Property indexes
    - Supports efficient lookup of the records with **WHERE** clause
  - ‘Kind’ indexes
    - Supports efficient lookup of queries of form **SELECT ALL**
- Configurable indexes
  - Composite index:
    - Retrieves more complex queries
- Query execution
  - Indexes with highest selectivity is chosen

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

And there are few more properties like automatically it creates indexes single property index or there is a kind index supports the efficient lookup queries of form select all type of things the configurable in indexes and there is a query execution indexes with highest selectivity is chosen, right. So, it is when we do the query execution.

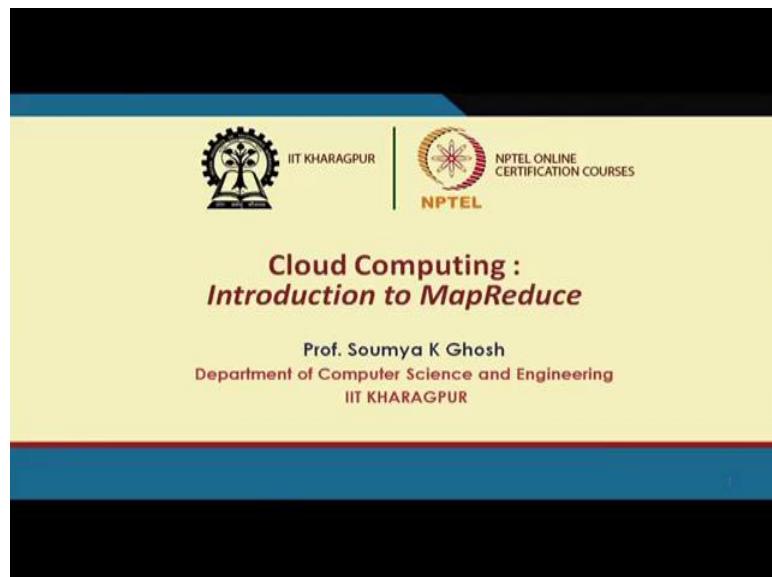
So, with this we will stop our discussion here. So, what we tried to discuss over see is there different aspects we have the notion of our traditional databases which is established, fault tolerant, efficient and there are different mechanism to do that. So, we have we have also already this parallel execution things and its present. So, when we deal with a large volume of data in the cloud which are likely to be there, then what is are the different aspects we need to look at. So, we may not be able to follow the this column oriented or tuple oriented relational database we need to a sorry row oriented database we need to four for column oriented data base and there are different file system like GFS, HDFS and over that this data store Dynamo and your simple DB and those things what which are being implemented by various inter cloud service providers CSPs for efficient storage access, nread write execution of very very large databases.

Thank you.

**Cloud Computing**  
**Prof. Soumya Kanti Ghosh**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**

**Lecture – 14**  
**Introduction to Map Reduce**

(Refer Slide Time: 00:40)



Hello, so we will continue our discussion on cloud computing. As in our previous lecture we discussed about data store or data how to manage data in cloud having an overview of the things. Now, we like to see that another programming paradigm which is called MapReduce right a very popular programming paradigm which is primarily level out by Google, but now being used for different scientific purposes. So, Google primarily developed it for their large scale searches search engines primarily to search on huge amount of volumes of documents which their Google search engines chants, but it becomes an important paradigm programming paradigm for this scientific world to work on to exploit this philosophy to efficiently execute for different type of scientific problems.

(Refer Slide Time: 01:30)

**Introduction**

- MapReduce: programming model developed at Google
- Objective:
  - Implement large scale search
  - Text processing on massively scalable web data stored using BigTable and GFS distributed file system
- Designed for processing and generating large volumes of data via massively parallel computations, utilizing tens of thousands of processors at a time
- Fault tolerant: ensure progress of computation even if processors and networks fail
- Example:
  - Hadoop: open source implementation of MapReduce (developed at Yahoo!)
  - Available on pre-packaged AMIs on Amazon EC2 cloud platform

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, map reduce is a programming model developed at Google, primarily objective was to implement large scale search, text processing on massively scalable web data stored in using big table or and GFS distributed file system. So, as we obtained that big data and GFS distributed file systems the data are stored. So, how to process this massively scalable web data that means, the huge volume of data are coming into play. Design for processing and generating large volumes of data via massively parallel computation utilizing tens of thousands of processor at a time.

So, I have a large pool of processors a huge pool of data and I want to do some analysis out of it. So, how can I do it? So, one very popular problem what we see is that if I have a huge volume of data and number of processors then how do say want to do some sort of word counting or counting the frequency of some of the words in that huge volume of data like I want to find out that how many times IIT Kharagpur appears in this a huge chunk of data, which are primarily stored in this HDFS or GFS or big table type of architecture.

So, and it should be fault tolerant, ensure progress of the computation even if processor fails and network fails right. So, because as there are huge volume huge number of processors and say underlining networks, so I do ensure fault tolerant. So, one of the example is Hadoop open source implementation of MapReduce developed at time volume had over initially developed at Yahoo and then became a open source. Available

in a pre packaged AMIs on Amazon EC 2 platform, right. So, we are what we are looking at is trying to give a programming provide a programming platform or programming paradigm which can interact with data basis which are stored in this sort of cloud data stores right, it can be HDFS, GFS type or managed by big table and so on so forth.

(Refer Slide Time: 04:00)

**Parallel Computing**

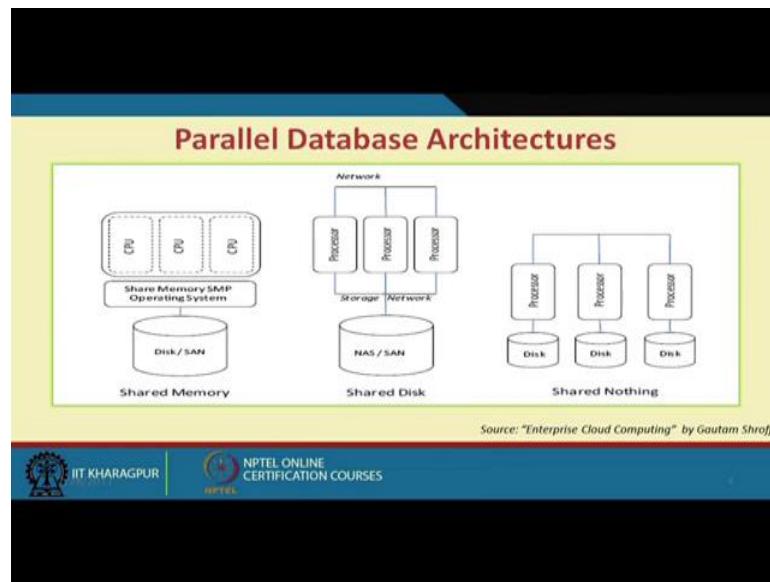
- Different models of parallel computing
  - Nature and evolution of multiprocessor computer architecture
  - Shared-memory model
    - Assumes that any processor can access any memory location
    - Unequal latency
  - Distributed-memory model
    - Each processor can access only its own memory and communicates with other processors using message passing
- Parallel computing:
  - Developed for compute intensive scientific tasks
  - Later found application in the database arena
    - Shared-memory
    - Shared-disk
    - Shared-nothing

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, if we look at again parallel computing as we have seen in our previous lectures, so different models of parallel computing it depends on the nature and evolution of the processor, multiprocessor computer architecture. So, it is shared memory model, distributed memory model, so these are the 2 popular thing. So, parallel computing developed for computing, intensive scientific tasks as we all know; later found application in data base arena or data base paradigm also, right.

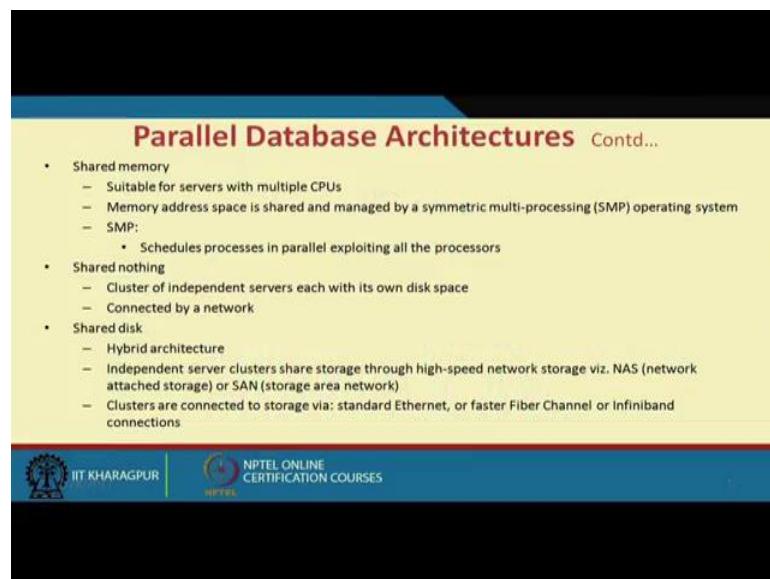
So, it was initially it is more of a doing a huge scientific task and later we have seen that it has a lot of application in the database domain too. And we have seen in our earlier lecture that we have three type of scenario one is shared memory, shared disk and shared nothing, right. So, whenever we want to do a programming paradigm or work on something which can work on this sort of parallel programming paradigm where the data stored in the different this sort of clouds storages, so we need to take care of that what sort of mechanism is there. Like, whether it is shared memory, shared disk or shared nothing type of configuration.

(Refer Slide Time: 05:28)



This is the picture already we have seen in our earlier lectures, so we do not want to repeat. So, it is a shared memory structure, shared disk and shared nothing, but the perspective we are looking at now is little different. There it is more of the storage where we are looking at. Now, we are trying to look at that how the programming can exploit this type of structure.

(Refer Slide Time: 05:52)



So, this is already we have seen; so, shared memory suitable for servers with multiple CPUs. Shared nothing cluster of independent server each with its own hard disk, so

connected by a high-speed network. And shared disk, so it is a hybrid architecture independent server cluster shares storage through a high speed network storage like NAS or SAN. Clusters are connected via to storage via standard Ethernet, fast fiber channel infini-band and so on and so forth.

So, whenever we do anything parallel or anything parallel computing or parallel storing and type of things, what is our back of the mind is to have efficiency right. So, you want to do parallelism to one of the major aspect is to have a efficiency. There may be other aspects of fault tolerance and full proof and failure register and type of thing, but primarily it should be efficient. Now, first of all the type of work we are doing there should be inherent parallelism into it. If there is no inherent parallelism, then it may not be fruitful to do using always a parallel architecture.

So, first of all they should be inherent parallel it should not be a sequence of operations and then you try to do a parallel. So, it is job 1, job 2, job 3, job 4 a sequence is there or in between some parallelism is there, but if you want to make a parallel operation, there may not be.

(Refer Slide Time: 07:22)

**Parallel Efficiency**

- If a task takes time  $T$  in uniprocessor system, it should take  $T/p$  if executed on  $p$  processors
- Inefficiencies introduced in distributed computation due to:
  - Need for synchronization among processors
  - Overheads of message communication between processors
  - Imbalance in the distribution of work to processors
- Parallel efficiency of an algorithm is defined as:

$$\epsilon = \frac{T}{P \cdot T_p}$$

**Scalable parallel implementation**

- parallel efficiency remains constant as the size of data is increased along with a corresponding increase in processors
- parallel efficiency increases with the size of data for a fixed number of processors

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, if a task takes time  $T$  in uni-processor, it should take  $T / p$  if executed in  $P$  processor ideally if the parallelism is there, and we are thinking that there is no cost in dividing distributing and type of things. So, it ideally  $T / p$  is a something ideal condition we can have. So, inefficiencies introduced in distributed computation due to need of

synchronization among the processors. So, I need to synchronize among the processor, it is not like that all processor has you may have the individual clocks and you need to synchronize that where things will be there. Otherwise if you if you divide the job into 2 where one executed now and one executed after couple of hours then it is it could have been better that is execute to in a one system. So, synchronization in between the processor is one of the important aspects. So, need to synchronize.

Overheads of message communication between the processors another aspect; imbalance in the distribution of work to the processors another, so it may not be equally divided and type of things. So, these are the different aspects which indirectly affect this efficiency or bring about inefficiency into this parallel implementation. So, parallel efficiency of an algorithm can be defined as  $\frac{T}{p \cdot T/p}$ . So, if it is scalable we say this is scalable or scalable parallel efficiency remains constant as the size of the data increased along with a corresponding increase of the processor.

So, what is happening when more data is coming, so you go on deploying more processor or you go on requesting from the cloud more processor and then your efficiency remains constant, the efficiency values does not change? So, then what we say that it is scalable. So, that if I increase both say for example, for linearly then it goes on in the constant thing. Parallel efficiency increases with the size of the data for a fixed number of processor, it increases with the size of the data; and if it is a fixed number of processor then we can have effectively more efficiency.

(Refer Slide Time: 09:46)

**Illustration**

- **Problem:** Consider a very large collection of documents, say web pages crawled from the entire Internet. The problem is to determine the frequency (i.e., total number of occurrences) of each word in this collection. Thus, if there are  $n$  documents and  $m$  distinct words, we wish to determine  $m$  frequencies, one for each word.
- Two approaches:
  - Let each processor compute the frequencies for  $m/p$  words
  - Let each processor compute the frequencies of  $m$  words across  $n/p$  documents, followed by all the processors summing their results
- Parallel computing is implemented as a distributed-memory model with a shared disk, so that each processor is able to access any document from disk in parallel with no contention

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Now, the example, which is there in that book you are referring also you will find the example in different literature, this sort of example not the same. Consider a very large collection of documents say the web document crawled by the entire internet. So, it is a pretty large it is large every day it is growing. The problem is to determine the frequency that is total number of occurrences of each word in this collection right. So, I want to determine the total number of what is the frequency of occurrences of each word in this document d. So, thus if there are n documents and m distinct words, we use to determine m frequency one for each word right. So, this is a simple problem may be true or may be more relevant for search engines and type of things.

So, we have two approaches let each processor compute the frequency for  $m/p$  words. So, each processors if there are p processors, if the m frequencies I need to calculate, I divide m by p, so many, so for example, I want to look for I have ten processors and I have no I am look for some 90 words, m equal to ninety. So, every processor does it chunk of ten right roughly if it is not divisible then you have to make some asymmetric division. So, it makes things and that at the end of things they again report the things together or in some through some system. So, other way let each processor compute the frequency of m words across n by p documents.

So, total number of documents say 10,000. So, 10,000 number of words I am looking for 90 number of processor I am having 10. So, one is 90 by 10 is the 9 words on an average given to the every processor and they count on the things.

Other thing what we are telling that each processor compute the frequency for all the 90 words, but on n by p document if that 10,000 words and then p 10 processors; so, some thousand document so each take 100 documents and do the processing and once that frequency of this m words by individual professors processors come out then I sum up this thing and aggregate and show the result that this is the thing right, by followed by all processors summing their results right. Parallel computing, now which one will be efficient based on this parallel computing paradigm, we need to look at right. So, parallel computing is implemented as a distributed memory model with a shared disk, so that each processor is able to access any document from the disk in parallel with no contention. So, this can be one of the implementation mechanisms.

(Refer Slide Time: 13:01)

**Illustration** Contd...

- Time to read each word from the document = Time to send the word to another processor via inter-process communication =  $c$
- Time to add to a running total of frequencies  $\rightarrow$  negligible
- Each word occurs  $f$  times in a document (on average)
- Time for computing all  $m$  frequencies with a single processor =  $n \times m \times f \times c$
- First approach:
  - Each processor reads at most  $n \times m/p \times f$  times
  - Parallel efficiency is calculated as:
  - Efficiency falls with increasing  $p$
  - *Not scalable*

$$\epsilon_a = \frac{nmfc}{pnmf} = \frac{1}{p}$$

 IIT KHARAGPUR |  NPTEL ONLINE CERTIFICATION COURSES

Now, time to read each word from the document say if let us assume that time to read each word from the document equal to time to send the word to another processor via inter processor communication and equals to  $c$ . So, making thing simple so it may be it should be means ideally in ideal case or in a real life case it will be different, but we make these scenarios. So, first approach so time for so time to add to a running total of

the frequencies negligible, so summing up is negligible. Once I find the frequencies of this m word then summing up is negligible.

Each what word occurs f times on the document on an average. So, if I for our calculation sake that each word that on an average, workers some f time. Time for compute all m frequencies with the single processor equal to then I have  $n \times m \times f \times c$ . So, this is the time to compute m frequencies with a single processor, if I have a single process this could have been the thing. So, if we do the first approach, first approach was this one, let each processor compute the frequency of  $m/p$  words, so that is a first approach.

So, each processor reads at most  $\frac{n \times m}{p \times f}$ . So, parallel efficiency is calculated as

$\frac{n \times m \times f \times c}{p \times n \times m \times f \times c}$ , so 1 by p very vanilla type consideration. So, we take that all are doing

all are morally same frequencies, all are negligible time for the any aggregation then the all time for the means read and write another operations we have to consider c, considering this we are getting  $\frac{1}{p}$ . So, efficiency falls with increasing p. So, if we

increase the p, then the efficiency falls. So, it is not constant. So, it is not scalable, it is one of the major problem is that though it is what we say easy to conceptualize etcetera, but there is a problem in the scalability so of the things. This one that let each processor compute frequencies per n by m words n by m words is not scalable.

(Refer Slide Time: 15:36)

**Illustration** Contd...

- Second approach
  - Number of reads performed by each processor =  $n/p \times m \times f$
  - Time taken to read =  $n/p \times m \times f \times c$
  - Time taken to write partial frequencies of  $m$ -words in parallel to disk =  $c \times m$
  - Time taken to communicate partial frequencies to  $(p - 1)$  processors and then locally adding  $p$  sub-vectors to generate  $1/p$  of final  $m$ -vector of frequencies =  $p \times (m/p) \times c$
  - Parallel efficiency is computed as:

$$\epsilon_p = \frac{nmfc}{p \left( \frac{n}{p} mfc + cm + p \frac{m}{p} c \right)} = \frac{nf}{nf + 2p} = \frac{1}{1 + \frac{2p}{nf}}$$

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Whereas, in the second approach, where that  $m$  words we divide into the different processes oh sorry we divide that document  $d$ , whereas every processor compute this for all the  $m$  words and then aggregate. So, apparently what it looks that this could me more costly. So, it is there is a aggregation thing then you are doing clubbing those processor, club means dividing the  $m$  set into different this whole documents set into different partitions and doing that, this could be in efficient than the first one. But let us see what is there. So, the number of read performs for each processor is  $n/p \times m \times f$  right the time taken to read is  $n/p \times m \times f$ . It is because you are having  $n/p$  amount of volume of the data and then want to calculate for  $m \times f \times c$ , so that number of time taken to calculate this read. Time taken to write partial frequency on of  $m$  words in parallel to disk is  $m(c \times m)$ .

So, once you are done you need to write on the parallel to the disk and that is that comes to be  $(c \times m)$  time taken to communicate partial frequency right to  $p-1$  processors. And then locally adding sub  $p$  sub vectors to generate  $1/p$  of the final  $m$  vector of frequencies then what we have  $p \frac{m}{p} c$ . So, what you need to do we are time taken to communicate partial frequencies right because you do not have the whole frequencies. So, partial frequency by different processor and  $p-1$  processor and then locally adding  $p$  sub

vectors to generate 1 by p here of the final m vector frequencies is this one. So, individually need to do.

So, if we adopt all those things in case of this second approach what we have this parallel frequency as this structure, so  $\frac{1}{1+2p/nf}$ , so that is if you if you look at it little minutely

if you consult the book, it is not a very difficult problem difficult to deduce. It is pretty easy just have to go by step by step. Now, this is an interesting phenomena. So, the term we are having here is  $1+2p/nf$ . So, in this case a p is many, many times less than nf. Efficiency of the second approach is higher than that of the first right here if it is p is many, many times less than nf, then this term that this will be tending towards one. And it can be seemed that there is much efficiency is much higher.

(Refer Slide Time: 18:57)

**Illustration** Contd...

- Since  $p \ll nf$ , efficiency of second approach is higher than that of first
- In fist approach, each processor is reading many words that it need not read, resulting in wasted work
- In the second approach every read is useful in that it results in a computation that contributes to the final answer
- Scalable
  - Efficiency remains constant as both  $n$  and  $p$  increases proportionally
  - Efficiency tends to 1 for fixed  $p$  and gradually increased  $n$

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

In the first approach, so there is a type it should be, let us in the in first approach each processor is reading many words than it needs to read resulting in wastage of time. What we have done in the first approach this many processor we have divided this m into different chunk. So, the processor say as we as we have taken the example that if I am having m as ninety and number of processor is p, so 90 by p is 10. So, everybody is getting 10, but when it is searching the whole document, so number of documents is reading where there is no hit, it is no success.

So, efficiency, so in the second approach every read is useful right. As it results in a computation and distributes to the final results. So, for in the second approach, every read is likely to be useful where it contribute to this result. So, it is scalable also. The efficiency remains constant at both n and p increases potentially, they proportionally. So, what we see what we have done there that if my data load increases I will increase the processor. So, if I proportionally increase the data processor then my efficiency remains constant in this case in the second case. Efficiency tends to one for fixed p and gradually increasing n. So, efficiency tends to 1, if the number of processor is fixed and gradually increased we are increasing n that means we are increasing the data load, number of processor fixed and it will basically approaches one.

(Refer Slide Time: 20:57)

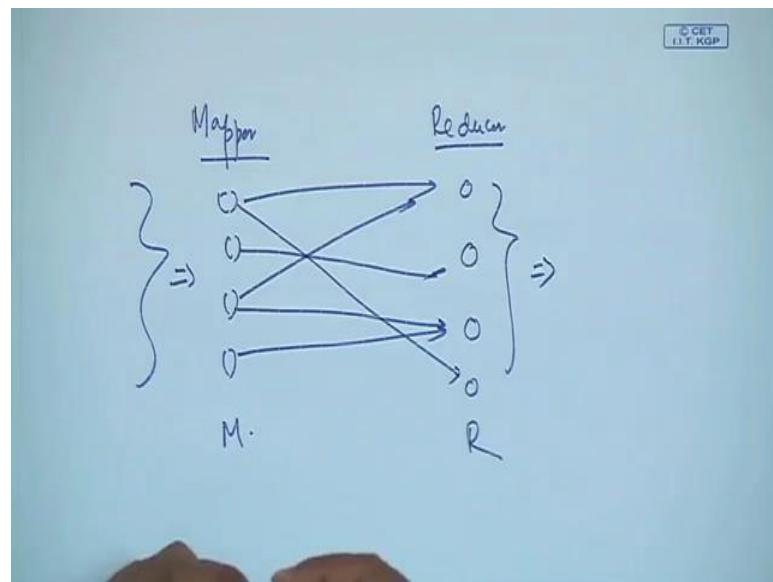
**MapReduce Model**

- Parallel programming abstraction
- Used by many different parallel applications which carry out large-scale computation involving thousands of processors
- Leverages a common underlying fault-tolerant implementation
- Two phases of MapReduce:
  - Map operation
  - Reduce operation
- A configurable number of M 'mapper' processors and R 'reducer' processors are assigned to work on the problem
- Computation is coordinated by a single master process

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, with these context or with these background of that which can be that this doing that individually then aggregating is becoming more efficient with this things, we look at that your map reduce model. So, it is a parallel programming abstraction used by many different parallel applications which carry out large scale computations involving thousands of processors; leverages a common underlining fault tolerant implementation. Two phases of map reduce map operation and reduce operation. A configurable number of M mapper - mapper processor and R reducer processors are assigned to work on the problem. Computation is coordinated by a single master process. So, what we are having now? There are different mapper processors like and there is a different reducer processor. So, whole process, I divide into two things.

(Refer Slide Time: 22:16)



Like I have a mapper, so different mapper processor, so there are M processor and there is reducer. So, there are different reducer processor. So, what we do is when the data come here it basically do some execution and then this reducer may be based on the type of problem it will go on different reduce things and do the execution. So, reducer will generate more of aggregated results right. So, what it tries to do is a parallel programming abstraction used by many parallel applications which carryout large scale computation involving thousands of processors. So, here the application come into play. So, it is a two phase process, one is a map operation, another is a reduce operation. So, that the configurable number of M mapper processor, R reducer processors, so it is configurable; that means, you can have more etcetera mapper and reducer.

(Refer Slide Time: 23:28)

**MapReduce Model** Contd...

- Map phase:
  - Each mapper reads approximately  $1/M$  of the input from the global file system, using locations given by the master
  - Map operation consists of transforming one set of key-value pairs to another:  
Map:  $(k_1, v_1) \rightarrow [(k_2, v_2)]$ .
  - Each mapper writes computation results in one file per reducer
  - Files are sorted by a key and stored to the local file system
  - The master keeps track of the location of these files

Map:  $(k_1, v_1) \rightarrow [(k_2, v_2)]$

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, map reduce phase. So, if we look at the map phase each mapper read approximately  $1/M$  of the input from the global file. So, it is not the whole data  $d$ , but a chunk of the data read. Map operation consists of transforming one set of value key value pair to another set of key value pair. So, what map does, it is a one set of key value pair to another set of key value pair. So, map( $k_1, v_1 \rightarrow [(k_2, v_2)]$ ). So, each mapper writes computational results in one file per reducer. So, what it does, it basically for every reducer it produces a file. So, it says if there are reducers R 1, R 2, R 3 a mapper m, I create three files based on the corresponding the reducer. So, the files are sorted by a key and stored in a local file systems right. The master keeps tracks of the location of these files. So, there is a master map reduce master, so which takes care of this location of the file, each mapper produces a one file for every reducers and the master takes care where the files are stored in the local disk etcetera.

(Refer Slide Time: 24:55)

**MapReduce Model**  
Contd...

- **Reduce phase:**
  - The master informs the reducers where the partial computations have been stored on local files of respective mappers
  - Reducers make remote procedure call requests to the mappers to fetch the files
  - Each reducer groups the results of the map step using the same key and performs a function  $f$  on the list of values that correspond to these key value:

$$\text{Reduce: } (k_2, [v_2]) \rightarrow (k_2, f([v_2])).$$

- Final results are written back to the GFS file system

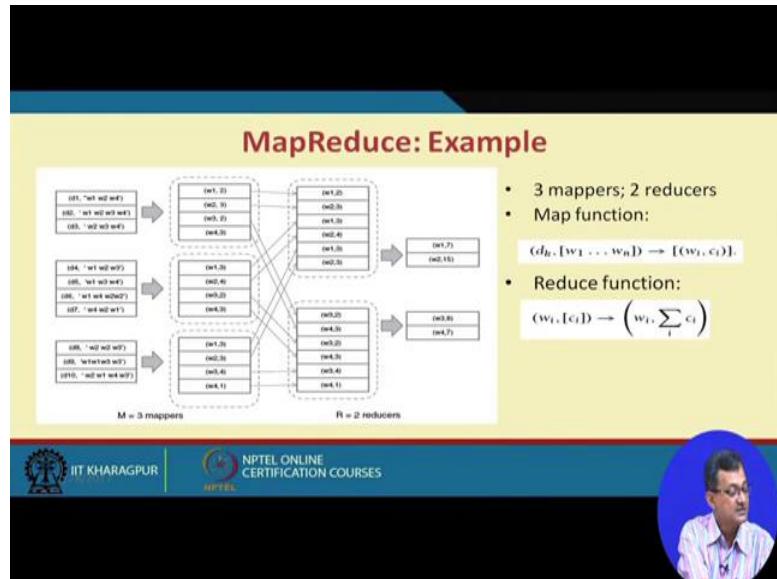
IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES



In the reduce phase, the master informs the reducers where the partial computation have been stored on local file systems of respective mappers; so that means, in the reducer phase the reducer consult this master which informs that where its related files are stored corresponding to the every mapper functions. Reducer makes remote procedure call to the mappers to fetch the files. So, reducer in turn make a remote procedure call for the mapper. So, mapper it is somewhere in the disk and the reducer there may be in different structure with different types of VMs etcetera running on the things ideally it is not far not geographically distributed then the things will not work. So, nevertheless it is working on that particular data which are produced by the mapper.

So, each reducer groups the results of the map step using the same key value key value function  $f$  etcetera, so  $(k_2, [v_2]) \rightarrow (k_2, f([v_2]))$ . So, here the aggregated functions in comes into play. In other sense, if we remember our problem. So, what we do that every doc, every key or every word we want to calculate the frequency, so the functional model is summing up the frequencies of the things, it can be different for different type of things. So, it does a  $k_2 v$  etcetera. So, it goes for another key value up here. Final results are return back to the GFS file system Google file system.

(Refer Slide Time: 26:36)



So, map reduce example. So, if we see there are 3 mapper, 2 reducer. So, map function in this in our case is that is the data  $d$  there are the set of word  $w_1, w_2, w_n$  and it produce for every  $w_i$  the count of the things, how much count the portion of the mapper it is having. So, every  $w_i$ , it counts the thing. So, if you see if  $d_1$ , it has  $w_1, w_2, w_4$ ;  $d_2$  these are the things and it counts this. So, every mapper does it, and then it basically stored in a intermediate space where the reducer reads. So, it generates every file for every reducer like this particular things is generate a particular file for a reducer. So, there are two reducer.

So, for two reducer every mapper generates the file. So, and the reducer in turns basically accumulate those. So, it says that  $w$  it has the thing  $w_1, w_2$ , so  $w_1$  as 7,  $w_2$  as something 15. In this case,  $w_3, w_4$  are the other two. So, the reducers reduces the thing from the inputs of the or from the outputs of the mapper getting the input from the mappers output.

(Refer Slide Time: 28:08)

**MapReduce: Fault Tolerance**

- Heartbeat communication
  - Updates are exchanged regarding the status of tasks assigned to workers
  - Communication exists, but no progress: master duplicate those tasks and assigns to processors who have already completed
- If a mapper fails, the master reassigns the key-range designated to it to another working node for re-execution
  - Re-execution is required as the partial computations are written into local files, rather than GFS file system
- If a reducer fails, only the remaining tasks are reassigned to another node, since the completed tasks are already written back into GFS

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, map reduce model is fault tolerance; there are different way to look at it, one is heart beat message. So, every particular time period, it says that whether it is a live and type of things. Communication exists, but no progress master if there are communication exists, but no progress master duplicate those tasks and assign the processor who are already completed or some free processors. If the mapper fails, the mapper reassigns key value designated to it to another work node on the re-execution. So, if it is a failure then it re-execute the thing. If the reducer fails only the remaining task need to be reassigned to another node. Since the completed tasks are already written back to Google file system. So, if the completed tasks are there, they are already in Google file systems only the remaining tasks need to be reassigned.

(Refer Slide Time: 29:04)

**MapReduce: Efficiency**

- General computation task on a volume of data  $D$
- Takes  $wD$  time on a uniprocessor (time to read data from disk + performing computation + time to write back to disk)
- Time to read/write one word from/to disk =  $c$
- Now, the computational task is decomposed into map and reduce stages as follows:
  - Map stage:
    - Mapping time =  $c_m D$
    - Data produced as output =  $\sigma D$
  - Reduce stage:
    - Reducing time =  $c_r \sigma D$
    - Data produced as output =  $\sigma \mu D$

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, if you want to calculate the efficiency of the MapReduce, so the general computation task on a volume of data  $D$ . So, takes  $wD$  time to uni-processor read time to read data from disk performing computation write back to the disk. Time to read write one word from to disk is  $c$ . Now, the computation task is decomposed into map reduce stages like map stage mapping time  $c_m D$  data producing and output  $\sigma D$ , reduce stage reduce time  $c_r \sigma D$  and data produced at the output is  $\sigma \mu D$ . So, this is not that difficult. So, mapping time how much that with  $D$  every mapper is doing data produced time is from the particular mapper which is how much time it is producing reduce reducers time in calculated with the every  $c_r$  and that finally, we have that reducer output.

(Refer Slide Time: 30:11)

**MapReduce: Efficiency** Contd...

- Considering no overheads in decomposing a task into a map and a reduce stages, we have the following relation:  $wD = cD + c_mD + c_r\sigma D + c\sigma\mu D$
- Now, we use  $P$  processors that serve as both mapper and reducers in respective phases to solve the problem
- Additional overhead:
  - Each mapper writes to its local disk followed by each reducer remotely reading from the local disk of each mapper
- For analysis purpose: time to read a word locally or remotely is same
- Time to read data from disk by each mapper =  $\frac{wD}{P}$
- Data produced by each mapper =  $\frac{\sigma D}{P}$

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, considering no overheads in decomposing the task into map and reduce stages, we can have the following relationship. So, if we forget the overhead in decomposing in mapping and reducing, so we can have this summation of the things. Now, if we had  $P$  processors that serve as both mapper and reducer right irrespective of the phases to solve problem. So, if we use  $P$  processor sometimes it acts as a mapper, sometimes act as a reducer. Then we have additional overhead each mapper writes to local we have some additional overheads writes to local disk followed by each reducer remotely reading to the disk. For analysis purpose time to read to a word locally or remotely, let us consider as same. Time to read a data from the disk is for each mapper is  $wD$  by number of with an if the number of processor is  $P$   $wD/P$  data producer is mapper is  $\sigma D/P$ .

(Refer Slide Time: 31:12)

**MapReduce: Efficiency** Contd...

- Time required to write into local disk =  $\frac{c\sigma D}{P}$
- Data read by each reducer from its partition in each of  $P$  mappers =  $\frac{\sigma D}{P^2}$
- The entire exchange can be executed in  $P$  steps, with each reducer  $r$  reading from mapper  $r + i \bmod r$  in step  $i$
- Transfer time from mapper local disk to GFS for each reducer =  $\frac{c\sigma D}{P^2} \times P = \frac{c\sigma D}{P}$
- Total overhead in parallel implementation due to intermediate disk reads and writes =  $\left(\frac{wD}{P} + 2c \frac{\sigma D}{P}\right)$
- Parallel efficiency of the MapReduce implementation:

$$\epsilon_{MR} = \frac{wD}{P\left(\frac{wD}{P} + 2c \frac{\sigma D}{P}\right)} = \frac{1}{1 + \frac{2c}{w}\sigma}$$

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, time required to write back to the disk because once you read then you have to after computation, you have to write back to that is that much. So, similarly data read by each reducer from its partition to each mappers  $P$  mappers are  $\sigma D/P/P$ . So, if we calculate like that we say that the parallel efficiency of the map reduce implementation

comes as this one,  $\frac{1}{1 + \frac{2c}{w}\sigma}$ .

(Refer Slide Time: 31:50)

**MapReduce: Applications**

- Indexing a large collection of documents
  - Important aspect in web search as well as handling structured data
  - The map task consists of emitting a word-document/record-id pair for each word:  $(d_k, [w_1 \dots w_n]) \rightarrow [(w_i, d_k)]$
  - The reduce step groups the pairs by word and creates an index entry for each word:  $[(w_i, d_k)] \rightarrow (w_i, [d_{i_1} \dots d_{i_m}])$
- Relational operations using MapReduce
  - Execute SQL statements (relational joins/group by) on large data sets
  - Advantages over parallel database
    - Large scale
    - Fault-tolerance

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Now, so this is what we get a parallel efficiency out here. Now, if the indexing map reduce there are several type of applications one is indexing a large collection of documents right, so that which is primarily one of the major motivation for Google. So, important aspect for web search as well as handling structured data. So, map task consists of emitting a word document, record id pair for each word like as we have seen  $wdk.w_1.n$  into map to one its map  $w_1$  into every word dks. So, I can have some sort of indexing reduce step groups the pair of words and creates entry into the thing.

So, there are applications in relational operations using map reduce. Execute SQL statements relational join, group by on large set of data. Advantages of parallel data base large scale fault tolerance we want to exploit and I can have those type of function like as we have seen that it is a group by clause and type of etcetera we can do, so that some sort of relational operations we can execute.

So, with these we come to this end of today's talk. So, what we try to do here to give you a overview of a MapReduce paradigm that how a problem can be divided into a set of parallel executions, which is a mapper node which creates intermediate results. And there is a set of reducer nodes which takes this data and create the final results right. And what we can which there are some of the things which is interesting that the mapper creates data file for every reducer. So, it is the data is created per reducer. So, the reducer knows that where the data is there.

Over and above there is a master controller or the map reducer master things which come to which knows where there things where the data is stored by the mapper and how the reducer will read. Not only that if the mapper node fails how to reallocate the things; if the reducer node fails, how to reallocate because the things or the reallocate the not executed data not executed things not executed yet to be executed operations and so on and so forth. So, with this we will stop our lecture today.

Thank you.

**Cloud Computing**  
**Prof. Soumya Kanti Ghosh**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**

**Lecture – 15**

**Open Stack**

Hello, so we will continue our discussion on cloud computing. Today we will discuss and so a demo on a open source cloud which is open stack; one of the very popular open source cloud. So, we will initially we will have some brief overview of open stack then we will; so, a demo on open stack; that how way open stack can be configured VM can be provision. Primarily, we have been used these open stack for IAS type of cloud infrastructure as a service type of cloud; that we will demonstrate today.

So, if you see open stack one of the very popular open source cloud; which you can download and install in a particular hardware configuration even with couple of species with you can install your open stack and see the performance of IAS type of cloud.

(Refer Slide Time: 01:28)

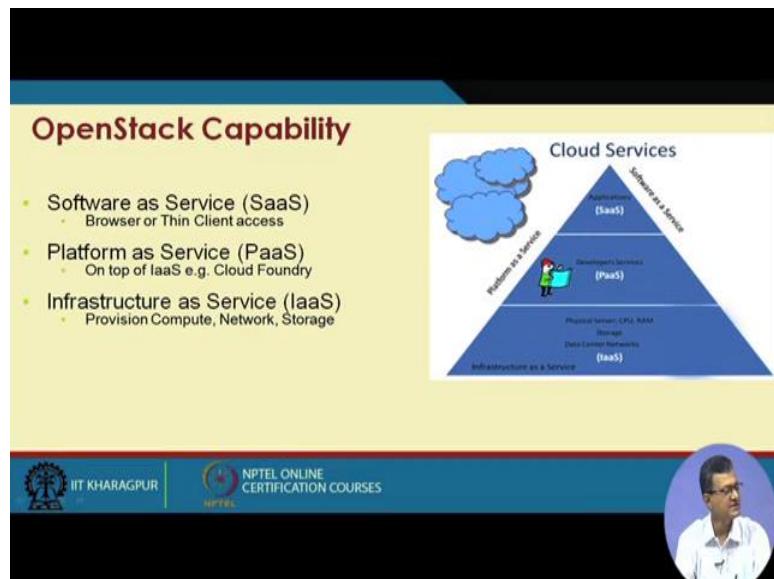


So, open stack is a cloud operating system that controls large pool of compute storage networking resources through a data centre; throughout a data centre, all manage through dashboard and gives administrator control and while empowering their users provisioning your resources through a wave interface.

So, what it says; it's say you have a set of resources it is a layer above those base this bare metal resources and it gives the administrator to control these resources, and the user can basically provisions VMs out of it. So, it access a IAS type of cloud and as it is as we mentioned it is a open source. So, you can download and install and give it provision of the things. Incidentally in IIT Kharagpur, we have done experimental; we have made experimental cloud called Meghamala, which is based on open stack and which has been installed over blade server.

But never the less open stack you can install over PC's; here also we have tried with PC's etcetera for small-small student projects. So, it gives a feel that how a particular infrastructure as a cloud works and also how we can work on those things.

(Refer Slide Time: 02:57)



So, as far as the open stack capability is there; so, it has a capability of all the services though primarily we use more as the infrastructure as a service. So, at the infrastructure as a service provision come compute network storage at the pass level on top of IAS cloud foundry and over the as the SaaS level that browser or thin client accesses. So, these are the whole pyramid of typical cloud services; which is accessible over internet. So, infrastructure as a service where you have physical serve as like CPU, RAM, storage, data center networks, etcetera; over that some developers service and over that some software as a service.

(Refer Slide Time: 03:44)

**OpenStack Capability**

- Virtual Machine (VMs) on demand
  - Provisioning
  - Snapshotting
- Network
- Storage for VMs and arbitrary files
- Multi-tenancy
  - Quotas for different project, users
  - User can be associated with multiple projects

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES



Now, if you look at the capability of open stack; so, primarily if you look at from point of view as infrastructure of service. So, it is a VMs, it can give VMs on demand; so, it is both provisioning and snapshotting is possible. So, it is virtual machine on demand, it has a provision for networking, it has storage for VMs and arbitrary files. So, you can have storage for virtual machines and other files and supports multi tenancy; quotas for different project users, user can be associated with multi projects and type of things.

So, in essence it gives you a full-fledged experience of a cloud and as it is a open source you are installing. So, you have the control both administrative and physical control over the whole thing, so the hardware is used and you are running on the things. So, it is a good thing or it is a very popular thing for individual user or group or lab to install and have a life experience and extremely useful for students with couple of PCs or even a couple of laptops to you can install open stack and see that how things are there.

As we mentioned earlier that there is a good amount of research going on; on resource management in cloud, resource management, power management and people are talking about green cloud and sort of things. So, this sort of lab scale implementation of open source clouds may help in having experience and experimentation of different type of parameters on the cloud. So, open stack is one of the very popular open source cloud.

(Refer Slide Time: 05:58)

## OpenStack History

Series	Status	Initial Release Date	Next Phase	EOL Date
Queens	Future	TBD		TBD
Pike	Under Development	TBD		TBD
Quota	Phase I - Latest release	2017-02-22	Phase II - Maintained release on 2017-08-28	2018-02-26
Newton	Phase II - Maintained release	2016-10-06	Phase III - Legacy release on 2017-10-09	2017-10-11
Mitaka	EOL	2016-04-07		2017-04-10
Liberty	EOL	2015-10-15		2016-11-17
Kilo	EOL	2015-04-30		2016-05-02
Juno	EOL	2014-10-16		2015-12-07
Icehouse	EOL	2014-04-17		2015-07-02
Havana	EOL	2013-10-17		2014-09-30
Grizzly	EOL	2013-04-04		2014-03-29
Folsom	EOL	2012-09-27		2013-11-19
Essex	EOL	2012-04-05		2013-05-06
Dalton	EOL	2011-09-22		2013-05-06
Cactus	Deprecated	2011-04-15		
Beta	Deprecated	2011-02-03		
Austin	Deprecated	2010-10-21		

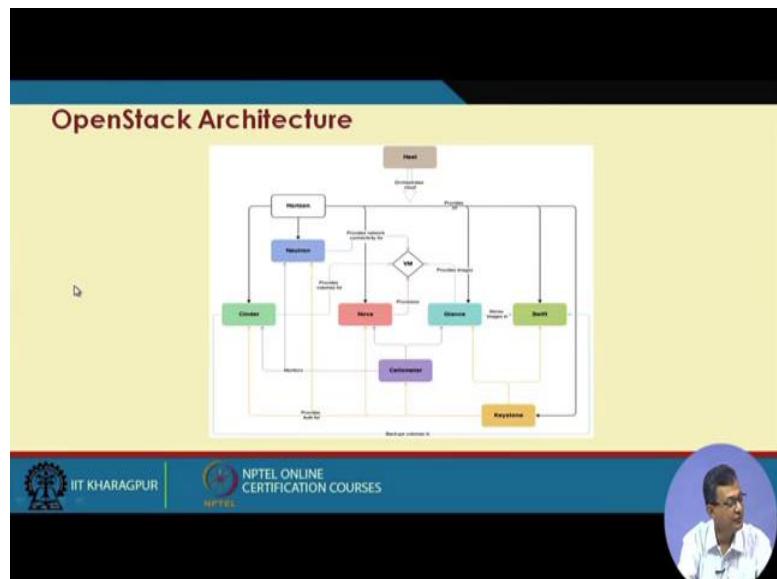
\*Started as a collaboration between NASA and Rackspace

 IIT KHARAGPUR |  NPTEL ONLINE CERTIFICATION COURSES

So, if you look at a history of open stack; it started with collaboration between NASA and Rackspace and over years, it has gone through different project mode and you can see that it started with around 2010 and we are having regular releases with over years.

So, if you see there is started with a project called Austin and go on going to that Newton and Pike and Queens which are to be declared.

(Refer Slide Time: 06:43)



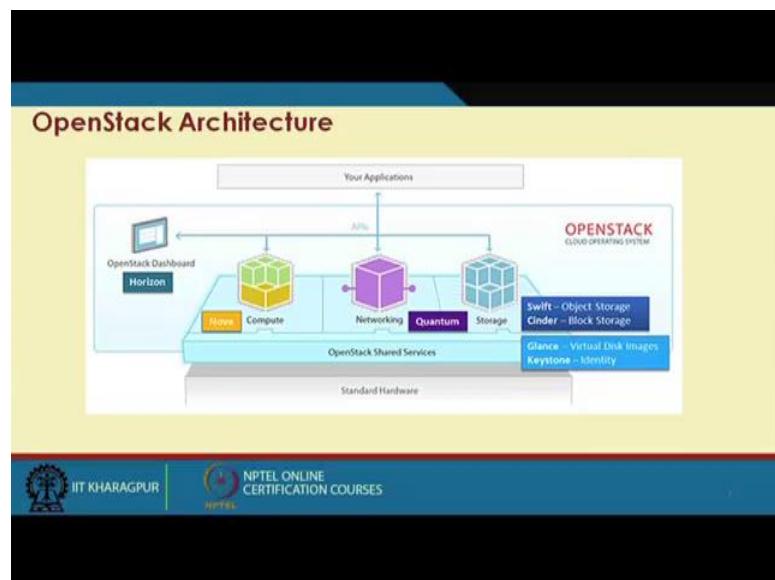
So, if we look at the overall architecture of open stack; there are a couple of components like one is that horizon which is primarily the dashboard; which invite the dashboard

project horizon. So, project newton which is the mostly look at the networking cinder which primary look at the block storage of open stack; nova is the compute, glance is the image services; the different type of different image services which can be hosted in the open stack is there.

Then we have other things like swift; which is object storage like we have cinder as block storage. So, object storage is swift; then ceilometer which is ceilometers, which is the telemetry services; keystone is the idea identity services. So, what we can see that this all different sort of services; which are required in a typical cloud are there in open stack.

So, though they developed in different project mode, but it comes as a bundle and we can have when we have installations, we have all these flavor into the things. So, it is good to have a feel of the cloud and you can have operational cloud using your own internal or in-house cloud using these type of open source. So, things it is less costly you have to pay for the infrastructure and if you have a one infrastructure as I mean already present or access infrastructure already present; which we can deploy it into the using open stack or any other open source cloud.

(Refer Slide Time: 08:44)



So, if you look at from the point of view; then you have the standard hardware at the backbone, then the open stack shear services and over that we have compute networking

storage services and so, this and then their open stack dashboard and the services we talked about.

So, the user applications come from the top and it goes on to this overall inform APIs and utilizes this different services to execute particular or realize particular job. I would like to mention here that we have taken most of these resources from again open stack site and other resources; so, these are the things which you can get in right that also. So, if you look at the major components as we are mentioning here; if you just go back like horizon, newton, cinder, nova glance, swift, ceilometer and keystone; so, these are the major component.

(Refer Slide Time: 09:54)

The slide has a yellow background with a black header and footer. The title 'OpenStack Major Components' is in red. Below it is a bulleted list:

- Service - Compute
- Project - Nova

A blue arrow points to the 'Nova' entry in the list. To the right of the arrow is a detailed description:

Manages the lifecycle of compute instances in an OpenStack environment. Responsibilities include spawning, scheduling and decommissioning of virtual machines on demand.

The footer contains the IIT Kharagpur logo, the NPTEL logo, and a portrait of a man.

So, we will just have a brief overview of these components before we see a demo on those things. So, one of the critical components compute which the service is compute and the project is a nova project; so, manages the life cycle of compute instances in a open stack environment. It manages the life cycle of a compute instances in a open stack environment, responsibilities includes spawning, scheduling, decommissioning of virtual machine on demand.

So, it primarily work with this VM scheduling, VM spawning and decommissioning or releasing the virtual machine on all on demand. So, if you look at it manages the whole compute cycle, so if you see that there are different type of flavor for VMs with different configuration in typical cloud environment; here also we can have different flavor of

VMs. So, I will user request that through the dashboard; the manager or the controller of the cloud can allocate different VMs to the user based on the requirement.

So, it comes with a storage; the ephemeral storage and a persistent storage which can be associated with this VMs; so, these are possible using this compute or nova services.

(Refer Slide Time: 11:44)

The slide has a yellow background with a blue header bar. The title 'OpenStack Major Components' is in red. A section titled 'Neutron' is shown with the following content:

- Service - Networking
- Project - Neutron
- Enables *Network-Connectivity-as-a-Service* for other OpenStack services, such as OpenStack Compute.
- Provides an API for users to define networks and the attachments into them.
- Has a pluggable architecture that supports many popular networking vendors and technologies.

At the bottom, there are logos for IIT Kharagpur and NPTEL, along with the text 'NPTEL ONLINE CERTIFICATION COURSES'.

Another aspect is the networking, another major service is the networking; which is under the neutron thing, enables network connectivity as a service for other open stack service such as open stack compute another things.

So, if you have other open stack services like compute or nova storage services and so on. So, this neutron or the networking provides a networking as a service to this different component of open stack. So, provides an API for users to define network and attachment to them. So, it provides API to the users to define networks and how to attach the things. So, if you look at any cloud infrastructure; so, two types of networks are prominent there; one is the internal network, which is internal to the cloud and there is an external network to which is external to the cloud.

So, like if as you are talking about that we have an open source cloud in our institute that what we call it Meghamala; the experimental open source using open stack with a very small experimental cloud which have been used by faculty and research scholars for their

computing primarily or computing needs, it comes with different flavor; we will show that example and how we do.

So, it has an internal network for the cloud whereas the external network for that cloud is basically the IIT network. So, as it is an in house cloud; so, this is not accessible from the external world; however, the cloud itself has an internal thing which is for communicating between this different component and providing services and an external link which gives a connectivity to the external one.

So, this is based on this neutron type of services; if you are using open stack. So, it has a pluggable architecture that supports many popular networking vendors and technology. So, it is interoperable between different networking vendors and technologies so that is feasible in open stack.

(Refer Slide Time: 14:04)

The slide has a dark blue header bar. Below it, a yellow section contains the title 'OpenStack Major Components' in red. Underneath the title, there is a bulleted list describing the Swift component:

- Service - Object storage
- Project - Swift

Below this, there is another bulleted list:

- Stores and retrieves arbitrary unstructured data objects via a RESTful, HTTP based API.
- It is highly fault tolerant with its data replication and scale-out architecture. Its implementation is not like a file server with mountable directories.
- In this case, it writes objects and files to multiple drives, ensuring the data is replicated across a server cluster.

At the bottom of the slide, there are logos for IIT Kharagpur and NPTEL, along with the text 'NPTEL ONLINE CERTIFICATION COURSES'.

The next one is the storage service; which comes under project swift, store and retrieves arbitrary structure unstructured data objects as a RESTful, HTTP based API.

So, it stores and retrieves unstructured data objects as a RESTful; I believe that you know you understand about RESTful services; RESTful and HTTP based API. So, it is highly fault tolerant with data replication and scale out architecture; its implementation is not like file server with mountable directories etcetera. So, it is a basically a fault tolerant and system with a scale out architecture, it is not they simple file server. So, it is much

more than that. And in this case, it writes objects and files to multiple devices ensuring the data is replicated across the server clusters.

So, in order to make fault tolerant it writes into the things; as if you remember when we are talking about data on cloud or data services on the cloud, in our lecture or when we discussed about those. So, the replication is the scenario if discuss with the 3 replications.

So, it replicate the data into 3 places, so whenever there is a right operations; so all these right will be having should sink with all the replicas. Whereas, in case of read over sense in any of the replica can response to the things; in never the less this storage service of swift also provide such structures to means fault tolerant and services.

(Refer Slide Time: 15:50)

**OpenStack Major Components**

- Service - Block storage
- Project - Cinder

- Provides persistent block storage to running instances.
- Its pluggable driver architecture facilitates the creation and management of block storage devices.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, another type of storage service is your block storage, which is provided by cinder provide persistent block storage to running instances. So, it is a persistent block storage to the running instances; it is pluggable driver architecture facilitate the creation and management our block storage. So, it is a; again it is a persistent storage and persistent block storage and provides a pluggable driver architecture to facilitate creation and management of block storage things.

(Refer Slide Time: 16:25)

## OpenStack Major Components

- Service - Identity
- Project - Keystone

- Provides an authentication and authorization service for other OpenStack services.
- Provides a catalog of endpoints for all OpenStack services.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Then another component we are having which is not directly under compute or storage, but plays a vital role in realizing the cloud is the identity service or what you say keystone under the keystone project.

Provides authentication and authorization service for other open stack services. So, it is a identity service and provides authorization and authentication of services for other things. Provides a catalogue of end points for all open stack services; so, also it along with that it provides a catalogue of end points of open stack services that how those services are defined and type of things.

(Refer Slide Time: 17:10)

**OpenStack Major Components**

- Service - Image service
- Project - Glance

- Stores and retrieves virtual machine disk images.
- OpenStack Compute makes use of this during instance provisioning.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, the next component is glance or what we say that image services. So, it basically open stack support separate type of images or image services where which can be loaded or which can be instantiated into that different VMs. So, I can have different VM and different images of say operating systems and other things can be instantiated on the difference VMs based on the user needs and requirement.

So, these sorts of services are provided by that; it is glance project or the glance service glancing service. So, it stores retrieve virtual machine disk images, so the virtual VM disk images are stores and retrieves by this glance. And open stack compute makes use of during the instance provisioning; as we are mentioning, whenever the open stack have instance provisioning; then they use this storage services.

Like you can have this image services; so, like you can have different images say for operating systems. So, you can have different flavor or images or other operating systems and when you instance; based on the requirement of the user, those are instantiated by this compute services; so, this image repositories are there in the open stack.

(Refer Slide Time: 18:54)

**OpenStack Major Components**

- Service - Telemetry
- Project - Ceilometer

- Monitors and meters the OpenStack cloud for billing, benchmarking, scalability, and statistical purposes.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

This is another service of telemetry; which is ceilometers monitors and meters the open stack cloud for billing, benchmarkings, scalability, statistical purposes.

So, this is important for overall metering of the cloud services as we have discussed in our initial lecture, as we understand that cloud is a meter service; that means, whatever uses and type of things are meter. So, that means this particular ceilometer or the telemetry services is in cloud, in open stack helps us in monitoring and metering the open stack cloud for billing, benchmarking, scalability, statistical purposes; it is not only for meeting, it is required that as you are having this measured services.

So, you can do benchmarking the address; the scalability resource and statistical analysis; like you want to know that what is the loading, how it be a, type of things. So, those type of things are available; so it is a: though it may not be directly contributing to the compute or storage or networking; which use anything that is the major component which allows us to work on those, allows us in this aspect.

But never the less it plays a vital role in having, in helping in realizing these metered service of the things along with statistical analysis or statistical measurement of things to charge the performance of the things, which not only help in building of the resource uses, also it helps in understanding the future requirement and how the infrastructure at the backbone need to be increased and type of things, it help in realizing those requirements also.

And maybe the train, maybe the periodical requirement; I can say that every one day this is the load; weekends load is less; however, some time; it is some particular time etcetera. So, this overall for analyzing over all the overall this performance of the things; what we required is more about different sort of data, which this telemetry service provides us.

(Refer Slide Time: 21:37)

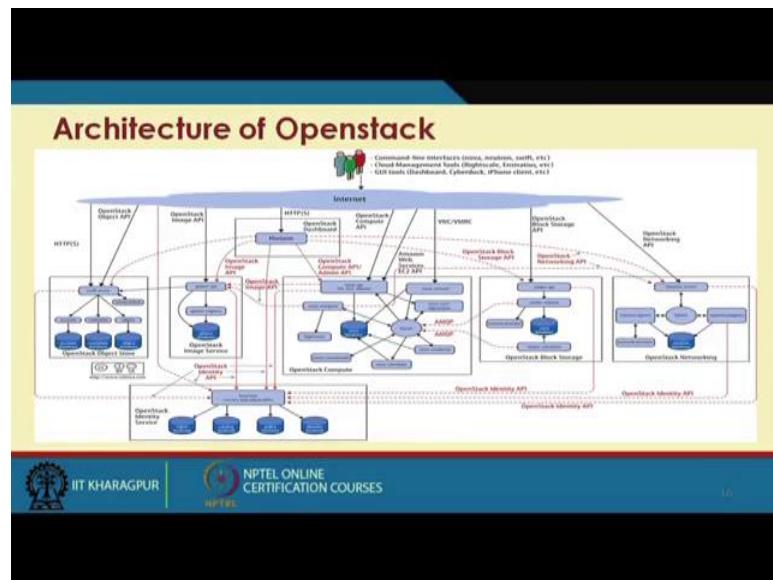
The slide has a dark blue header and footer. The main content area is yellow. The title 'OpenStack Major Components' is in red. Below it, there are two sections: 'Service - Dashboard' and 'Project - Horizon'. The 'Service - Dashboard' section contains a bulleted list: 'Provides a web-based self-service portal to interact with underlying OpenStack services, such as launching an instance, assigning IP addresses and configuring access controls.' At the bottom, there are logos for IIT Kharagpur and NPTEL, along with the text 'NPTEL ONLINE CERTIFICATION COURSES'.

Then we have the component of open stack dashboard; that is what we say under project horizon. So, its provides a web base self service portal to interact with underlining open stack services; such as launching an instance, assigning IP addresses and configuring access control.

So, it is a dashboard under project horizon provides a well web base self service portal; to interact with underline in open stack services such as launching, instance launching, assigning IP address and configuring; actually we will show in our demo that how we are using this open stack dashboard to user management, to resource management. Like we want to assign a VM or assign IP address or if an means configured the access controls, loading images etcetera all can be done using this dashboard.

So, it is also important aspects and it is the; what we can say frontend interface for the administrator to manage the cloud. So, it is extensively used for management of the cloud.

(Refer Slide Time: 23:09)

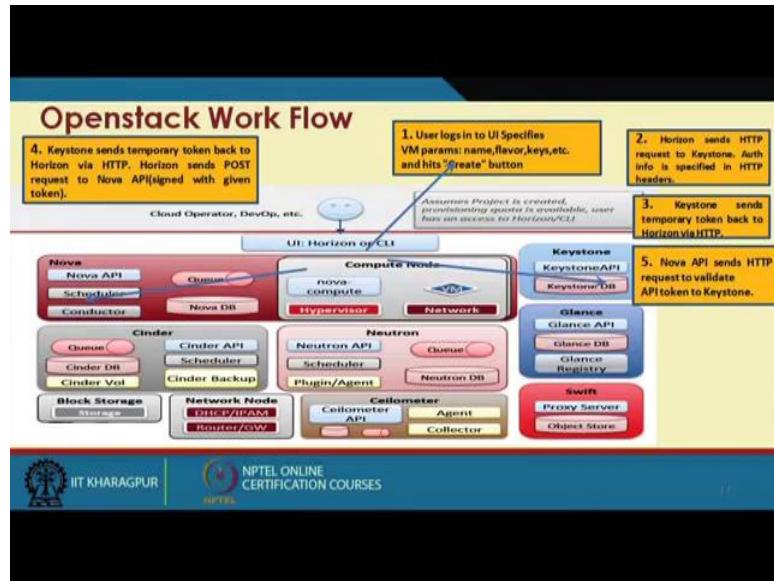


So, if we look at the overall architecture then what we see that we have different components. So, this is open stack object storage as we have discussed that is which is realized to swift; that open stack image services which is used to glance, open stack compute services which is realized to nova; realized by nova, there are block storage which is cinder; open stack networking; which is the neutron.

And open stack identity services which is keystone. So, there are different type of services and if we have seen that horizon which allows us to look at the dashboard. So, this that open stack dashboard and this is the internet that user comes on the command line interfaces; nova neutron swift etcetera, cloud management tools like right scale; other tools GUI tool dashboard, cyber duck and so on and so forth.

So, there is the user interface for the things and then it basically; based on its requirement it words into this. So, it shows that over how this overall at the top level; open stack modules are interconnected, how the process flows goes on into this open stack module. So, this is the; a overview of the thing; there are few more site of the individual component, we will not go to the much nitty-gritty of the thing, but nitty-gritty of the material, but try to see that what are the different type components.

(Refer Slide Time: 24:55)

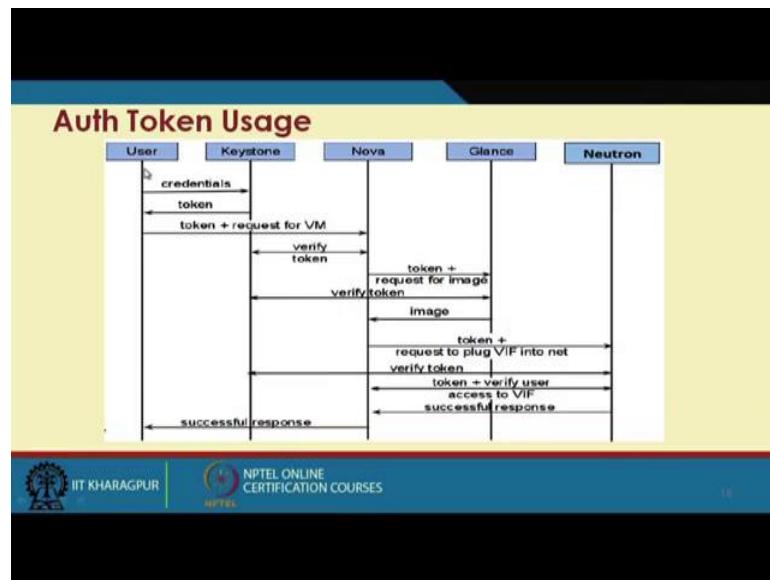


Like, if you look at the open stack workflow; so, these are the different components are there nova for compute, cinder for block storage; these are the compute nodes in the nova. There are other components within the nova, there is a cinder also have its own component. Neutron is the networking part; networking aspects; so, neutron API scheduler plug in, queues etcetera these are all networking related component.

We have ceilometers; which is primarily that the metering service or the telemetric component and there are keystone for the identity services. So, user log scene to UI specific VM parameters like name, flavor, keys etcetera and hits the create button. So, they do it on the means when user log scene with the particular creative VM. Then the horizon sends the HTTP request to keystone; authenticate information as specified in the HTTP header.

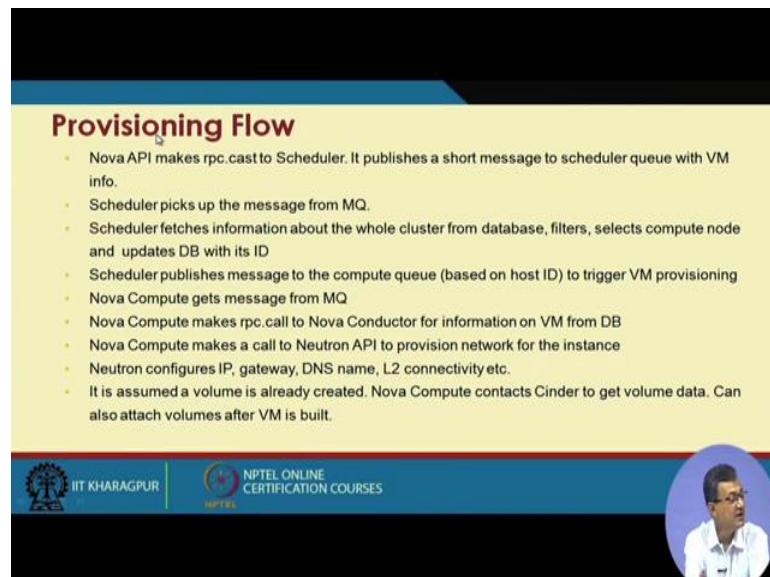
So, horizon or the dashboard service sends the information to this; I will re service then we are at the keystone sends the temporary token back to horizon via HTTP and nova API sends HTTP request to; now after that keystone sends the HTTP, then the keystone says the temporary token back to the horizon, via HTTP; horizon sends the post request to the nova API and send them. So, once that authentication is there then the horizon sends the things to the compute server; which in turned as it other operations. Now, finally the nova API sends HTTP equates to validate API token to case 2; so, this way the whole thing goes on.

(Refer Slide Time: 26:53)



And if we look at the token exchange mechanism; so, the user to the keystone it is; for the identification or authentication, then it goes to the nova, then the glance and the neutron. So, this is what we say; if you look at that process flow of that particular authentication token uses.

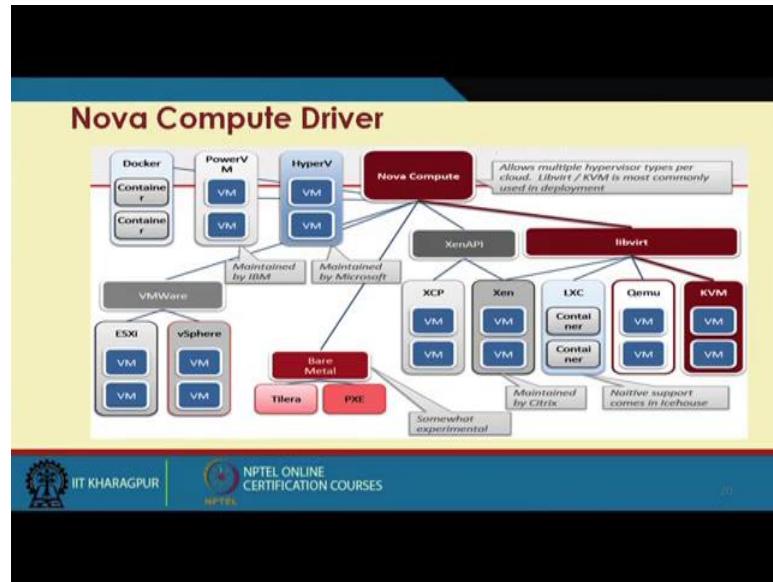
(Refer Slide Time: 27:16)



Similarly, there is a provisioning flow nova appearing makes a rpc cast to scheduler; scheduler picks up the message from the message queue, scheduler fetches information about the whole plaster from the databases, data base filters; selects compute nodes

scheduler, publishes message to the compute queues. Nova compute gets message from the nova message queue MQ and nova compute makes rpc called to nova conductor and so on and so forth. So, that is how that provisioning is made into the open stack.

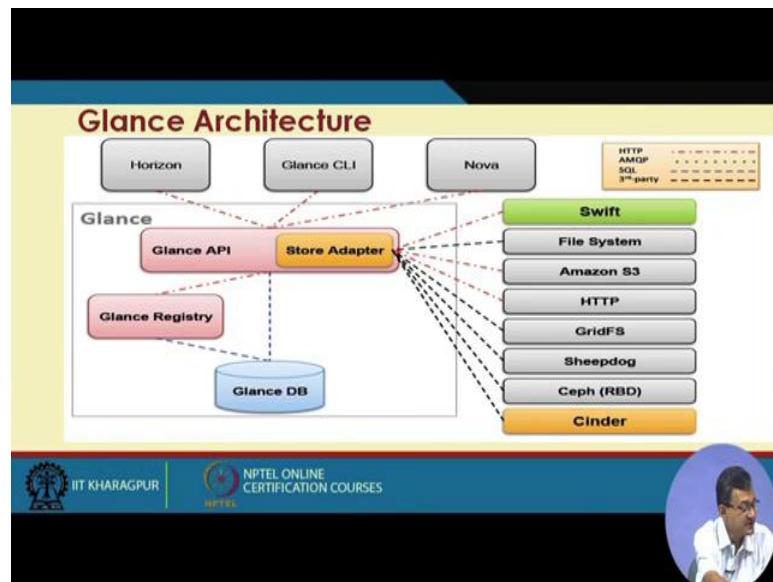
(Refer Slide Time: 27:54)



And then we have individual component like compute driver; which are expansion of the things. We are not going detail into the open stack thing; those who are interested in particular things can refer to their resources.

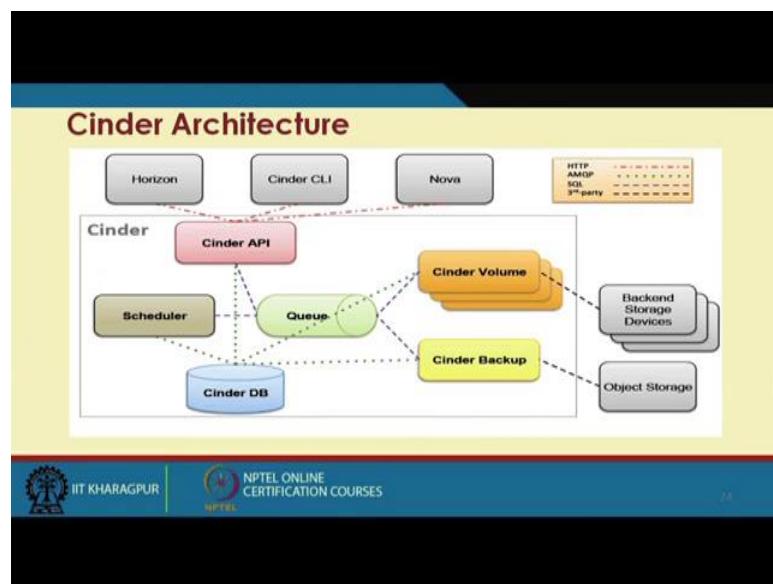
Similarly, we have neutron architecture which is the networking architecture and as such neutron also have several component into; a under its folder.

(Refer Slide Time: 28:27)



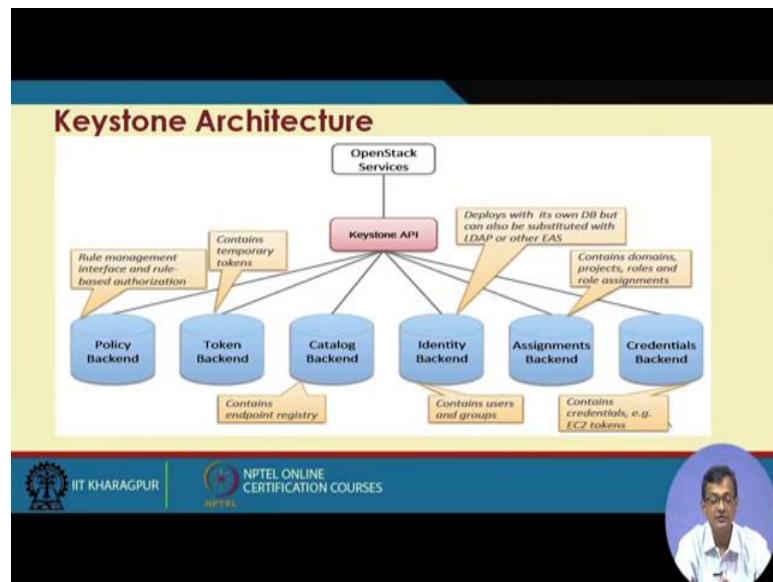
The glance which is for image service; so, it has glance databases, can registry that what sort of services is there; then glance API and storage adopter. So, we have again few components under the glance.

(Refer Slide Time: 28:52)



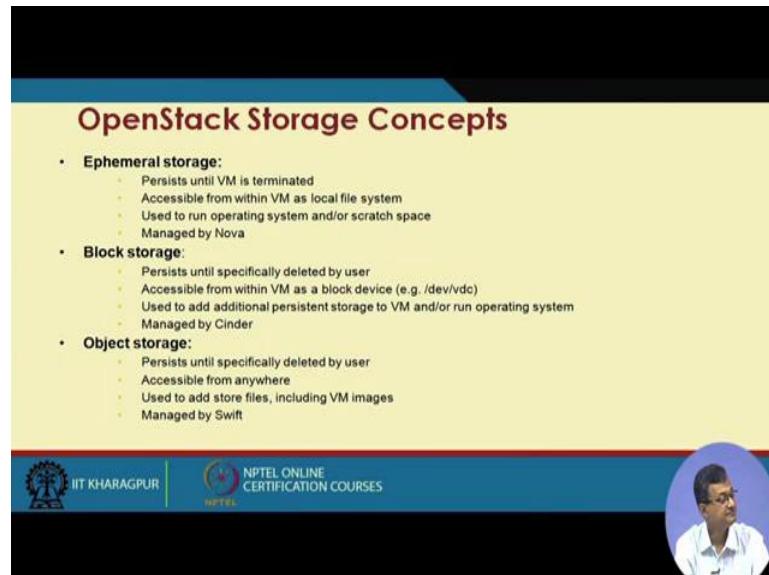
Similarly, cinder architecture which is cinder architecture which is called block storage; so these are the major component of the cinder; that is cinder databases, scheduler, API, volumes and a backup service of this cinder.

(Refer Slide Time: 29:10)



Keystone as we have already discussed is the identity type of service and it has different modules or different components. Policy back in token, tokenization, catalogue service, identity service, assignment backend and credentials backend, so these are the different type of component or services under this keystone identity service.

(Refer Slide Time: 29:38)



So, if you look at the open stack storage concepts; it is in the similar line with the other cloud storages; like ephemeral storage; persists until the VM terminated, accessible from within VM as local file systems; this is the ephemeral storage. So, once the VM is

terminated; it also goes off. Used to run operating system and or scratch space like as it is ephemeral, so it is primarily used to run operating system which are loaded as and when things are required. And it is also can be user a scratch space and managed by the nova; that we have seen block storage persist until specifically deleted by user.

So, it is a block storage, so it persists until specifically deleted by the user accessible within VM as a block service. Used to add additional persistent storage to VM and to operating systems, so it a used to add additional storage to the things; otherwise you are with the VM taking as storage only and it is managed by cinder.

Then we have a object storage, which is managed by swift persists until specifically deleted; accessible from anywhere, used to add and store files including VM images. So, this VM images which are managed by some of the images managed by the glance services are also used to add store files into the object storage and this and the object storage is managed by as we have discussed earlier is by swift.

(Refer Slide Time: 31:34)

The slide has a yellow background with a black header and footer. The title 'Summary' is in red. A bulleted list details the VM creation process:

- Users log into Horizon and initiates VM creation
- Keystone authorizes
- Nova initiates provisioning and saves state to DB
- Nova Scheduler finds appropriate host
- Neutron configures networking
- Cinder provides block device
- Image URI is looked up through Glance
- Image is retrieved via Swift
- VM is rendered by Hypervisor

At the bottom, there are logos for IIT Kharagpur and NPTEL, with the text 'NPTEL ONLINE CERTIFICATION COURSES'.

So, in summary if you have a quick overview of the whole this open stack, the user logs in horizon and initiates the VM creation. Keystone authorizes it, nova initiates provisioning and saves state to the database. Nova scheduler finds the appropriate host. Neutron configures the networking aspects, cinder provides block devices. Image URI is looked up through glance.

Image is retrieved via swift and VM is rendered by a hypervisor. So, this is in a overall that a brief overview of these open stack; open source cloud. So, what we will do next is a demo on these things; I will try to give a live demo of our open stack installation at IIT Kharagpur, as I mentioned and then we will see that how VM can be created and all those things.

Thank you.

**Cloud Computing**  
**Prof. Soumya Kanti Ghosh**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**

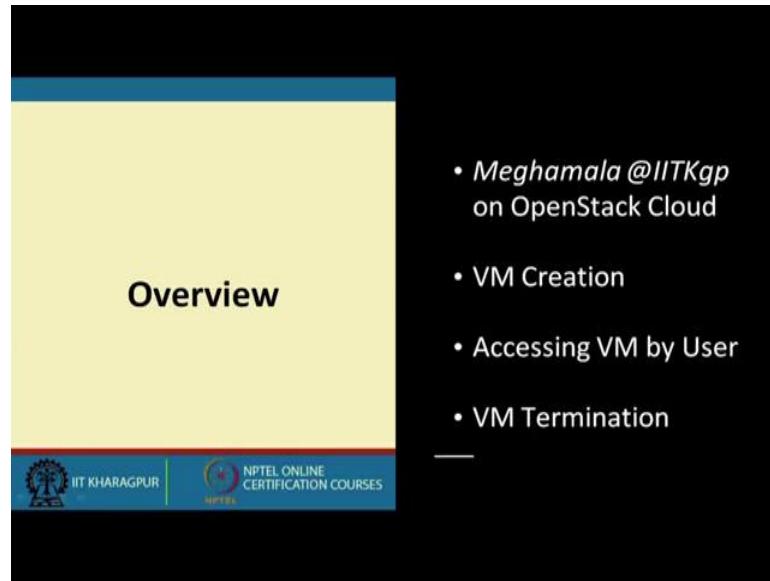
**Lecture – 16**  
**Open Source Cloud: Openstack Demo**

Hi, so we will continue our discussion on open stack cloud a open source cloud. And we will so, a sort demo how the open stack work. The primary objective is a open source cloud you can easily download those this open stack in your local systems we if you have couple of systems. And realize this cloud and see it is different aspect. As we have discussed the different type of services like compute storage image and other services that we will see that how these are realized, right.

So, as I mentioned that in IIT Kharagpur we have install a experimental cloud using open source platform open stack. So, a so a demo on that which is in our cloud we called it Meghamala. So, we will so, a demo on that it is primarily a open stack based cloud. So, with me Rajesh is there. So, Rajesh is primarily a administrator of Meghamala. So, he will so the how AVMs is created allocated how to run a particular job in that VM, how to dallocate and type of things as simple or the some of the operations on Meghamala.

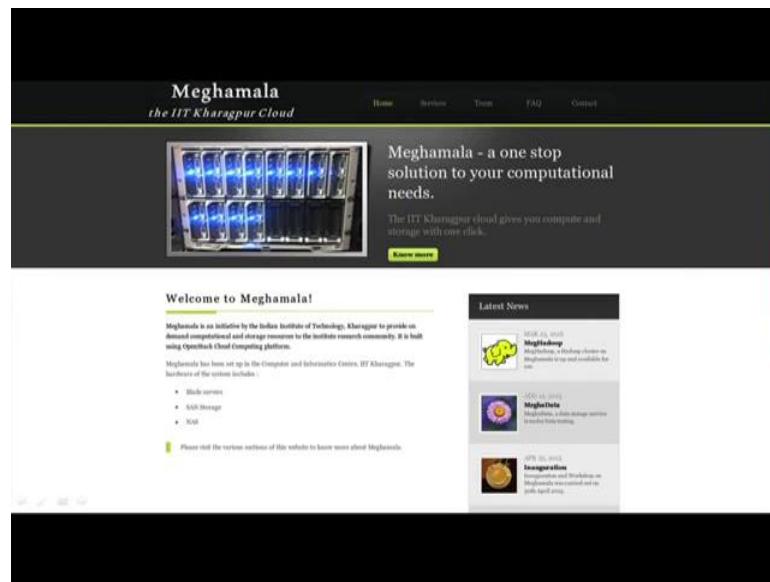
So, before we I hand it over to Rajesh for the demonstration on live demonstration on open stack that is Meghamala, I will just go through couple of slide to just give you a overview.

(Refer Slide Time: 01:55)



So, it is say open stack base cloud which we called that Meghamala. So, what we have VM creation what will.

(Refer Slide Time: 02:12)



So, the VM creation accessing VM by user and VM Termination. Meghamala gives different type of flavors of VM that to what Rajesh, we will show and this is a typical our Meghamala portal which has different aspects.

(Refer Slide Time: 02:22)

The screenshot shows the homepage of the Meghamala website. At the top, there is a dark header bar with the text "Meghamala" and "the IIT Kharagpur Cloud". Below the header, there are navigation links for "Home", "Services", "Team", "FAQ", and "Contact". The main content area has a white background. On the left, there is a section titled "Services offered by Meghamala" which lists various service offerings. On the right, there is a "Latest News" sidebar with four news items, each with a small thumbnail image, a date, and a brief description.

- MEGHADEEP: Meghamala, a hadoop cluster in Meghamala is up and available for use.
- Meghamala: Meghamala, a virtual machine service to easier file sharing.
- APRIL 25, 2012: Inauguration: Meghamala Inauguration Workshop on Meghamala installed version 20th April 2012.
- JULY 11, 2012: Installation Complete: Meghamala and Meghadata installed. Testing in progress.

A overview of the thing and these are the different type of flavors which Meghamala gives like IIT KGP regular with 2 vCPU, 4 GB RAM for different 45 GB ephemeral storage. If you remember that a few ephemeral storage and also we have a provision for persistent storage. Typically we give 20 GB persistent storage 20 dB or 60 dB different on the requirement IIT KGP large and IIT KGP extra large

So, these are the 3 flavors and the these are 3 operating systems which are they are in Meghamala. Along with that we give we are started giving some other services like meghadata of data services Meghadoop which is running over Meghamala, but primarily we will be hovering around these hat VM creation And so on and so forth.

(Refer Slide Time: 03:13)

The screenshot shows a web-based request form titled "VMs4U - Request form". The form fields include:

- Name of faculty
- Department
- Designation
- Phone/Mobile no.
- E-mail
- Program
- Preferred VM Name

Below these fields are radio buttons for "VM Type":  IITKgp\_update,  IITKgp\_Req,  IITKgp\_Sign.

Fields for "Number of VMs" (set to 1) and "Operating system" (set to Ubuntu 14.04) are present. A checkbox for "Persistent storage of virtual required" is checked.

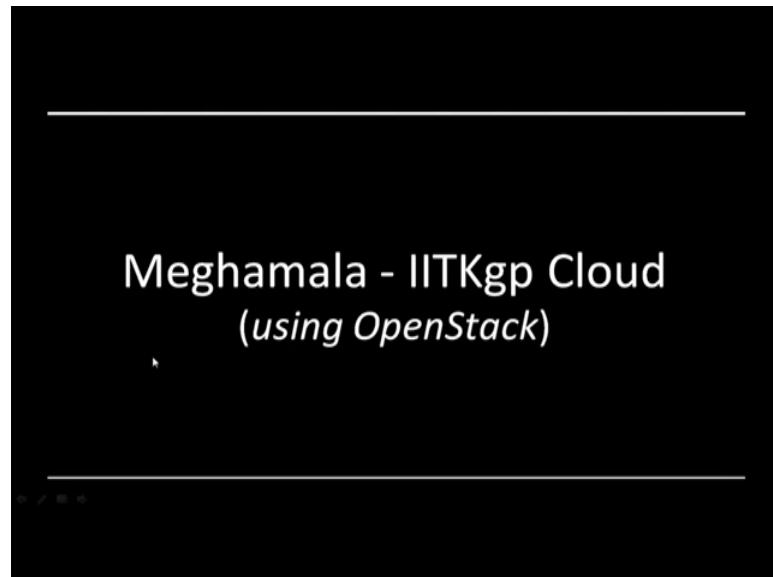
A "VM request ID" field contains the value "knyqbg".

A note at the bottom states: "Please note that the VMs should be used only for academic purposes. Avoid the Highend VMs as it'll hamper the performance for the contents of your VMs. It is important to highlight that the process of computation-needed may lead to termination/termination of the VMs".

At the bottom right is a "Submit Query" button.

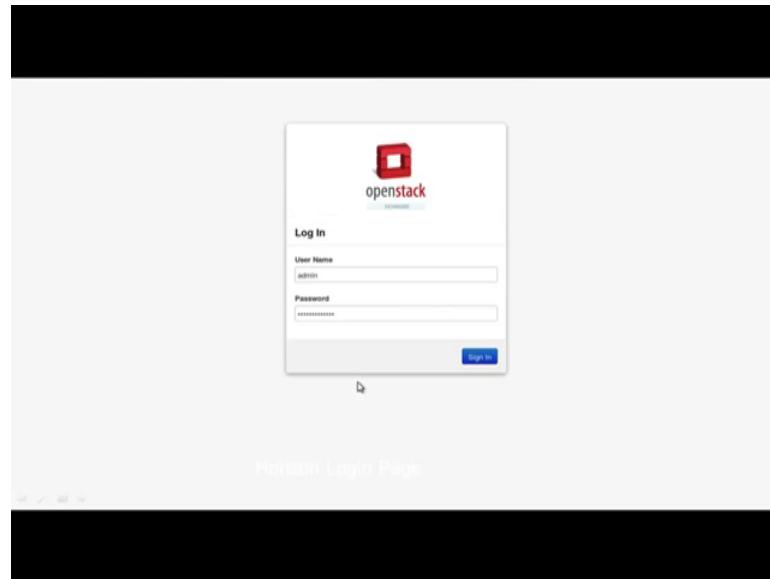
And this whole thing is based on open stack, and this is a typical request form by which a user can request for AVM.

(Refer Slide Time: 03:26)



And these are the different people who are involved in this Meghamala. So, so, if we look at.

(Refer Slide Time: 03:30)



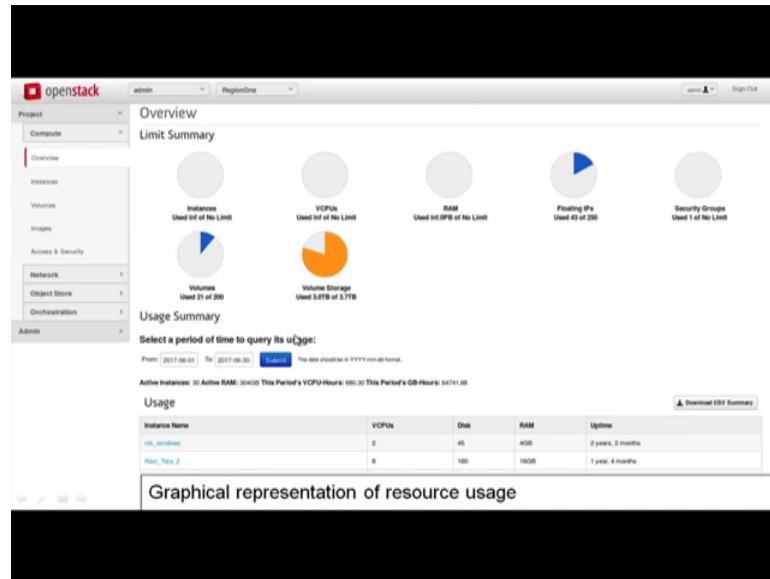
It is this is the log in screen or the open stack by which you can enter the open stack dashboard for management.

(Refer Slide Time: 03:40)

A screenshot of the 'Overview' page in the OpenStack Horizon dashboard. The top navigation bar shows 'admin' and 'RegionOne'. On the left, a sidebar menu includes 'Project' (selected), 'Compute', 'Overview', 'Instances', 'Volumes', 'Images', 'Access &amp; Security', 'Network', 'Object Store', 'Orchestration', and 'Admin'. The main content area is titled 'Usage Summary' and contains a sub-section 'Select a period of time to query its usage:' with date inputs 'From: 2017-06-01' and 'To: 2017-06-30', a 'Submit' button, and a note 'The date should be in YYYY-mm-dd format'. Below this is a summary table: 'Active Instances: 30 Active RAM: 304GB This Period's VCPU-Hours: 679.47 This Period's GB-Hours: 64662.52'. A 'Download CSV Summary' button is located at the top right of this table. The table has columns: Project Name, VCPUs, Disk, RAM, VCPU Hours, and Disk GB Hours. One row is shown: admin, 128, 2855, 30408, 679.47, 64662.52. The bottom of the page shows a progress bar with the text 'Displaying 1 item'.

And these are different aspects of the things like giving a overall summarization of the users summary.

(Refer Slide Time: 03:46)



Giving a overall representation of the resource uses in terms of different graph by graphs.

(Refer Slide Time: 03:54)

The screenshot shows the OpenStack Compute dashboard under the 'Compute' project. The main header includes 'RegionOne', 'admin', and navigation links for 'Compute', 'Project', 'Instances', 'Volumes', 'Images', 'Access & Security', 'Network', 'Object Store', 'Orchestration', and 'Admins'. The left sidebar has sections for 'Overview', 'Instances', 'Volumes', 'Images', 'Access & Security', 'Network', 'Object Store', 'Orchestration', and 'Admins'. The 'Instances' section lists several instances with their details: 'rhv\_virtinst' (CentOS\_7\_0\_0, IP 192.168.0.2, 10.4 G, 408 RAM, 2 VCPUs, 40.008 Disk, Active, nova, None, Running, 2 months, 2 weeks, Create Snapshot, More...); 'TestDiskPartition' (Ubuntu\_14\_04\_xlarge\_4G, IP 192.168.0.2, 10.4 G, 408 RAM, 2 VCPUs, 45.008 Disk, Active, nova, None, Running, 3 months, 2 weeks, Create Snapshot, More...); 'centosParity' (CentOS\_8.5\_0\_0, IP 192.168.0.3, 10.4 G, 408 RAM, 2 VCPUs, 40.008 Disk, Standby, nova, None, Shutdown, 7 months, Start Instance, More...); 'GL1\_R\_SERVER1' (Ubuntu\_New\_X20e, IP 192.168.0.4, 10.4 G, 3000 RAM, 4 VCPUs, 60.008 Disk, Active, nova, None, Running, 9 months, 1 week, Create Snapshot, More...); 'Hannu\_Uuvan\_LARGE' (Ubuntu\_14\_04\_xlarge\_800, IP 192.168.0.5, 10.4 G, 800 RAM, 8 VCPUs, 80.008 Disk, Active, nova, None, Running, 1 year, 2 months, Create Snapshot, More...); 'rhv\_virtinst\_2' (Ubuntu\_14\_04\_xlarge\_4G, IP 192.168.0.6, 10.4 G, 408 RAM, 2 VCPUs, 40.008 Disk, Standby, nova, None, Shutdown, 1 year, 2 months, Start Instance, More...); 'MegashapePowerMaster' (CentOS\_8.5\_0\_0, IP 192.168.0.7, 10.4 G, 800 RAM, 8 VCPUs, 80.008 Disk, Active, nova, None, Running, 1 year, 4 months, Create Snapshot, More...); and 'Megashape\_10' (CentOS\_8.5\_0\_0, IP 192.168.0.8, 10.4 G, 800 RAM, 8 VCPUs, 80.008 Disk, Active, nova, None, Running, 1 year, 5 months, Create Snapshot, More...). A callout box highlights the 'Instances' section.

And what are the different instance running at any point of time, volumes and snapshots of things which are maintained by cinder as we as we are discuss some time back.

(Refer Slide Time: 04:07)

The screenshot shows the OpenStack Compute interface under the 'Compute' project. The 'Images' tab is selected. A search bar labeled 'Glance' is highlighted with a red box. The table below lists the following images:

Image Name	Type	Status	Public	Protected	Format	Actions
Magnitude_ubuntu_ready	Snapshot	Active	Yes	No	QCOW2	[Launch] [More]
CentOS_8.1_OL	Image	Active	Yes	No	QCOW2	[Launch] [More]
Stackpole_1_10_4_2_30_000001	Snapshot	Active	No	No	QCOW2	[Launch] [More]
stackpole_working	Snapshot	Active	No	No	QCOW2	[Launch] [More]
Ubuntu_14_04_xlate_600	Image	Active	Yes	No	QCOW2	[Launch] [More]
Ubuntu_14_04_xlate_600	Image	Active	Yes	No	QCOW2	[Launch] [More]
Ubuntu_14_04_xlate_200	Image	Active	Yes	No	QCOW2	[Launch] [More]
Ubuntu_New_X200	Image	Active	Yes	No	QCOW2	[Launch] [More]
Windows_7_x64	Image	Active	Yes	No	QCOW2	[Launch] [More]
Fedora_28_OL	Image	Active	Yes	No	QCOW2	[Launch] [More]
Centos_7_OL	Image	Active	Yes	No	QCOW2	[Launch] [More]

This is the different images which is managed by the glance service, neutron is the networking aspects of the things.

(Refer Slide Time: 04:12)

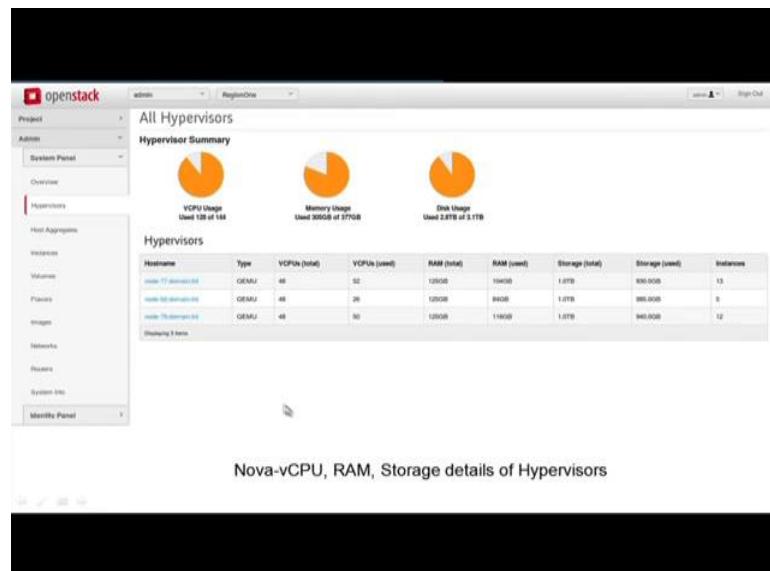
The screenshot shows the OpenStack Compute interface under the 'Compute' project. The 'Access & Security' tab is selected. The 'Manage Security Group Rules' page for the 'default' security group is displayed. The table shows the following rules:

Direction	Ether Type	IP Protocol	Port Range	Remote	Actions
Egress	IPv4	Any	-	0.0.0.0/0 (CIDR)	[Delete Rule]
Ingress	IPv4	Any	-	default	[Delete Rule]
Ingress	IPv6	Any	-	default	[Delete Rule]
Egress	IPv6	Any	-	-/0 (CIDR)	[Delete Rule]
Ingress	IPv4	ICMP	-	0.0.0.0/0 (CIDR)	[Delete Rule]
Ingress	IPv4	TCP	1-65535	0.0.0.0/0 (CIDR)	[Delete Rule]
Ingress	IPv4	TCP	3009 (TCP)	0.0.0.0/0 (CIDR)	[Delete Rule]
Ingress	IPv4	TCP	27017	0.0.0.0/0 (CIDR)	[Delete Rule]

Neutron- Network Access Rules of a Security Group

We are using all IPv4 structure.

(Refer Slide Time: 04:22)



And the hypervisors, nova vCPU is RAM storage details other hypervisors. Different flavors of compute server.

(Refer Slide Time: 04:29)

The screenshot shows the Nova Flavors dashboard under the 'admin' user. A table lists various VM flavor configurations:

Flavor Name	vCPUs	RAM	Root Disk	Ephemeral Disk	Swap Disk	ID	Public	Actions
ext4tiny	1	512MB	1GB	0GB	0MB	1	Yes	<a href="#">Edit Flavor</a> <a href="#">Delete</a>
ext4small	1	2048MB	2GB	0GB	0MB	2	Yes	<a href="#">Edit Flavor</a> <a href="#">Delete</a>
ext4medium	2	4096MB	4GB	0GB	0MB	3	Yes	<a href="#">Edit Flavor</a> <a href="#">Delete</a>
ETHDP_regular	2	4096MB	4GB	0GB	0MB	4	Aborted	<a href="#">Edit Flavor</a> <a href="#">Delete</a>
RamOverCommitTest	2	16384MB	20GB	0GB	0MB	5	209648x2-0f8e432a (Item: 61a8014747a)ice	<a href="#">Edit Flavor</a> <a href="#">Delete</a>
ETHDP_larger	4	8192MB	4GB	0GB	0MB	6	af209a50-4651-4482-8468-1a6d443bd51	<a href="#">Edit Flavor</a> <a href="#">Delete</a>
ext4large	4	8192MB	8GB	0GB	0MB	7	Yes	<a href="#">Edit Flavor</a> <a href="#">Delete</a>
Meghamala	4	8192MB	9GB	0GB	1024MB	8cc397a0-787b-4139-851a-e70a8fe02d4	Yes	<a href="#">Edit Flavor</a> <a href="#">Delete</a>
Meghamala_new	4	8192MB	9GB	0GB	1024MB	dc1a5a0b-0f48-425a-9394-7173006c9512	Yes	<a href="#">Edit Flavor</a> <a href="#">Delete</a>
ETHDP_charge	8	16384MB	8GB	0GB	0MB	9	36031a4f-12ab-4fb6-9d43-2215a7553d4	<a href="#">Edit Flavor</a> <a href="#">Delete</a>
ext4large	8	16384MB	16GB	0GB	0MB	10	Yes	<a href="#">Edit Flavor</a> <a href="#">Delete</a>

A note at the bottom states: "Nova- Different flavors of VMs in Meghamala".

That is a nova compute servers, like you can see that different category of nova.

(Refer Slide Time: 04:39)

Image Name	Type	Status	Public	Protected	Format	Actions
Meghamala_ancestor_metal	Snapshot	Active	Yes	No	QCOW2	<a href="#">Edit</a> <a href="#">More</a>
CentOS_8.1_01	Image	Active	Yes	No	QCOW2	<a href="#">Edit</a> <a href="#">More</a>
Stackware_1_30_A_2_01_100001	Snapshot	Active	No	No	QCOW2	<a href="#">Edit</a> <a href="#">More</a>
Stackware_weling	Snapshot	Active	No	No	QCOW2	<a href="#">Edit</a> <a href="#">More</a>
Ubuntu_14_04_x86_64	Image	Active	Yes	No	QCOW2	<a href="#">Edit</a> <a href="#">More</a>
Ubuntu_14_04_x86_64	Image	Active	Yes	No	QCOW2	<a href="#">Edit</a> <a href="#">More</a>
Ubuntu_14_04_x86_64	Image	Active	Yes	No	QCOW2	<a href="#">Edit</a> <a href="#">More</a>
Ubuntu_New_X200e	Image	Active	Yes	No	QCOW2	<a href="#">Edit</a> <a href="#">More</a>
Windows_7_64	Image	Active	Yes	No	QCOW2	<a href="#">Edit</a> <a href="#">More</a>
Fedora_30_01a	Image	Active	Yes	No	QCOW2	<a href="#">Edit</a> <a href="#">More</a>
Centos_7_64	Image	Active	Yes	No	QCOW2	<a href="#">Edit</a> <a href="#">More</a>
Ubuntu	Image	Active	Yes	No	QCOW2	<a href="#">Edit</a> <a href="#">More</a>

Compute server imaging, instances and overall compute services in Meghamala.

(Refer Slide Time: 04:41)

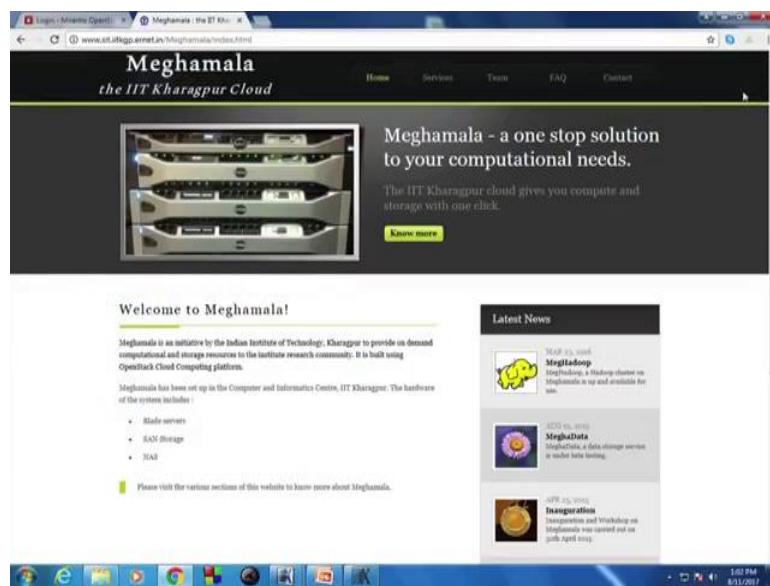
Name	Host	Zone	Status	State	Updated At
nova-conductor	node-01.domain.lid	internal	enabled	up	0 minutes
nova-scheduler	node-01.domain.lid	internal	enabled	up	0 minutes
nova-api	node-01.domain.lid	internal	enabled	up	0 minutes
nova-compute	node-77.domain.lid	nova	enabled	up	0 minutes
nova-compute	node-02.domain.lid	nova	enabled	up	0 minutes
nova-compute	node-79.domain.lid	nova	enabled	up	0 minutes
nova-compute	node-01.domain.lid	internal	enabled	up	0 minutes

So, with this what I will do I will switch over the control over to Rajesh to So write like directly a demo on Meghamala, which will you a idea that if you install your a open stack on your system. So, how it is likely to behave. So now, it is over to Rajesh. So, we will now start the demo on this open stack cloud what we have install which is installed in our institute that is Meghamala. So, it is basically open stack cloud and Rajesh is with me to show the demo. So, Rajesh will be showing making a walk through these

Meghamala the open stack cloud. So, primarily looking at the more on the VM creation, termination on other type of aspects. So, it is over to Rajesh to he will start with that Meghamala web portal to go to that dash dashboard a open stack and going inside the VM case and etcetera.

So, over to Rajesh, rajesh thank you sir happen.

(Refer Slide Time: 06:01)



So, I will continue from here. So, this is the homepage of our institute cloud which is Meghamala. So, as you can see the services we offer is not only the infrastructure cloud which is provided by Meghamala, but also some other services which is built on top of Meghamala. Like on top of open stack like Meghadoop which is a Hadoop cluster, Meghadata which is a personal cloud. This is a kind of a drop box like thing. So now, come back to coming back to Meghamala which is the open stack implementation of open stack cloud. So, you can see here we are offering 3 types of virtual machines. So, one is IIT KGP regulator, IIT KGP large and IIT KGP extra large.

So, these are the specifications of this 3 type of virtual machines that we provide. And apart from that we currently provide 3 types of operating systems to be loaded in the virtual machines, which are Ubuntu, CentOS and Fedora.

So now,

So we will now go to our open stack installation and see how does it look like, from an administration power point of view. So, this is the dashboard of open stack. So, I am logging into it. So, this is the overview that you get when you login to log in as an administrator to the open stack cloud.

So, you can see total we have total 134 vCPUs is use, 2945 GB of disk, 316 GB of RAM and this much time of vCPU hour that is currently being used. So, this is an overall description of the cloud which is running. So now, we coming to instances which is where we will find what are the VM's that are currently running. So, see,

So, these are the different VM's running as of now.

Now, as you have seen in the previous video that this volume will provide you this is actually the cinder part of the cloud, which will list the number of volumes that are currently being used, images is actually the glance. So, these are the images that we currently have, but as you as you have seen you only provide 3 types of VM.

So, most of them we are not using we are not giving the public, these are for internal purposes. So, in access and security here you will have security groups. So, this is firewall kind of concept in respect to the cloud. So, what will have is that, there will be rules in a defined in each security group. Kind of the rules of the network I mean this means which type of traffic is allowed in a VM and which type of traffic is not allowed so.

Both we will respect to I mean incoming and outgoing traffic.

Yeah both incoming and outgoing based on.

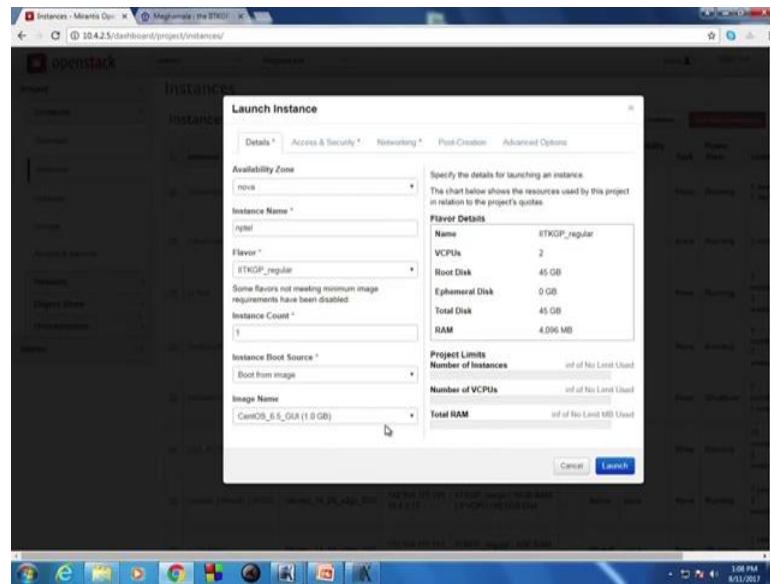
So, different port yeah type of services, right.

So, how with these helps is that when you create a VM a new VM you do not have to configure it is firewall independently. So, you can just assign the security group to it and automatically the firewall rules will apply. So now, if you come to the administration page. So, this was the user page as of the project page. So, admin user is also a tenant or a tenant of the open stack cloud and also it is an administrator. So, we got 2 components in the dashboard to one is project and one is admin. So, kind most of the things will be

same here, but the things which will differ I will show you that is hypervisors. So, here you will get the number of physical machines that are installed in our open stack cloud.

So, we have 3 physical machines each with 48 vCPUs, 20 GB of 25 GB of RAM. And 100 and one and currently 101 GB is used for the first one. Have you can see and 14 instances such running in the first compute node. So now we will try to create a VM, new virtual machine in our open stack cloud. So, coming to instances, and you will see here and there is a tab called launch instance. So, I am clicking that. So, currently we have only one availability zone which is nova.

(Refer Slide Time: 11:34)



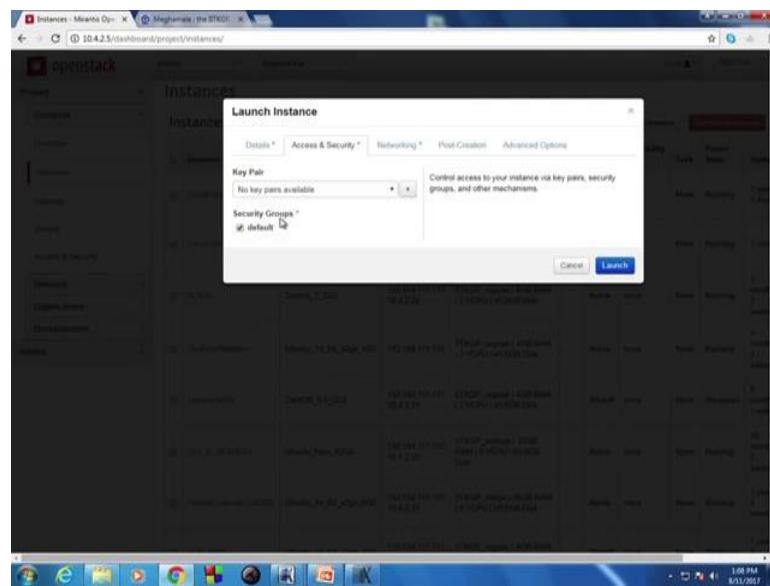
And just in typing a. NPTEL.

Typing the NPTEL.

VM name and the flavor. So, as small whatever yeah regular. So, I am giving IIT KGP regular.

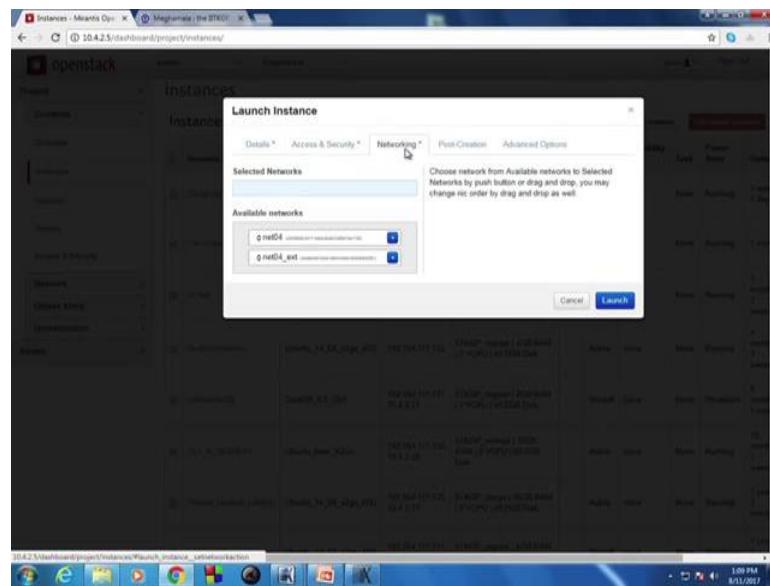
So, number of inst number of VM's that we want of this flavor you have putting one. So, instance boot source is where I am selecting image and pointing to the waste image that will be loaded in this VM.

(Refer Slide Time: 12:19)



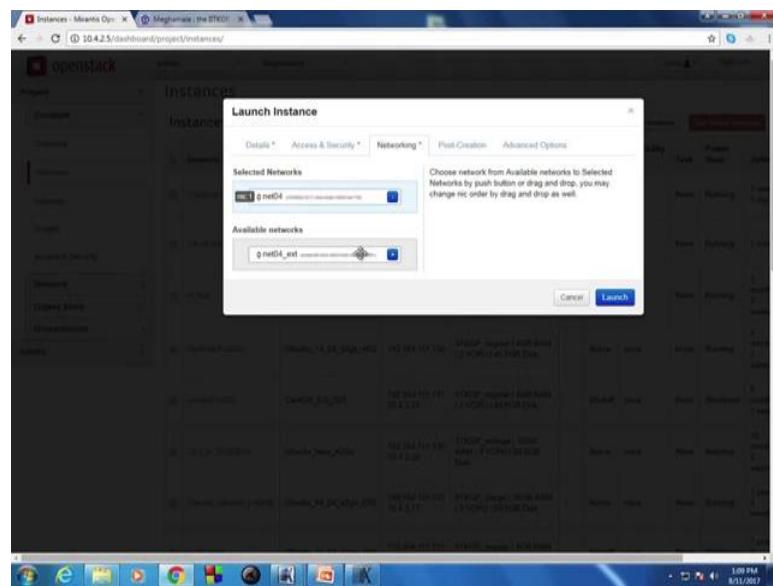
So, say I put CentOS. Now in access and security there is nothing to do you will can see as we have only one security group. So, it is currently selected. In networking part there are 2 networks one is external and one is internal to the cloud.

(Refer Slide Time: 12:24)



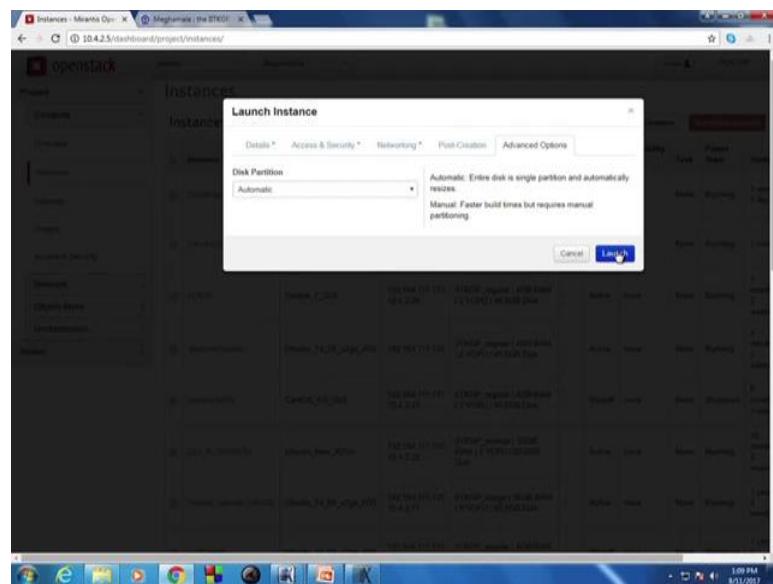
So, currently I will select the internal network and I will come back to this external network later.

(Refer Slide Time: 12:36)



So, post creation script you can give I am not getting anything here and this partition is automatic.

(Refer Slide Time: 12:44)



So, I will now press launch to launch the VM.

So, as you can see new VM came up here and it is current status is build.

So, it is building it will take some time.

(Refer Slide Time: 13:22)

Instance Name	Image Name	IP Address	Size	Key Pair	Status	Availability Zone	Task	Power State	Uptime	Actions
np01	CentOS_6.5_GUI	192.164.111.151	ITKGPU regular   4GB RAM   2 vCPU   45.0GB Disk	-	Build	nova	Spanning	No State	0 minutes	Associate Floating IP More ?
Cloud_np01t.2	Ubuntu_New_X2Go	192.164.111.151	ITKGPU large   8GB RAM   4 vCPU   160.0GB Disk	-	Active	nova	None	Running	1 week, 5 days	Create Snapshot More ?
Cloud_np01t.1	CentOS_6.5_GUI	192.164.111.150	ITKGPU regular   4GB RAM   2 vCPU   45.0GB Disk	-	Active	nova	None	Running	1 month	Create Snapshot More ?
esTest	Centos_7_GUI	192.164.111.153	ITKGPU regular   4GB RAM   2 vCPU   45.0GB Disk	-	Active	nova	None	Running	3 months, 3 weeks	Create Snapshot More ?
TestDiskPartition	Ubuntu_14_04_2Gps_45G	192.164.111.132	ITKGPU regular   4GB RAM   2 vCPU   45.0GB Disk	-	Active	nova	None	Running	4 months, 2 weeks	Create Snapshot More ?
centos01t01	CentOS_6.5_GUI	192.164.111.131	ITKGPU regular   4GB RAM   2 vCPU   45.0GB Disk	-	Shutoff	nova	None	Shutdown	8 months, 1 week	Start instance More ?
CL1_R_SERVER1	Ubuntu_New_X2Go	192.164.111.130	ITKGPU xlarge   32GB RAM   8 vCPU   160.0GB Disk	-	Active	nova	None	Running	10 months, 2 weeks	Create Snapshot More ?

So,

Rajesh has created 2 more VM's earlier. So, that the time can be the NPTEL 1 and 2 NPTEL already there you continue.

So, as I was saying that there is there are 2 networks one is internal and one is external. So now, you can assign the see now here. The internal network from internal network it has got the IP address. So, I will now allocate and external network IP address. So, that it can be accessed from outside the cloud.

(Refer Slide Time: 13:42)

Instance Name	IP Address	Size	Key Pair	Status	Power State	Uptime
np01	192.164.111.151	ITKGPU regular   4GB RAM   2 vCPU   45.0GB Disk	-	Build	No State	0 minutes
Cloud_np01t.2	192.164.111.151	ITKGPU large   8GB RAM   4 vCPU   160.0GB Disk	-	Active	Running	1 week, 5 days
Cloud_np01t.1	192.164.111.150	ITKGPU regular   4GB RAM   2 vCPU   45.0GB Disk	-	Active	Running	1 month
esTest	192.164.111.153	ITKGPU regular   4GB RAM   2 vCPU   45.0GB Disk	-	Active	Running	3 months, 3 weeks
TestDiskPartition	192.164.111.132	ITKGPU regular   4GB RAM   2 vCPU   45.0GB Disk	-	Active	Running	4 months, 2 weeks
centos01t01	192.164.111.131	ITKGPU regular   4GB RAM   2 vCPU   45.0GB Disk	-	Shutoff	None	8 months, 1 week
CL1_R_SERVER1	192.164.111.130	ITKGPU xlarge   32GB RAM   8 vCPU   160.0GB Disk	-	Active	Running	10 months, 2 weeks

So, we have some allocated IP address, we have some IP address see if we are finished with this list we can add this and new IP address from the pool will be generated.

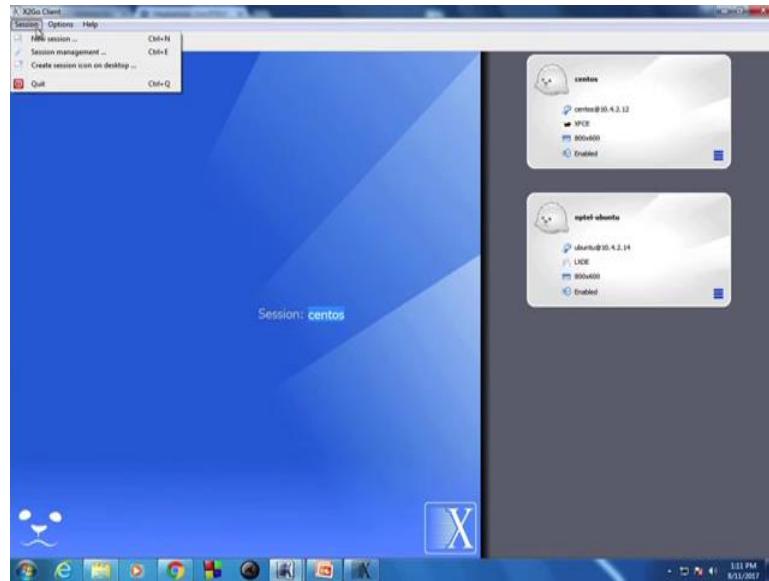
So, let us say this and I click on associate. So, as you can see, this new IP address yeah. So, this new external IP address is also associated with the VM. So now, I will show you how to connect to this VM from GUI frontend. So, for that I am using a software called x2go.

(Refer Slide Time: 14:39)



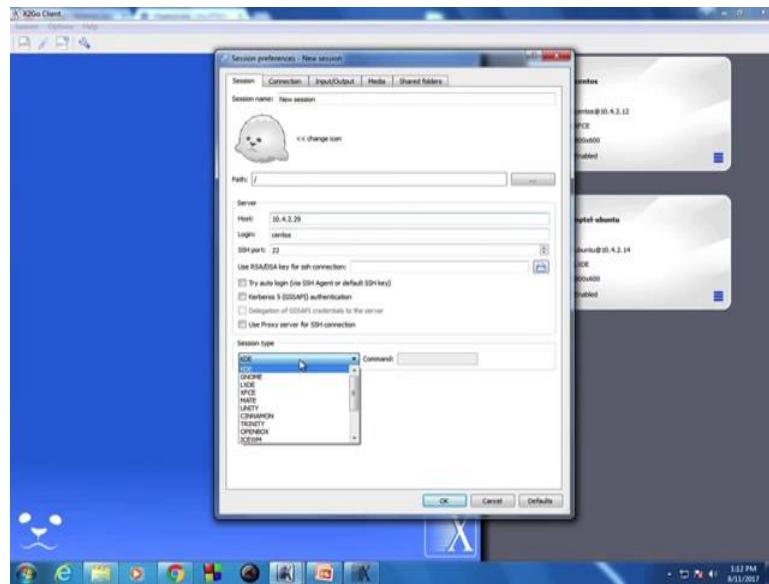
So, in our website in our cloud website, we have we have put the link and how to use it for the users. So, I have installed it and I am just showing you how do I connect with.

(Refer Slide Time: 14:52)



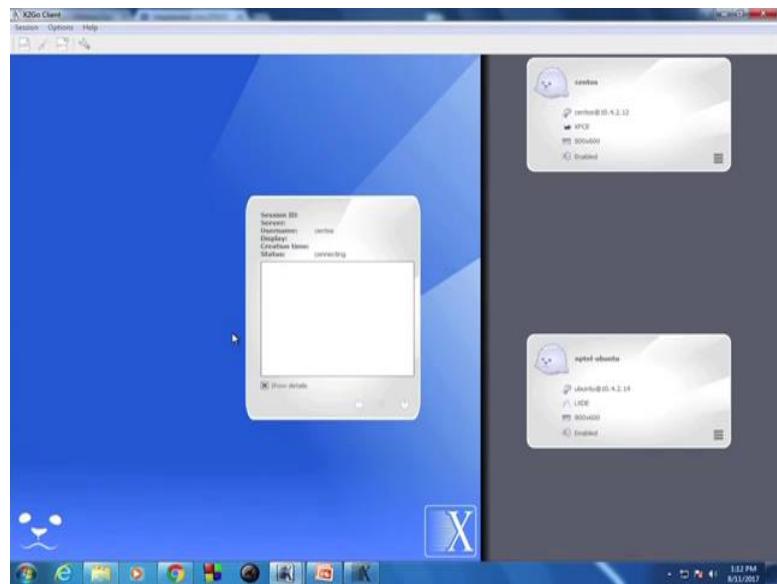
So, this is the x2go client. So, the server part is installed already installed again the VM. So, here what we have to do is or you have to create a new session and we have to put the credentials and the host ID and login ID of the VM.

(Refer Slide Time: 15:06).



For example our new VM was 10 4 2 29. So, this was the public IP address of the VM. So, and the login ID was CentOS and the session type for CentOS type of VM, we are we are having xfc installed as that desktop. So, I am selecting it.

(Refer Slide Time: 15:47)



So now, so entering the password and this will land us to the VM desktop. So, let us see. So, this is the VM that we have just created. And you can see the as usual options which you getting a CentOS machine are there. So, we are say opening at terminal and let say we want to check the internet connection why.

So, you know your enter that VM.

Yeah now you are using the VM yeah, yeah now I am inside the VM yeah.

So, from the x2go yeah from the go claim you like the enter the meghamala that is sent wise VM right say.

So, let us check whether internet. So, inside the VM. So, as you can see the from inside the VM we can access internet and it is as usual like any other machine. So, this is no different from any other machine running CentOS or whatever operating system you have chosen so.

So, it when it come comes up you can show some other aspects of the now what I am seeing that from the this you know open stack dashboard that. So, tangential will come that before that like it is running down it is same way yeah that NPTEL. So, it is any character yeah more characterizations are there yeah, yeah.

So, when you build it we saw that the status first build. So now, that it has been it has been build and built. So now, the status changed to active and the power state is running. So, here are other few options from here which are useful when you are administrating for example, shut down, but that logging itself yeah was So that yeah activities of the VM yeah and terminate the instance we will actually delete the VM and will I mean delete the VM from the cloud.

So, it releases that yeah. So, it will releases the all the resources that that was the allocated do it. So, just check that whether that your youtube and alright this is alright. So, as you can see youtube is running in this VM that is the introductory yeah this is also the introductory. So, it is running over the VM I mean using the VM that going to truth.

So, you can do other computing and etcetera yeah everything is same. So now, as we have used VM. So now, let us see how to terminate the VM and release all the resources. So, here I am clicking on the terminating step, step. And the option is straight forward terminating. So, this will terminate the our instance and release the resources as you can see this will be no longer available here ok.

So, schedule termination of instance NPTEL it may takes sometime, but it is deleting yeah it is deleting. So, that is resource will be released yeah. So, that is one. So, that is a overall I mean quick demo on the things again what we wanted to show you that you can have your own small scale a open source like in this is case of open stack installation. And you can do lot of experiment. So, you can have a field of as a administrator how things are work also you can have a feel that as a user how people can work.

So, it will it is it will be nice that if you have couple of systems and install the open stack. And there are lot of nitty-gritty you need to follow the open stack installation thing which is true for any installation, but it is a good exercise to have a open source cloud of you.

Thank you, thank you.

**Cloud Computing**  
**Prof. Soumya Kanti Ghosh**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**

**Lecture – 17**  
**Case study with a commercial cloud: Microsoft Azure**

Hi, let us continue our discussion on cloud computing. Today what we will discuss is one of the one some demo on a commercial cloud namely Microsoft azure. So, to see that how it works. So, as such I need to I let me mention that there is no specific preference for using azure though it is a very popular and use worldwide. So, the more of a it is showing a demo on the things as my students are comfortable on this that we thought that there demo and you will be good. And Secondly, we will be using a free login of this is azure. So, that it will be easy for you to replicate those who were interested to you can basically create your own login, and try this type of things. And I have a fill that how really a commercial cloud works, right.

So, that is the major motivation and any other type of things any other type of commercial or open source cloud you can use there is no, no nothing binding on the thing, but just to show that how to work on. The thing azure as we know that it is popularly a popularly a more popular for a platform as a service maintained by Microsoft, have data centers Microsoft a data center across the globe, and as a huge user base for the things, right.

(Refer Slide Time: 01:43)

The slide has a dark blue header and footer. The main content area is yellow. At the top left, it says 'Microsoft Azure : An overview'. Below that is a bulleted list:

- Microsoft Azure is a growing collection of integrated cloud services which developers and IT professionals use to build, deploy and manage applications through a global network of datacenters.
- With Azure, developers get the freedom to build and deploy wherever they want, using the tools, applications and frameworks of their choice.

At the bottom right of the slide, it says 'Source: <https://azure.microsoft.com/en-in/>'.

In the footer, there are logos for IIT Kharagpur and NPTEL, along with the text 'NPTEL ONLINE CERTIFICATION COURSES'.

So, if you look at Microsoft Azure over view the it is a growing collection of integrated cloud services, which developers and IT professional used to build deploy and manage applications through a global network of datacenter right. Maintained by Microsoft in this case. With Azure developers gets the freedom to build and deploy, wherever they want using the tools application and frameworks of their choice.

So that means, it gives ubiquitous access to that for developing the things as we mentioned it is a primarily a PaaS type of services platform as a services it is extremely useful for IT professional. And developers to develop apps over the this over this platform. And any type of further matter it is true for any type of PaaS type of cloud, where the developers have a this developing platform, looking of developing different type of apps on the thing, right. With without having installation and infrastructure on their own their own premises. Microsoft Azure you can use there was a on the cloud or on one premise that sort of absence are also there.

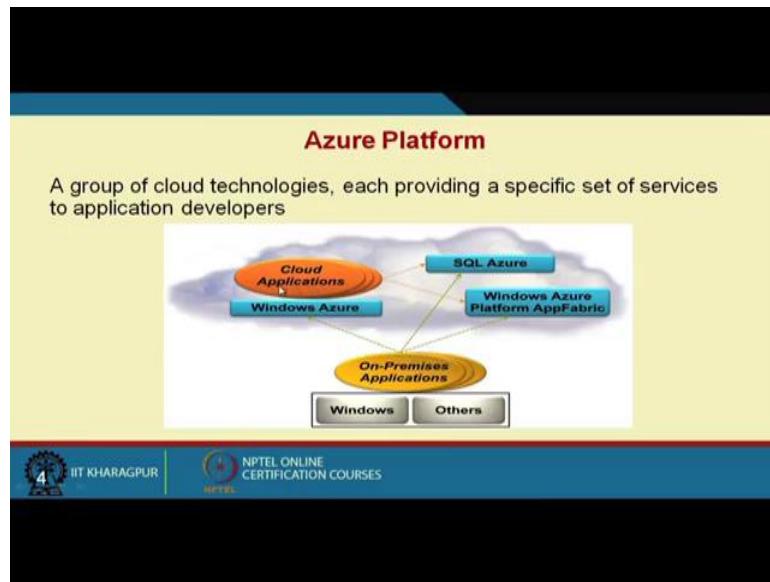
(Refer Slide Time: 02:56)



So, as I was mentioning there Microsoft as a global presence and of data centers. And this picture we have taken again most of the things we have taken from Microsoft azure website.

So, these are this picture maybe little figure may be little old than maybe there are different type of data center. The idea is to say that is a global presence of data sensors which are interconnected and user can hooking to the azure as a cloud without bothering that where the it is actual application or the platform is like.

(Refer Slide Time: 03:32)



So, if you look at the azure platform it is a group of cloud technologies is providing a specific set of services to the applications. So, one the core is that azure or the Microsoft azure thing where the cloud applications are launched there is a SQL azure which gives a SQL server type of support there is a windows azure platform, windows azure platform fabric right. And the windows or other users on premise applications they can hook into this components.

(Refer Slide Time: 04:04)

The slide has a yellow background with a black header bar. The title 'Major Components' is centered in red. Below the title is a bulleted list of three items:

- MS Azure
  - Provides a Windows-based environment for running applications and storing data on servers in Microsoft data centers
- SQL Azure
  - Provides data services in the cloud based on SQL Server
- AppFabric
  - Provides cloud services for connecting applications running in the cloud or on premises

At the bottom left, there are logos for IIT Kharagpur and NPTEL. At the bottom right, there is a circular profile picture of a man.

So, if you look at the major components as we mentioned one is the Microsoft azure, provides a windows based environment for running application and storing data on servers in Microsoft data centers. SQL azure provides data services in the cloud based SQL server, and app fabric provides cloud services for connecting application running in cloud or on premises. So, it is important that there is app fabric it is say provides cloud services for connecting applications which may be running on the cloud, or on premise on premise servers and every things. So, it gives this a versatile way of connecting to different type of applications with a backend databases. So, it is it is truly a development platform, and paas is suppose to be a developed platform.

(Refer Slide Time: 05:01)

### MS Azure

- Customers use it to run applications and store data on Internet-accessible machines owned by Microsoft
- Those applications might provide services to businesses, to consumers, or both

The diagram illustrates the Microsoft Azure architecture. At the bottom is a blue bar with the IIT Kharagpur logo and 'NPTEL ONLINE CERTIFICATION COURSES'. Above this is a yellow section titled 'MS Azure' containing the bullet points. To the right is a white diagram showing 'Businesses' and 'Consumers' represented by icons of people. Arrows from both groups point to a central box labeled 'Windows Azure'. This box contains 'Applications' (orange) and 'Data Store' (blue). Below the central box is the text 'Microsoft Data Centers'.

So, look at the other it look at a Microsoft azure customer, use to run application and stored data on data in internet accessible many machines or the which can be connected through the network connection and owned by or supported maintained by the Microsoft.

In case of Microsoft azure for other service provider there are other things those application might provide services to businesses to consumers or both right. So, provide services to business says services to consumer and look like here the picture shows that can be business oriented things like whole of your or a development things or partial of the development platform etcetera you leverage on the azure platform or it can be the individual consumer or customers which can connect to the things and here the azure fabric is there basic azure and which connect to it is backbone of bare metal systems, and it has data repository or data store. And then the applications which allows you to connect to the external world. So, and we have a broad network connectivity for serving the thing.

(Refer Slide Time: 06:16)

The diagram shows a central oval containing three horizontal bars labeled 'Compute' (red), 'Storage' (blue), and 'Fabric' (green). To the left, a small icon of a computer monitor with a bar chart is connected by arrows to a central box labeled 'Windows Azure'. This box is also connected to the 'Compute' and 'Storage' components within the oval.

**Azure**

- **MS Azure** is a foundation for running applications and storing data in the cloud
  - Provides *compute* and *storage* services for cloud applications

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

NPTEL

So, azure is a foundation for running applications and storing data in the cloud. So, it provides as a as a cloud service provider, it is a foundation for running application and storing in the cloud provides compute, and storage services for cloud applications right. So, similarly it is compute storage and fabric. So, fabric is primarily binding the things with the bare metal that different data centers the resources etcetera. So, it is it keeps a layer above the things which allows for virtualization another type of supports.

(Refer Slide Time: 06:54)

The diagram shows a central oval containing three horizontal bars labeled 'Compute' (red), 'Storage' (blue), and 'Fabric' (green). The 'Compute' bar has a sub-section labeled 'Running applications'. The 'Storage' bar has a sub-section labeled 'Storing and accessing data'. The 'Fabric' bar has a sub-section labeled 'Managing resources'.

**Components**

- Compute
  - Running applications
  - Support applications that have a very large number of simultaneous users and that can scale out
- Storage
  - Storing and accessing data
  - Applications require storage as simple blobs, a more structured way to store information, or a way to exchange data between different parts of an application
- Fabric
  - Managing resources
  - Providing a common way to manage and monitor applications that use this cloud platform

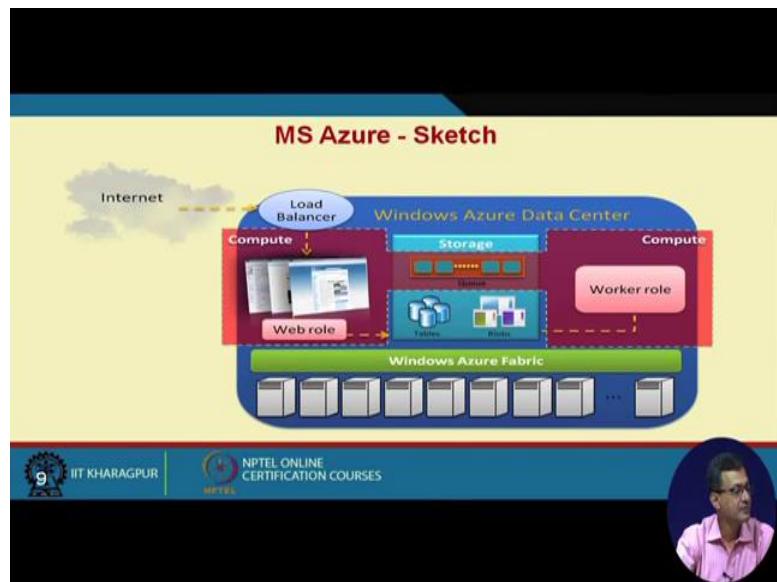
IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

NPTEL

So, if we look at the component wise that uses the one is the compute component run running applications. Support applications that have a very large number of simultaneous users and that can scale out, right. So, it can scale up scale down and type of things.

Storage storing and access accessing data, applications request storage as a simple as simple blobs a more structure way to store information or a way to exchange data between different parts of the applications. So, it is more of a storage type of services and the fabric managing resources. So, whatever the underlying resources there the azure fabric allows to manage these resources. So, provide a common way to manage monitor applications that use the cloud platform. So, it as a monitoring tool to see that how that uses are there and basically managing the underlining resources or the bear metal resources that the backbone.

(Refer Slide Time: 07:53)

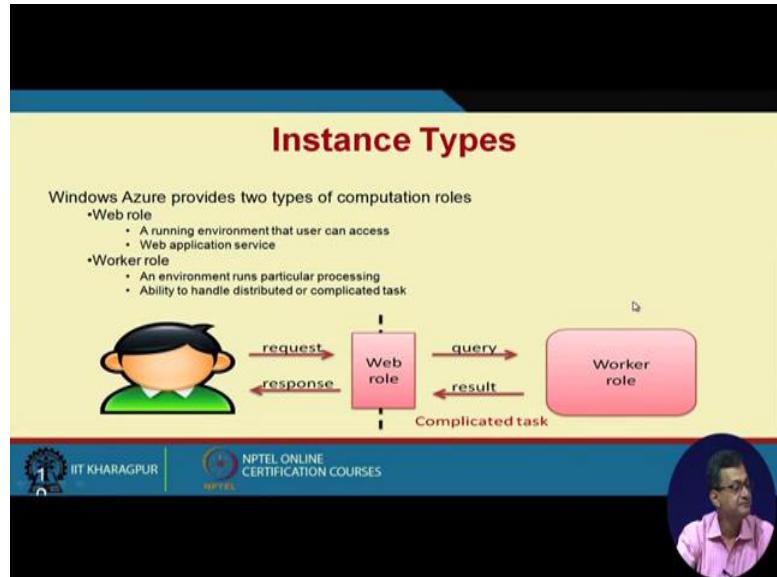


So, if you have the overall sketch. So, you have this bear metal at the things this azure fabrics which is the which basically interact with this resources or manages the resources. And then we have they are different type of thing compute storage compute and there are there are different things web role worker role, then tables blobs which are which are more about the storage more component of the storage part.

So, user are connected user customers or businesses connects to the things through a through the internet there is a load balancers, which look at that which distribute the load

which has the web role which allows you to the connector to the worker role of the compute.

(Refer Slide Time: 08:43)



So, if you look at the instance type as we have discussed in the last slide. So, azure provides 2 types of computational role, one is the web role a running environment that users can access right. So, the user can access through internet. And web application service. So, these are the 2 things which are the web role and the worker role. And environment running particular processing alright. And ability to handle distributed or complicated tag or complex task. So, this is the worker rule. So, user request it is a web role it goes to the query as a worker role and re results and response comes into other way.

So, it is more of a acts as a interface between the complex task or completed task with the other ended the user of the consumer.

(Refer Slide Time: 09:36)

The slide has a yellow header bar with the title 'Instance Types' in red. Below the title is a list of four bullet points:

- Any service must include *at least one role* of either type, but may consist of any number of web roles or work roles
- Worker role can *communicate* with Web role using the Windows Azure storage *queues*
- Each VM contains an *agent* to allow the application to interact with the Windows Azure fabric

At the bottom of the slide, there is a footer bar with the IIT Kharagpur logo, the text 'NPTEL ONLINE CERTIFICATION COURSES', and a circular profile picture of a man.

So, instance type any service must include at least one role of either type, but may consist any number of a web roles or worker roles right. So, that is important worker role can communicate with web role using azure storage queues right. So, the worker rule can communicate with the web role in the storage queues which we have seen in this picture see this is set of queues here it has a if this in the storage section of the thing or storage component. Each VM contains an agent a contains an agent to allow the application to interact with the azure fabric. So, each VM has a agent to which interact has a which can interact with the azure framework, which interact communicate with the underlining resources.

(Refer Slide Time: 10:30)

**Deploy anywhere with your choice of tools**

- Connecting cloud and on-premises with consistent hybrid cloud capabilities and using open source technologies

Build your apps, your way      Connect on-premises data and apps      Extend the cloud on-premises

Source: <https://azure.microsoft.com/en-in/>

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, what tries to give right. So, it deploy anywhere with your choice of tools. So, it gives a flexibility to give to deploy your application anywhere with the choice of tools. So, it connecting cloud an on premises with consistent hybrid capabilities and using open source technologies right. So, it has a ability it of cloud and on premises with consistent hybrid capabilities and using open source technology. So, it is what is a time. So, if you look at. So, it is one part of the looking at it is a build your apps your own way right. So, you can have your own apps connect on premise data and apps right. And extend the cloud on premises right. So; that means, one way is that you can develop your apps one way is that connect on premise data and app applications if required. And extend the cloud on premises right. So, you can extend those things pull it to this on premise applications.

So, what Microsoft try to focus on or try to give services that it is a it is a versatile of developing apps, it is a it and connect to your own on premise data and applications and you can extend that cloud component to with the on premise type of this. There are these are required for different type of applications or data which are something maybe a proprietary in nature, some are legacy applications running on the things. And some things which may not which may not be so much so much high level of security or high level of proprietary things are there which can always you can run on the cloud and have more versatile or quick access one things. So, they also provide some of the protection

mechanisms helps to protect assets through rigorous methodology and focus on security privacy compliance and transparency, right.

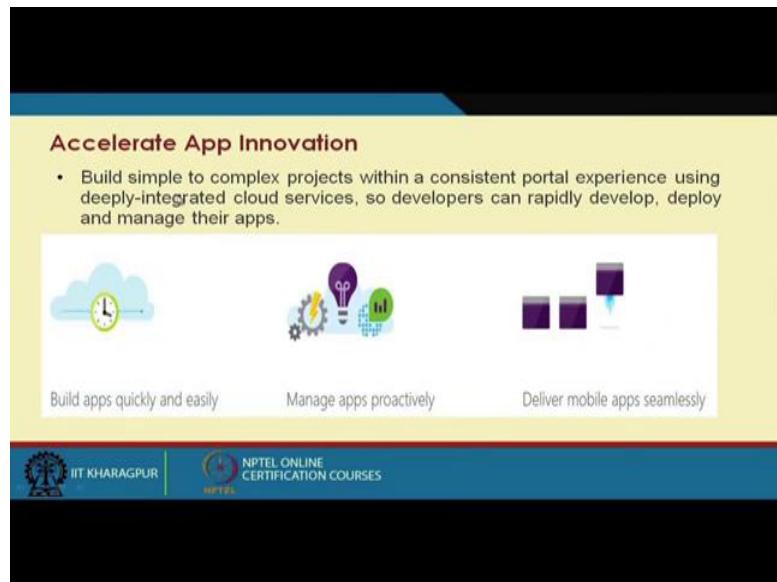
(Refer Slide Time: 12:41)



So, achieve global scale in local region detect and mitigate threats and rely on most trusted cloud.

So; that means, what it is tries to gives is that different security features privacy compliance transparency which gives a trust or a confident in the user base to use the assumes. So, that the thing and that is expected for any cloud provider that the selection of the cloud provider depends on how much how other than how much versatile or what type of services we are providing what type of security to my data and applications are there So that they try to pull it.

(Refer Slide Time: 13:26)



The other things are accelerate app innovations we will simple to complex projects within consistent portal experience using deeply integrated cloud service. So, developers can rapidly develop deploy and manage their applications.

So, that that exactly as we have discussed earlier also that exactly the one of the reasons we are going for cloud right. I need to rapidly develop deploy and manage applications. Now as azure is primarily a platform as a service. So, if there is the development of the things is much more had it been a software as a service. Then uses of the things may be more and had it been IAS type of service then virtual means availability or leveraging the virtual machines put have been the management focus. So, it build apps quickly and easily manage apps proactively. So, this management of the apps development of the apps management of the apps and deliver mobile app seamlessly. So, this apps can be quickly develop manage and deliver.

(Refer Slide Time: 14:36)

**Analytics and Data Services**

- Uncover business insights with advanced analytics and data services for both traditional and new data sources. Detect anomalies, predict behaviors and recommend actions for your business.

Add intelligence to your apps      Predict and respond proactively      Support your strategy with any data

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Analytics and data services, so is also what the claim that provide and not only they ask. So, uncover business insights with advanced analytics and data services for both traditional and new data source. So, detect anomalies predict behaviors and recommend actions for the businesses right. So that means, it gives several type of other Meta applications what we say is, not only or libraries which allows you to have some sort of a data analytics and data services out of it. So, which in turn allow to add integers to your apps. Predict and respond proactively on some situations. And supports your strategy with any data. So, this sort of things are provided by azure, so that means, it has a reach set of data analytics and data service type of components or laborites which are allows you to component.

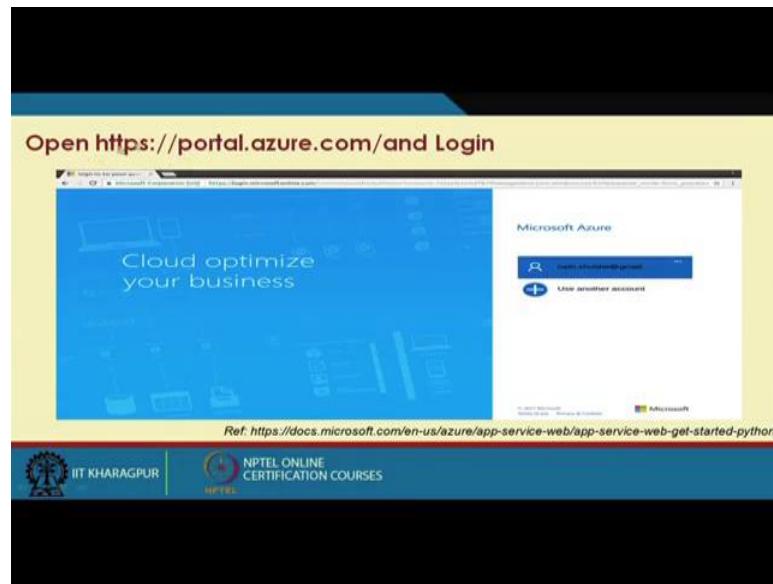
(Refer Slide Time: 15:39)



So, next thing what we will be doing to show you a very short demo on creation of a python web app in Microsoft area azure. So, it is it may not be that a very complex thing, but it is the idea is to show that how to use azure. So, what will be doing here, we will be primarily using a open means free license or free login of the things. The idea is that So that you can have a you can yourself try without before going to paid login which has much more resources and might most applicable means much more applications development resources. So, before that you can basically try out your hand that how the azure think will be there; so, before one of my scholar TA of this course.

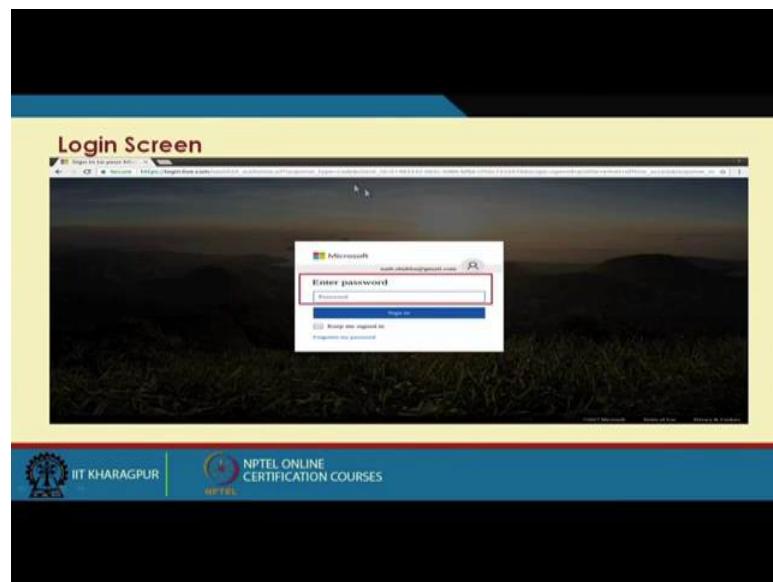
So, you are running on demo on these particular things, I will just quickly go through that some of the steps. So, there may be little variant variation on the actual steps where we showing, but it will be easy to look at it.

(Refer Slide Time: 16:55)



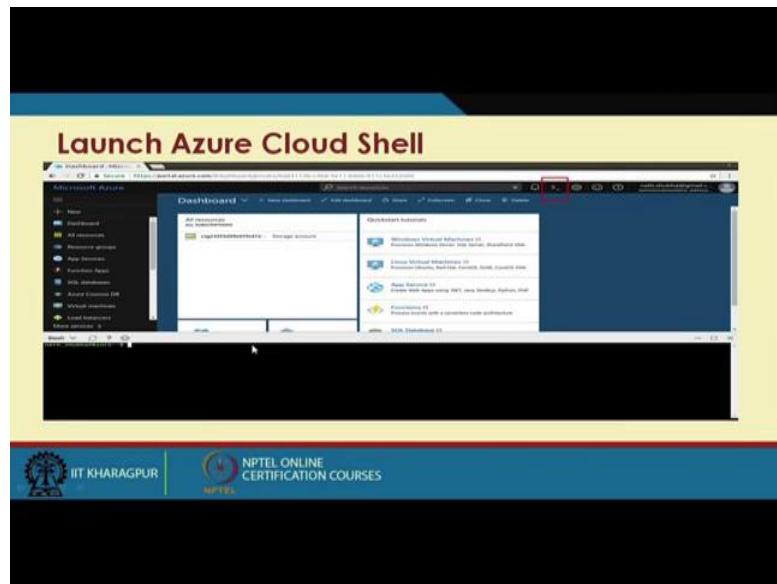
So, I can you can open this portal dot azure dot com I will log in to the things. So, with your login and password.

(Refer Slide Time: 17:03).



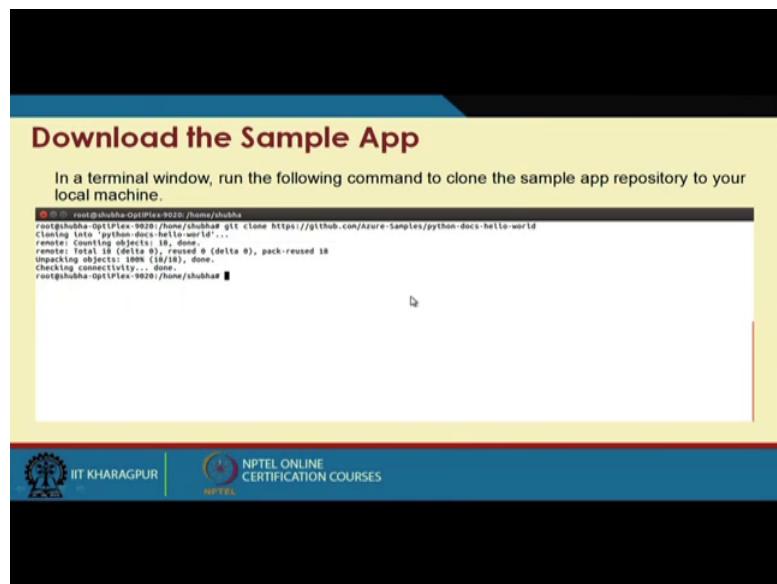
So, the login screens somewhat looks like this. And you can launch azure cloud cell after login into the this azure platform.

(Refer Slide Time: 17:07)



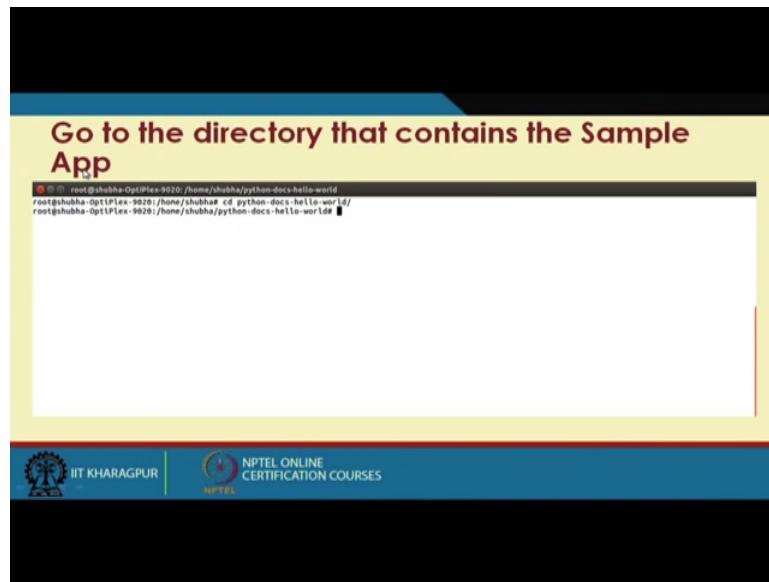
And you can basically download a sample app. So, in the terminal window run the following command to clone this sample app repository to your local machines.

(Refer Slide Time: 17:16)



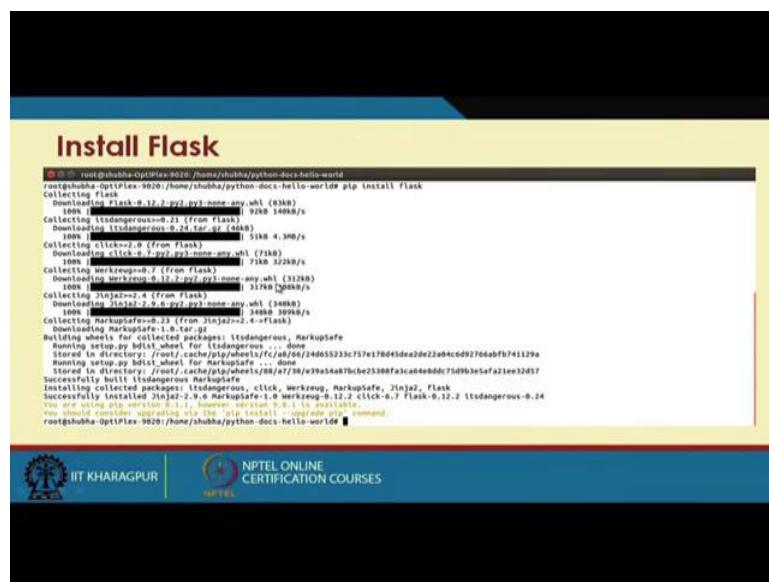
So, idea is that you can download this a particular sample app, which will allow which will be easier for the beginners to update the tabs according to your need, and then upload the thing same to the azure platform and run on the thing.

(Refer Slide Time: 17:47)

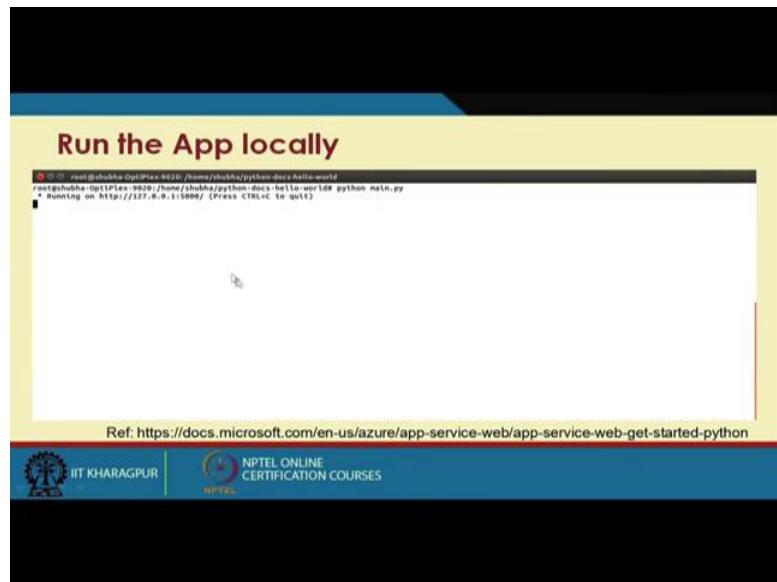


So, go to the directory that contains this downloaded sample app, in your local directory here where you go to the directory and then install the flask.

(Refer Slide Time: 17:55)

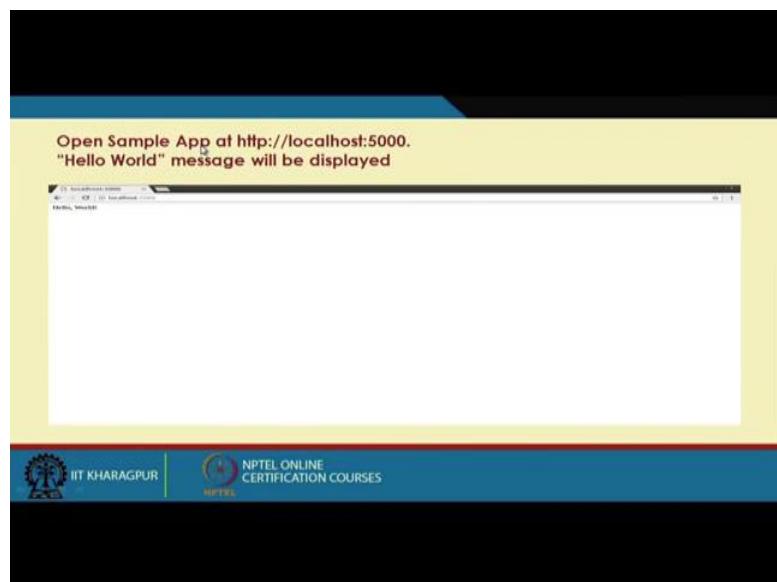


(Refer Slide Time: 17:58)



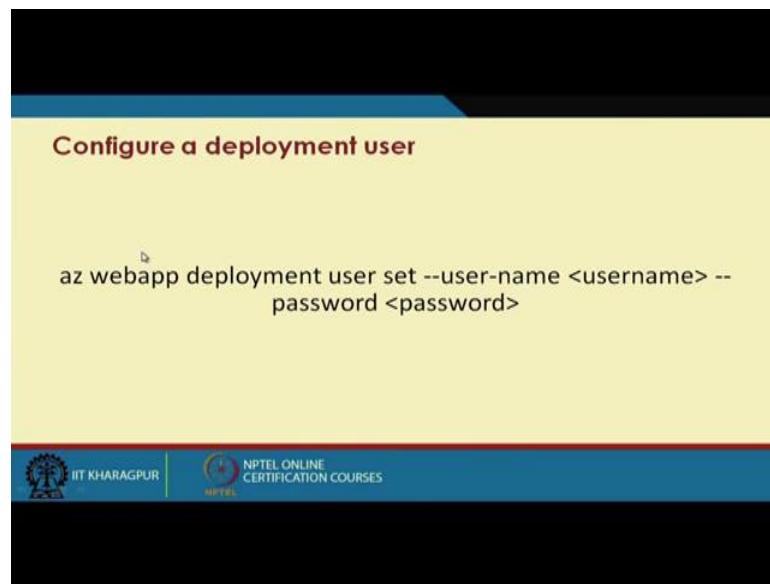
So, run the app locally once you check it that whether it is running locally.

(Refer Slide Time: 18:06)



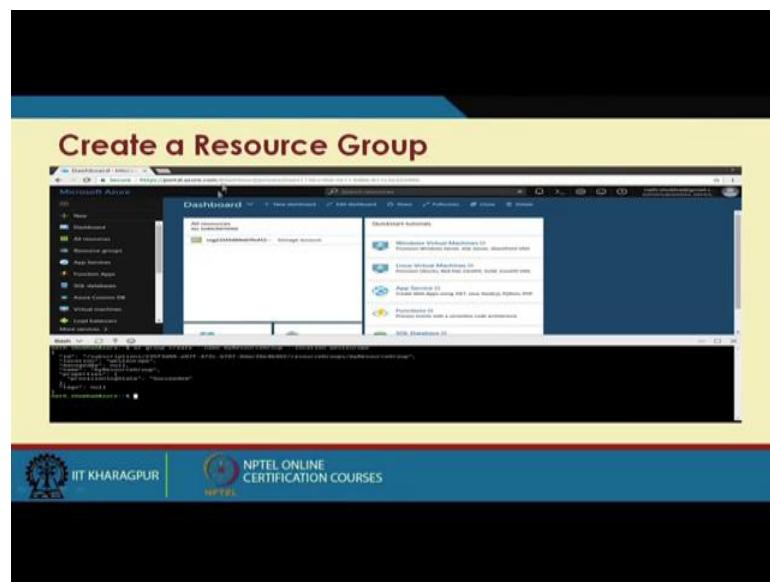
So, open the app in the by local dot by this particular things on a 5000 port and hello world message will be displayed that you can test. Then configure a deployment user by using this is the comment that show you.

(Refer Slide Time: 18:16)



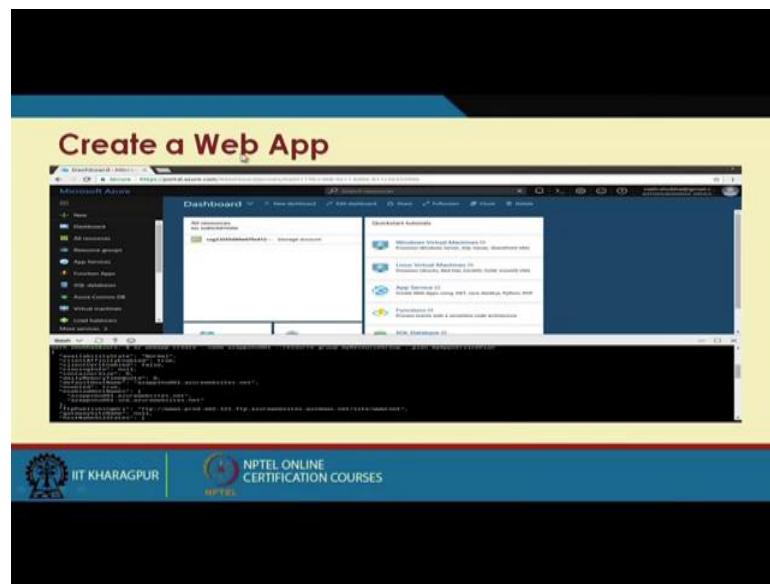
Creates a resource group in the dashboard of the azure. Create an azure app service plan.

(Refer Slide Time: 18:24)

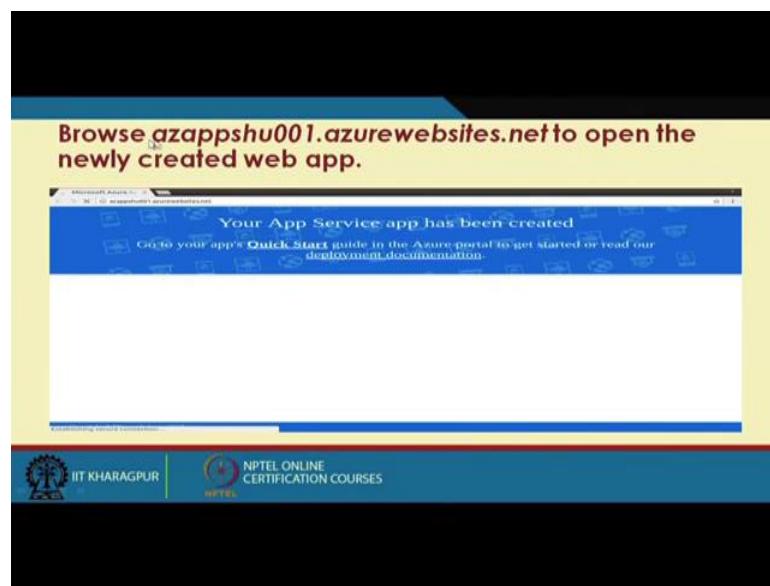


Create a web app of your own. And then browse through this particular URL to open the newly created app correct.

(Refer Slide Time: 18:38)



(Refer Slide Time: 18:42)



So, that you basically you, so you create your own app and learn on the as your platform.

(Refer Slide Time: 18:56)



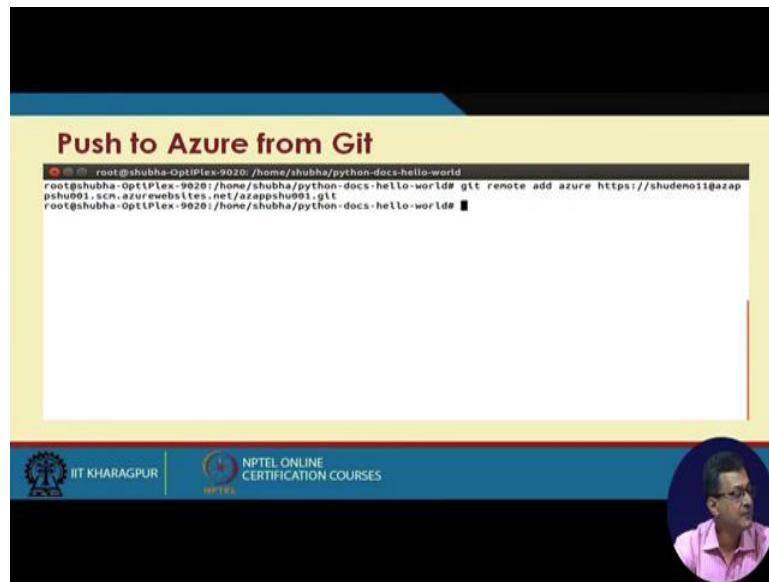
So, then you configure python in the in the particular using the azure dashboard.

(Refer Slide Time: 19:08)



Configure local kit which will allow you to sing the data between the between your local machine and the that in the azure and the azure platform.

(Refer Slide Time: 19:19)



So, push to azure form get that will show you that that these are the commands. Push to the azure from the remote to deploy the particular app, and browse this again through this particular run it, right.

(Refer Slide Time: 19:26)



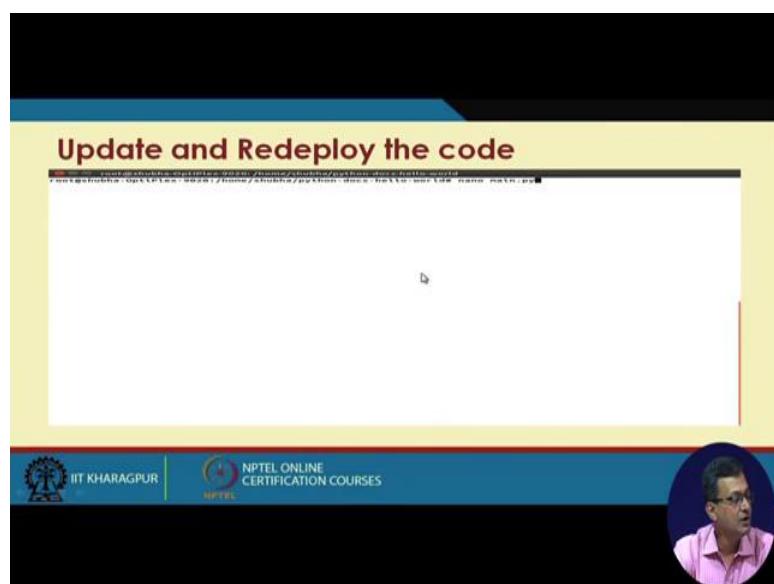
So, and update and re-deploy the, but code.

(Refer Slide Time: 19:31)



So, use a local text editor open main dot py that python file in the python app to make the changes, so whatever you want to make changes the python think.

(Refer Slide Time: 19:37)



(Refer Slide Time: 19:41)



Using a local text editor, open main.py file in the Python app to make changes

```
root@shubha-OptiPlex-9020:~/home/shubha/python-docs-hello-world
GNU nano 2.5.3 File: main.py
Modified

from flask import Flask
app = Flask(__name__)
@app.route('/')
def hello_world():
 return 'Welcome to the NPTEL course on Cloud Computing!!'
if __name__ == '__main__':
 app.run()
```

Get Help | Write Out | Where Is | Cut Text | Uncut Text | Justify | To Linter | Cur Pos | Go To Line | Prev Page | Next Page

IIT Kharagpur | NPTEL ONLINE CERTIFICATION COURSES



Like welcome to NPTEL course on cloud computing may be the simple a state forward text which is there, you can do other complex applications if you want.

(Refer Slide Time: 20:03).



Commit the changes in Git

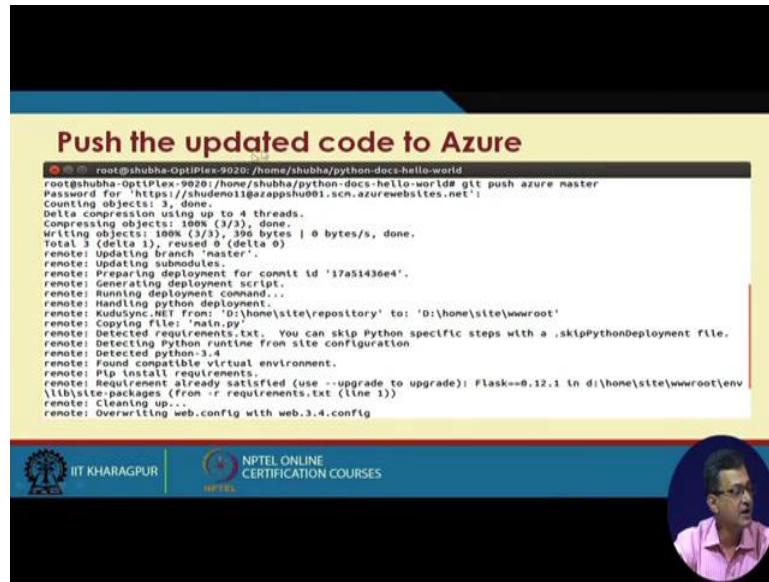
```
root@shubha-OptiPlex-9020:~/home/shubha/python-docs-hello-world
root@shubha-OptiPlex-9020:~/home/shubha/python-docs-hello-world# git commit -am "updated output"
[master 17a5143] updated output
 1 files changed, 1 insertions(+), 1 deletion(-)
root@shubha-OptiPlex-9020:~/home/shubha/python-docs-hello-world#
```

IIT Kharagpur | NPTEL ONLINE CERTIFICATION COURSES



So, commit the changes in the GIT. So, so that it is a synced with that as you push and upload the code to the azure.

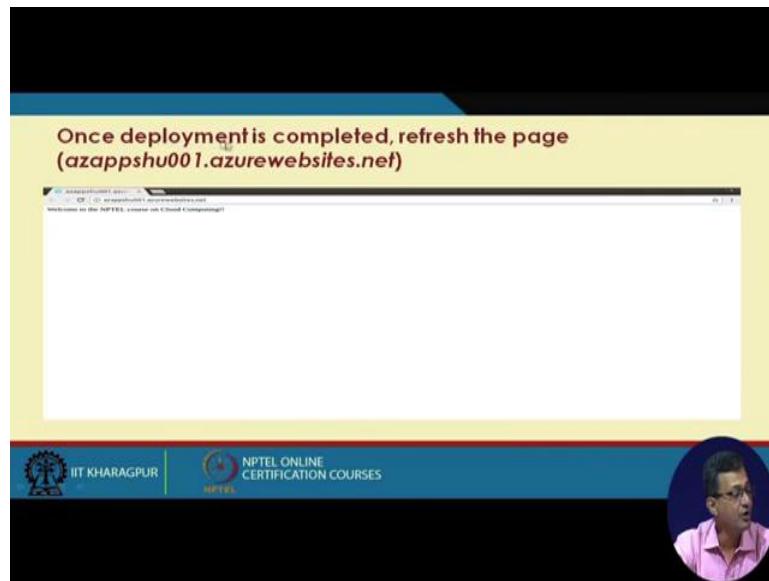
(Refer Slide Time: 20:11)



```
root@shubha-OptiPlex-9020:/home/shubha/python-docs-hello-world# git push azure master
Password for 'https://shudemo1@azappshu01.scm.azurewebsites.net':
Compressing source files... done.
Delta compression using up to 4 threads.
Compressing objects: 100% (3/3), done.
Writing objects: 100% (3/3), 163 bytes | 0 bytes/s, done.
Total 3 (delta 1), reused 0 (delta 0)
remote: Updating branch 'master'.
remote: Updating submodules...
remote: Generating deployment...
remote: Running deployment command...
remote: KuduSync.NET From: 'D:\home\site\repository' to: 'D:\home\site\wwwroot'
remote: Copying file: 'main.py'
remote: Detected requirements.txt. You can skip Python specific steps with a .skipPythonDeployment file.
remote: Ignoring Python runtime from site configuration
remote: Detected python-3.4
remote: Found compatible virtual environment.
remote: Requirements already satisfied (use --upgrade to upgrade): Flask==0.12.1 in d:\home\site\wwwroot\env\lib\site-packages (from -r requirements.txt (line 1))
remote: Cleaning up...
remote: Overwriting web.config with web.3.4.config
```

So, these are the commands which again will be shown on that means, one a life demo.

(Refer Slide Time: 20:21)



Once the deployment is completed replace the page. So, that once you replace the page you can see that that particular URL that will come to NPTEL course in the cloud computing correct. So, what you will do now we will we will show you a particular the same exercise maybe little bit a few more screen will be there step by step, that how you can run a simple app in using Microsoft azure. So, the major idea is to so, that you can have a feel of how a commercial cloud there. So, if you if you use any other open source

cloud you of paas type of things there the command line may be different the type of steps maybe little bit different, but never the less the basic philosophy will be the same. So, there is the idea is to give have a direct feel of the things. Thank you, so what we will do we will continue with the with the demo subsequently.

**Cloud Computing**  
**Prof. Soumya Kanti Ghosh**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**

**Lecture - 18**  
**Demo**

Hi. So now, in continuation with our Demo on Azure; so we will be showing a live demo on as Azure system with a free login. So, the idea is as I mentioned that to; so, that you can also have a fill of a commercial cloud how it works etcetera. So, the application what we are going to demonstrate here and will be a very simple python web app the idea is that how you can develop your other this sort of for other apps in a in azure, right. So, with me Shubhabrata with be there.

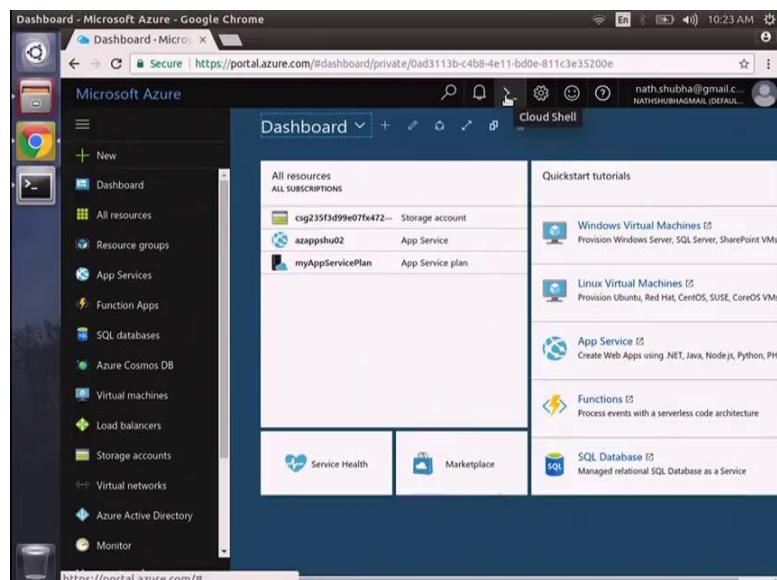
Hello.

Shubhabrata will be showing a live demo on the on a system. So, that you can follow the step though there are different variant of doing that, but this is one of the one of the standard step 1 of the standard procedure to look at the; to run any Azure app. So, I will give request Shubhabrata to start that.

Yes.

Demo show

(Refer Slide Time: 01:33)



Hello everyone, in this demo we are going to present the creation of a python web application in Microsoft Azure. Microsoft web applications are highly scalable self patching and it also provides us to develop our local application, the portal azure portal that is.

Portal.

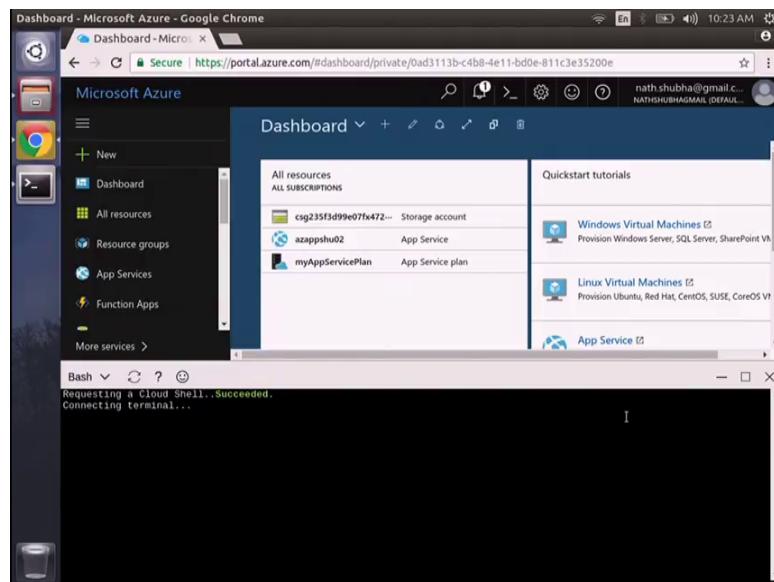
Portal dot azure dot com; so then we need to give our login credentials over here. So, here we need to give our user name and password, right.

So, Shubhabrata is have already having a login. So, he is using his own login otherwise you need to create it your own your; create your new login.

Right it is redirecting me to that Microsoft account sign in page; there I need to provide the password, right. So, now, we will be seeing the dashboard; all right, this is the dashboard for Microsoft azure. So, now, we need to launch a terminal that is azure terminal. So, that is this one the cloud shell.

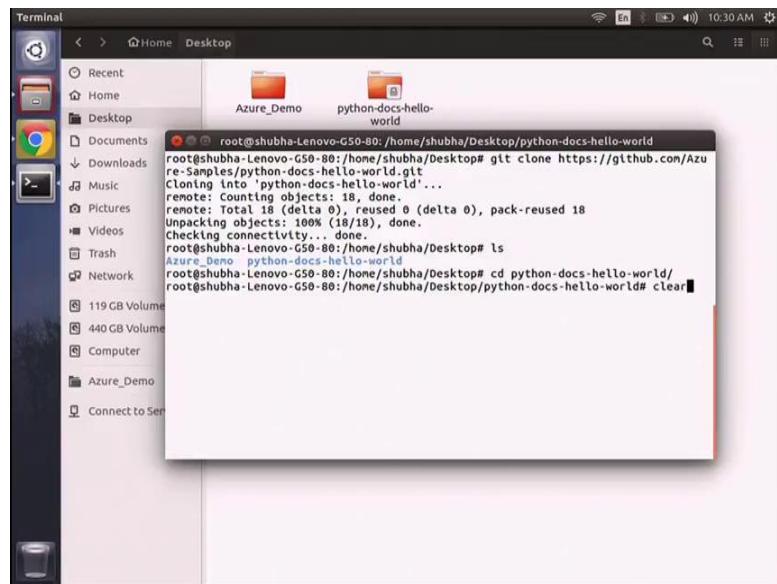
Cloud shell.

(Refer Slide Time: 03:51)



Right; so, that has been completed now it is connecting terminal, right. So, now, we need to do a local development of this python web application.

(Refer Slide Time: 04:12)



So, for that I will be downloading a existing project from Git. So, the command for that is git clone https colon double slash github dot com. So, here we have a demo for hello world.

Right; so, it is cloning the project directory into my local machine.

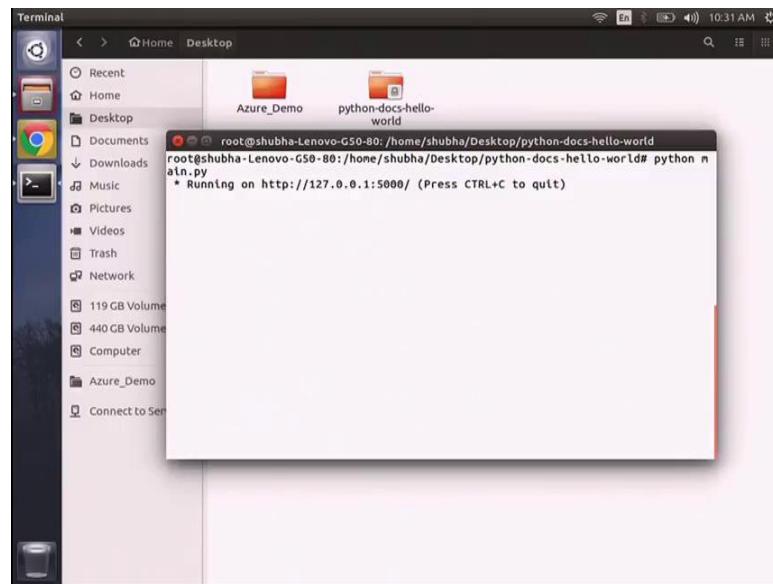
Need to revert it back.

Now, we need to go to that project folder. So, as you can see the project folder is this one, so cd right. So, here we have the contents. So, if I do cat of this main dot python file we can see.

Hello, World!

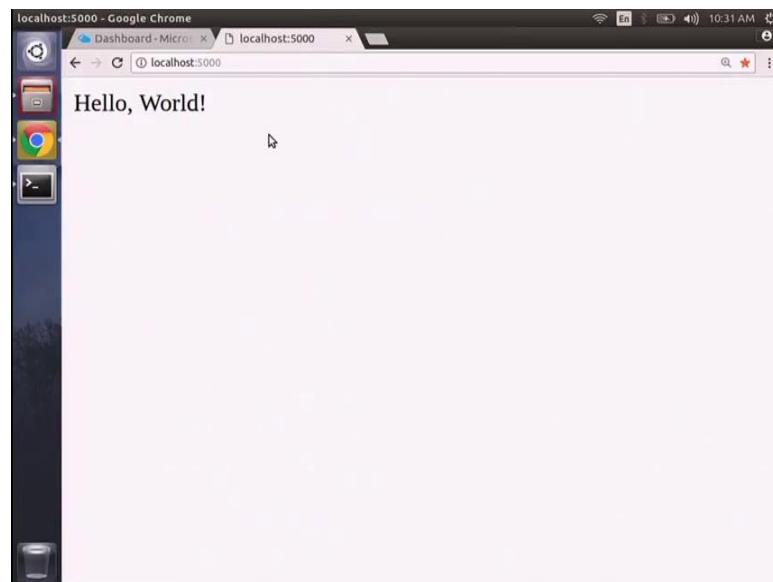
That hello world code has been written over here now in order to do this running of this in order to run this application we need flask; flask is a library which python uses and it provides us a framework that is microwave framework in python. So, I will be installing flask.

(Refer Slide Time: 06:03)



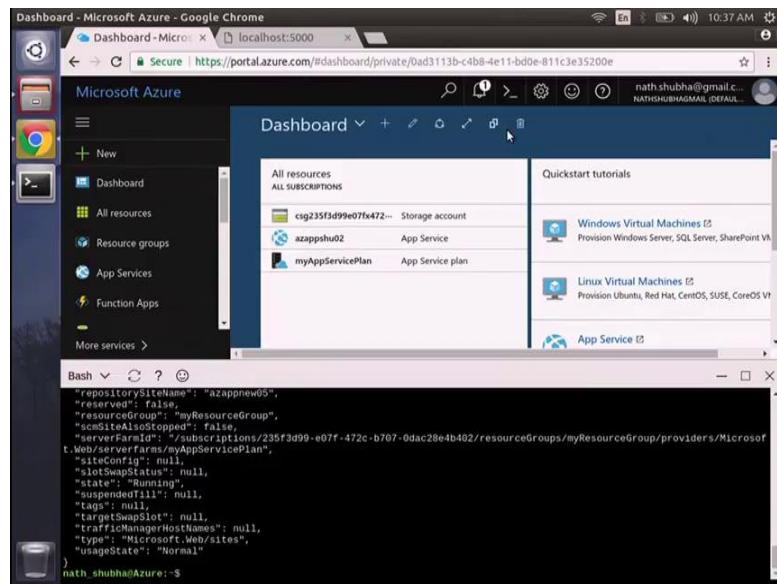
Now, I will be running this web application right in the browser I write localhost 5000.

(Refer Slide Time: 06:18)



So, it is showing me hello world, right. So, I will be pushing this project to our Microsoft as your, this portal so that we can host our local project in Microsoft azure.

(Refer Slide Time: 06:28)

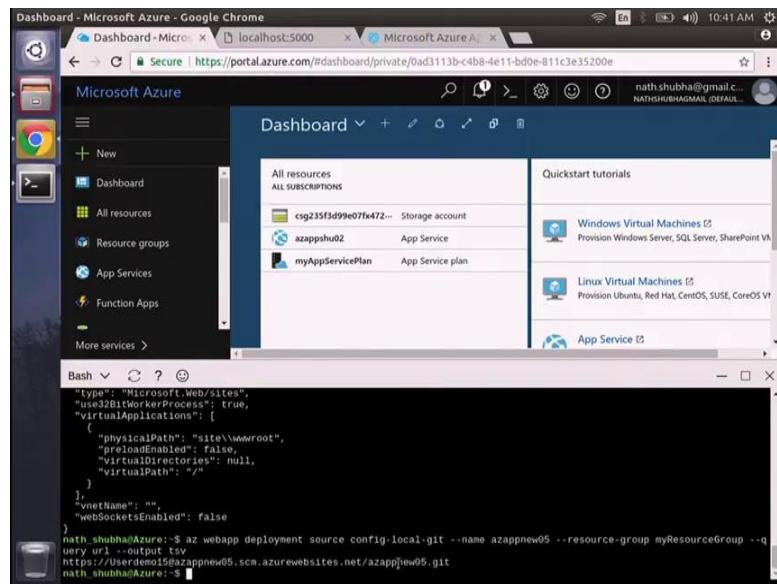


Now, in the Azure client we need to set a deployment user. So, this deployment user is required for FTP and local grid deployment to a web application. So, the command for this is a web app deployment user; set user name, I need to give; this will create a deployment, right. So, deployment user has been created with this user name and password. Now we recur a resource group. So, resource group is a logical container into which as your resource says like web applications data bases and storage accounts are deployed and managed. So, the command for that is az group create and your name for the resource group location, right. So, this will create a resource group resource group.

So, you been initiated a deployment user followed by a resource group. So, resource group is created.

Right; now we require an app service plan. So, an app service plan specifies the location app service plan specifies the location size and features of the web server that is that hosts our application, right. So, the command is az app service. This will create an app service plan, right. Now the web application generally provides a hosting space for our code and provides a url to view the deployed app. Now I will be creating an web application the command is az web app create name, right. So, now, if I go to the web browser and type the web app name followed by the dot azure websites dot net I will be able to view the webpage.

(Refer Slide Time: 10:02)

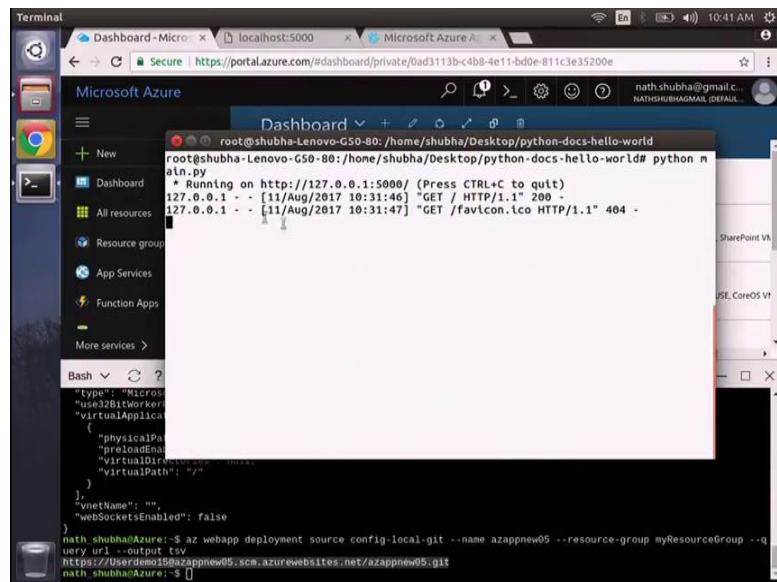


So, that now as this code is python based, we require we require to configure this as your; in order to use python. So, so I will be using a command this command will configure the corresponding python version with this web application. So, the command is az web app config space z python.

So, setting the python version this way uses the default container provided by this platform. Now we need to configure; configure local git deployment. So, this app service supports several ways to deploy content to a web app such as FTP local git; git hub, visual studio routine services etcetera for this demo we will be deploying using local git, Local git.

That means we will deploy by using the git command to push from our local repository to as your repository. So, the command is az web app deployment, right. So, we need to copy this URL and you need to go to terminal I need to open another terminal git remote add.

(Refer Slide Time: 11:38)



The screenshot shows a terminal window with a background image of the Microsoft Azure dashboard. The terminal is running as root on a Lenovo G50-80 laptop. It displays the output of a Python application running on port 5000, showing log entries for requests to / and /favicon.ico. Below that, a command is shown to deploy a web app named 'azappnew05' to an Azure resource group 'myResourceGroup'. The command uses 'az webapp deployment source config-local-git' and specifies the URL 'https://userdem01@azappnew05.scm.azurewebsites.net/azappnew05.git'.

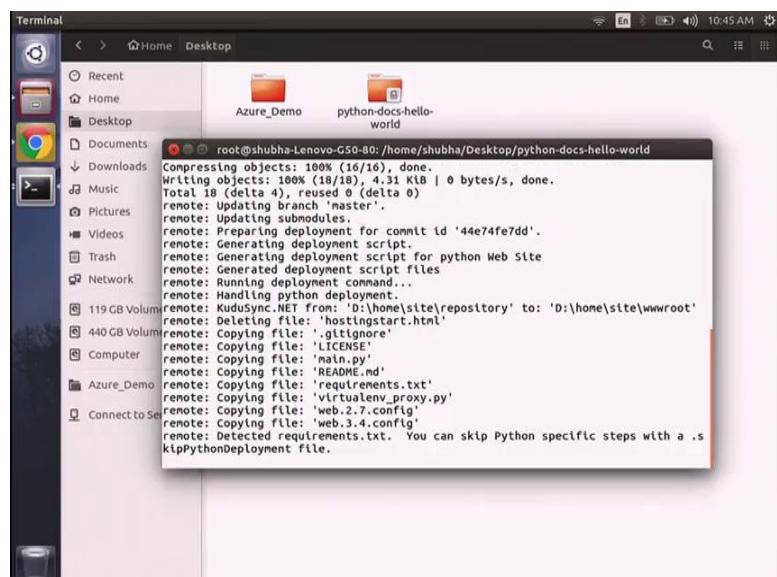
```
root@shubha-Lenovo-G50-80:~/home/shubha/Desktop/python-docs-hello-world
root@shubha-Lenovo-G50-80:~/home/shubha/Desktop/python-docs-hello-world# python m
aln.py
 * Running on http://127.0.0.1:5000/ (Press CTRL+C to quit)
127.0.0.1 - - [11/Aug/2017 10:31:46] "GET / HTTP/1.1" 200 -
127.0.0.1 - - [11/Aug/2017 10:31:47] "GET /favicon.ico HTTP/1.1" 404 -
nath.shubha@Lenovo-G50-80:~$ az webapp deployment source config-local-git --name azappnew05 --resource-group myResourceGroup --q
uery url --output tsv
https://userdem01@azappnew05.scm.azurewebsites.net/azappnew05.git
nath.shubha@Lenovo-G50-80:~$
```

Azure. Azure. That URL.

That URL; for this we need to issue this command git, sorry, git push azure master, right. So, here it will prompt us to give a password. So, the password should be the password which we have given during the application of the deployment user, right. So, this command is asking for the corresponding password.

Password; so, that is we need to specify here up to we are going; now this will push to the azure remote to deploy our web application.

(Refer Slide Time: 13:07)



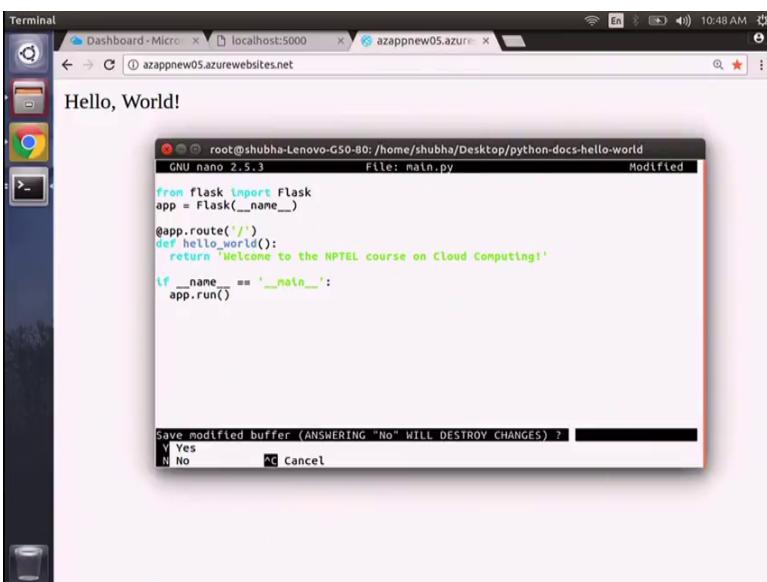
The screenshot shows a terminal window with a background image of a file explorer. The terminal is running as root on a Lenovo G50-80 laptop. It displays the output of a 'git push' command to an Azure repository. The command pushes changes from the local 'master' branch to the 'master' branch on the remote 'D:\home\site\repository'. The output shows the files being copied, including 'hostingstart.html', 'LICENSE', 'main.py', 'README.md', 'requirements.txt', 'virtualenv\_proxy.py', 'web.2.7.config', and 'web.3.4.config'. A note at the end of the output says 'remote: Detected requirements.txt. You can skip Python specific steps with a .skipPythonDeployment file.'

```
root@shubha-Lenovo-G50-80:~/home/shubha/Desktop/python-docs-hello-world
root@shubha-Lenovo-G50-80:~/home/shubha/Desktop/python-docs-hello-world# git push
Compressing objects: 100% (16/16), done.
Writing objects: 100% (16/16), 4.31 KB | 0 bytes/s, done.
Total 1B (delta 4), reused 0 (delta 0)
remote: Updating branch 'master'.
remote: Updating submodules.
remote: Preparing deployment for commit id '44e74fe7dd'.
remote: Generating deployment script.
remote: Generating deployment script for python Web site
remote: Generated deployment script files
remote: Running deployment command...
remote: Handling python deployment.
remote: KuduSync.NET from D:\home\site\repository' to: 'D:\home\site\wwwroot'
remote: Deleting file: 'hostingstart.html'
remote: Copying file: 'LICENSE'
remote: Copying file: 'main.py'
remote: Copying file: 'README.md'
remote: Copying file: 'requirements.txt'
remote: Copying file: 'virtualenv_proxy.py'
remote: Copying file: 'web.2.7.config'
remote: Copying file: 'web.3.4.config'
remote: Detected requirements.txt. You can skip Python specific steps with a .skipPythonDeployment file.
```

So, it will take some time. Then you recreate the application.

Right, so it has deployed successfully now we need to go to this for me that message hello world right, right. So, now, I will do a small change to our local project folder. So, I will open that main dot python file; we will do a change over here, I will change this message, let us say that is this one.

(Refer Slide Time: 14:48)



```
root@shubha-Lenovo-G50-80:/home/shubha/Desktop/python-docs-hello-world
GNU nano 2.5.3 File: main.py Modified
from flask import Flask
app = Flask(__name__)

@app.route('/')
def hello_world():
 return 'Welcome to the NPTEL course on Cloud Computing!'

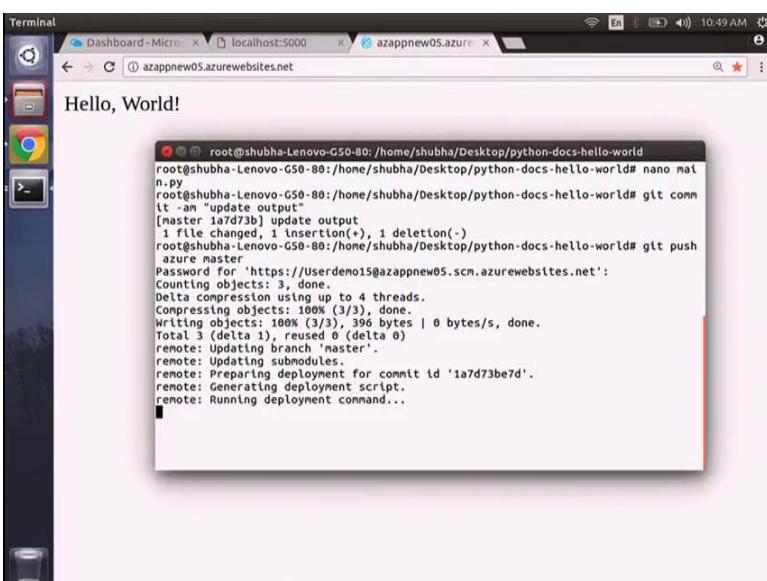
if __name__ == '__main__':
 app.run()

Save modified buffer (ANSWERING "No" WILL DESTROY CHANGES) ?
```

Y Yes  
N No  
^D Cancel

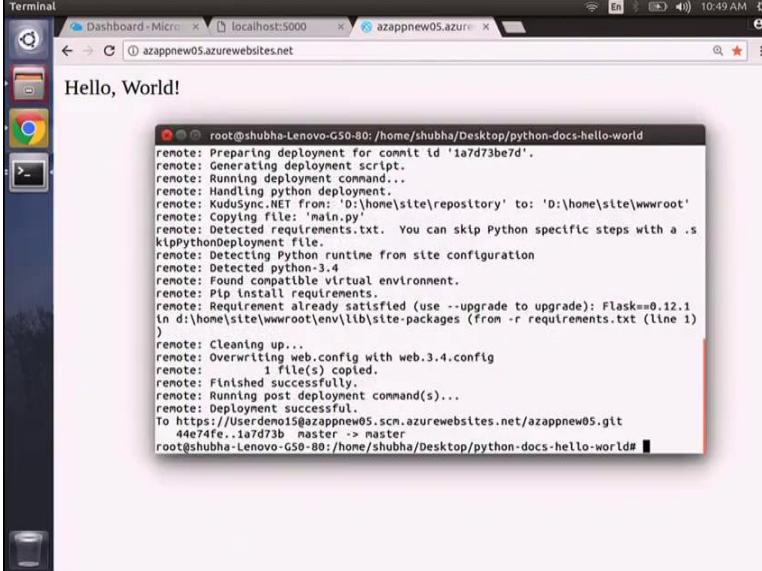
Welcome To the NPTEL Course on cloud computing, right, now we need to comment these changes in our git; so, for that I will use git commit. So, it has been able to update.

(Refer Slide Time: 15:27)



```
root@shubha-Lenovo-G50-80:/home/shubha/Desktop/python-docs-hello-world
root@shubha-Lenovo-G50-80:/home/shubha/Desktop/python-docs-hello-world# nano main.py
root@shubha-Lenovo-G50-80:/home/shubha/Desktop/python-docs-hello-world# git commit -am "update output"
[master 1a7d73b] update output
 1 file changed, 1 insertion(+), 1 deletion(-)
root@shubha-Lenovo-G50-80:/home/shubha/Desktop/python-docs-hello-world# git push azure master
Password for 'https://Userdemo15@azappnew05.scm.azurewebsites.net':
Counting objects: 3, done.
Delta compression using up to 4 threads.
Compressing objects: 100% (3/3) done.
Writing objects: 100% (3/3), 366 bytes | 0 bytes/s, done.
Total 3 (delta 0)
remote: Updating branch 'master'.
remote: Updating submodules.
remote: Preparing deployment for commit id '1a7d73be7d'.
remote: Generating deployment script.
remote: Running deployment command...
```

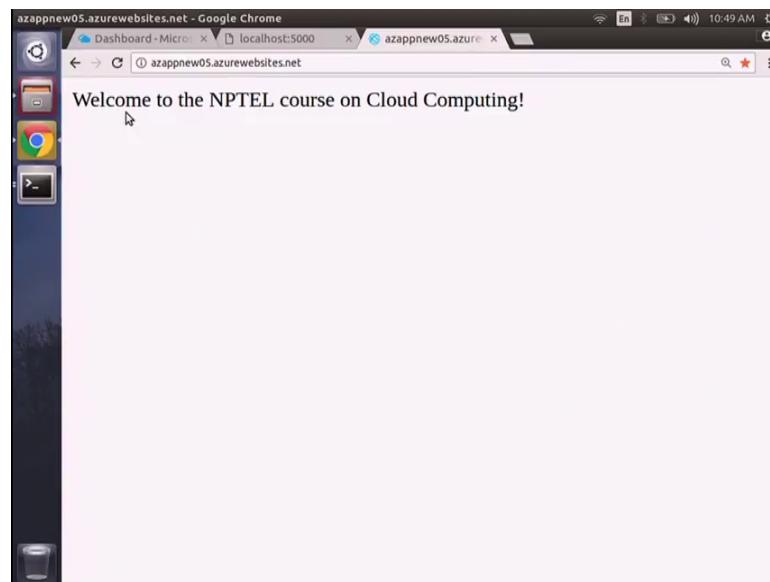
(Refer Slide Time: 16:24)



```
root@shubha-Lenovo-G50-80:/home/shubha/Desktop/python-docs-hello-world
remote: Preparing deployment for commit id '1a7d73be7d'.
remote: Generating deployment script...
remote: Running deployment command...
remote: Handling python deployment.
remote: KuduSync.NET from: 'D:\home\site\repository' to: 'D:\home\site\wwwroot'
remote: Copying file: 'main.py'
remote: Detected requirements.txt. You can skip Python specific steps with a .skipPythonDeployment file.
remote: Detecting Python runtime from site configuration
remote: Detected python 3.4
remote: Creating temporary virtual environment.
remote: Pip install requirements.
remote: Requirement already satisfied (use --upgrade to upgrade): Flask==0.12.1
in d:\home\site\wwwroot\env\lib\site-packages (from -r requirements.txt (line 1))
)
remote: Cleaning up...
remote: Overwriting web.config with web.3.4.config
remote: 1 file(s) copied.
remote: Finished successfully.
remote: Running post deployment command(s)...
remote: Deployment successful
To https://Userdemo15@azappnew05.scm.azurewebsites.net/azappnew05.git
 44e74fe..1a7d73b Master -> master
root@shubha-Lenovo-G50-80:/home/shubha/Desktop/python-docs-hello-world#
```

Now, you need to push the code changes to Microsoft azure, git push azure master, right, again I need to give the password of the deployment user. So, the app changes has been deployed successfully now we need to refresh this webpage.

(Refer Slide Time: 16:42)



So, this will give me the message welcome to NPTEL. Course on Cloud computing.

So, as we can see that what Shubhabrata has shown from that step by step procedure, he has used local git, right.

Yes.

For sinking with azure; so, you can develop your own web app in an host in those thing in azure, right. So, similarly you can develop other apps also. So, it is again, it is just to give a feel of how a commercial cloud work and as you can see; this is sort of things any which any type of pass type of platform like azure; we will have somewhat similar characteristics, right.

So, with these let us end our discussion today on a commercial cloud than Microsoft azure. We will continue with other things in the next lecture.

Thank you.

Thank you.

**Cloud Computing**  
**Prof. Soumya Kanti Ghosh**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**

**Lecture - 19**  
**Case Study: Google Cloud Platform (GCP)**

Hi. Let us start our discussion or continue our discussion on Cloud Computing. As we mentioned that there are several commercial cloud and open source cloud are available in providing services at various level like IaaS, PaaS or SaaS. So, one of the one of the popular such cloud we will discuss. So, again the major objective is to I just show you how a commercial cloud works and so that you can try yourself and see that have the flavor of the cloud right.

So, today we will discuss about Google cloud platform briefly we will discuss and we will also give you a short demo how to develop a app or host your web app into the global Google cloud platform which is very user friendly and easy to use. So, again there is no immediate there is no particular motivation in basically having working with some commercial cloud, but it is just to use, it as a use case or test case where you can practice and see that how things works. And we are using free account a demo account, so that you can also try at your end and see that how things works

So, as we understand Google as worldwide presence and their data centers are across the globe, and this Google cloud platform also if you see, they are in various regions right like North American region, UK region, and Asia and so on and so forth. So, they are distributed regions, every region have some zones and so that these are divided into geo graphically spread right. So, there are several services which has a global view, which has a zonal view and more infrastructure wise a view. So, keeping all this at the backbone, we will try to see that what Google cloud platform provides.

(Refer Slide Time: 02:47)

**What's Google Cloud Platform?**

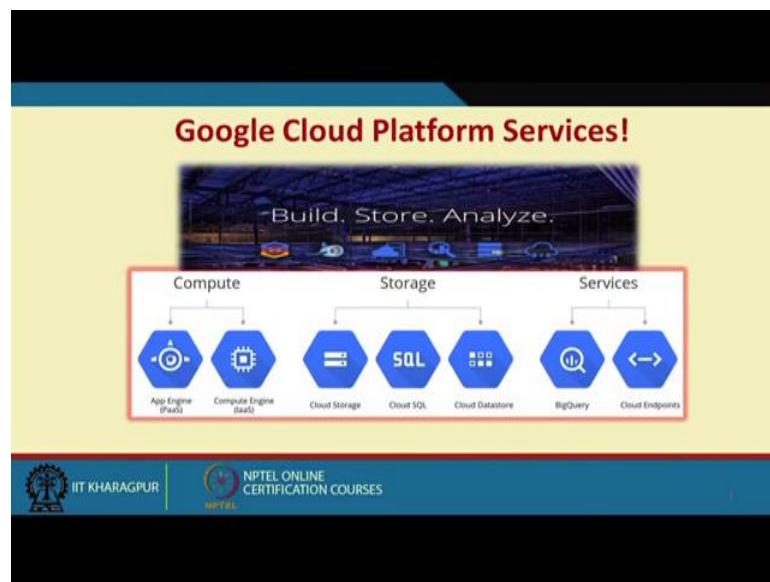
- **Google Cloud Platform** is a set of services that enables developers to **build, test and deploy** applications on Google's reliable infrastructure.
- **Google cloud platform** is a set of modular cloud-based services that allow you to create anything from simple websites to complex applications

Source: <https://cloud.google.com/>; <https://developers.google.com>

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, if you see like it say state of services what they provide is a set of services that enables developers to build, test, deploy application on Google's reliable infrastructure. So, we have taken these thing from their web resources, there one is cloud dot Google dot com and developers Google dot command related resources. So, what they claim that it say set of they provide larger set of services that enable developers to build, test, deploy application on Google infrastructure. Google cloud platform is a set of modular cloud base services that allow you to create anything from simple website to complex applications; that means, you can have a you can host your website or you can have a complex application running on the cloud.

(Refer Slide Time: 03:36)



So, if we look at little broad aspect, so it build, store, analyze, so this is the three motto. And in order to have, so we have Google have a computer services which is provided by Google app engine which is primarily a PaaS and compute engine which provides IaaS type of services other than that it has a storage services cloud storage cloud SQL and cloud data store. So, they are cloud storage services in Google and there are other services like query, cloud endpoint type of services. Rather if you look at their website the portal services are listed there and they are implemented or increasing or modified overtime right.

(Refer Slide Time: 04:36)



So, what they claim or what Google try to support is run your application, host your site on Google's infrastructure build on the same infrastructure that allow Google to return billions of search results in milliseconds that means they are basic infrastructure, they serve around 6 billion hours of YouTube video per month and some 425 million Gmail users. So, the same sort of infrastructure you can use when you are using thing. So, it is a globally connected over the network, highly redundant. So, it is fault tolerant in that respect and Google go on what they claim that go on doing innovation. So, you have innovative infrastructure in place, so that is what even one use Google platform what Google wants to provide them.

(Refer Slide Time: 05:35)

**Google Cloud Platform? (contd..)**

*Focus on your product*

Rapidly develop, deploy and iterate your applications without worrying about system administration. Google manages your application, database and storage servers so you don't have to.

- ✓ Managed services
- ✓ Developer Tools and SDKs
- ✓ Console and Administration

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, from the user point of view they are focus on their application or the product rapid development, deploy, iterate your applications without worrying about the system administration etcetera. So, all those things are taken care by the backend service provider, in this case Google. Google manages your application database storage server. So, you do not have to manage all those things so manage services right developers tools and SDKs are available, console and administration for management of the things. So, these are the things which are comes with you when we use Google platform.

(Refer Slide Time: 06:18)

**Google Cloud Platform? (contd..)**

**Mix and Match Services**

Virtual machines. Managed platform. Blob storage. Block storage. NoSQL datastore. MySQL database. Big Data analytics. Google Cloud Platform has all the services your application architecture needs.

- ✓ Compute
- ✓ Storage
- ✓ Services

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES



So, other aspects what they claim or what you can have is mix and match of services. So, you have virtual machine, manage platform, blob storage, block storage, NoSQL data store, MySQL databases, big data analytics, Google cloud platform has all the services your application architecture needs. So, it says you can have a number of services if you and you can have a mix and matches of the services to develop or launch your application right. So, compute storage and services are the core of the things and you can mix and match of several services to realize your particular application or particular objective of your particular application.

(Refer Slide Time: 07:10)

**Google Cloud Platform? (contd..)**

**Scale to millions of users**

Applications hosted on Cloud Platform can automatically scale up to handle the most demanding workloads and scale down when traffic subsides. You pay only for what you use.

**Scale-up:** Cloud Platform is designed to scale like Google's own products, even when you experience a huge traffic spike. Managed services such as App Engine or Cloud Datastore give you auto-scaling that enables your application to grow with your users.

**Scale-down:** Just as Cloud Platform allows you to scale-up, managed services also scale down. You don't pay for computing resources that you don't need.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

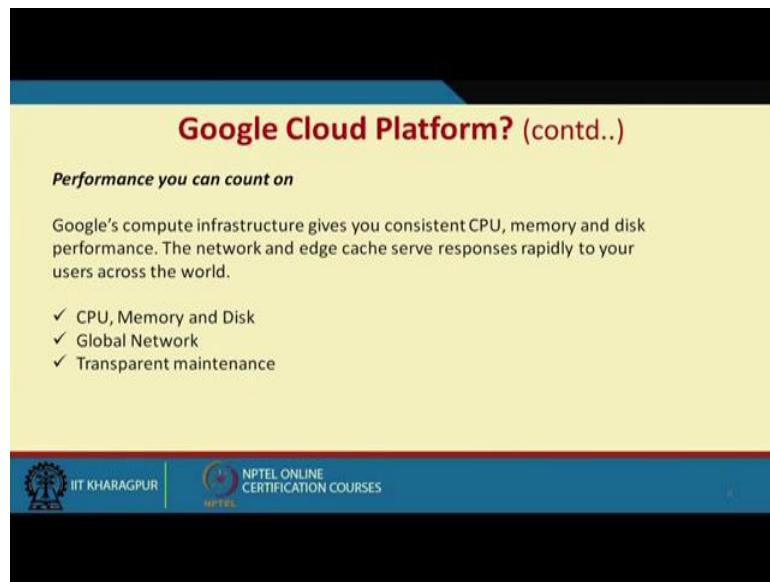


So, the other advantages what they offer is the scale to millions of users. As such Google has they have use user base and they have the infrastructure, which can support this user base and as you are using the same infrastructure as you will be used this same infrastructure. So, we can basically scale to this millions of users. So, applications hosted in cloud platform can automatically scale up to handle most demanding workloads and scale down when the traffic is updates. So, scale up, scale down is both possible. So, you pay only you use right pay as you go model.

Scale up cloud platform is designed to scale like your own products like if you use Google other products whenever the demand is high scale up and so forth right. Even when you experience a huge traffic spike then it actually what they propose or what they offer is that the type of things they are doing with their own products the same type of support or management they do for when you somebody launches something on the Google cloud platform. So, manage services searches Google app engine or cloud data store give you, auto-scaling that enables user grow with your user.

So, these are the manage service like GIE Google app engine or cloud data store which provides auto scale whenever the load is higher, it goes up or other way out. Scale down just as cloud platform as allows you to scale up, manage services also scale down right you do not pay for computing resources you do not. So, if you have lesser load. So, you can scale down, so that release of the or deactivating the computing resources are in place.

(Refer Slide Time: 09:09)



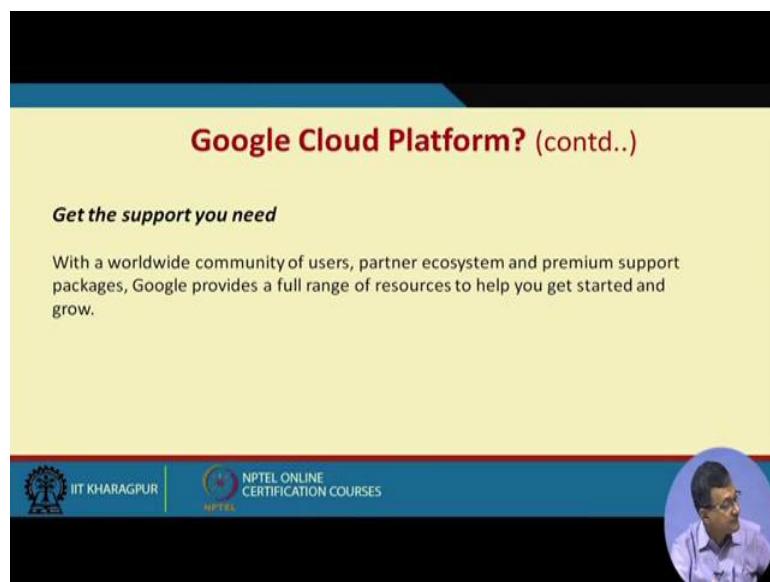
The slide has a black header and footer. The main content area is yellow. The title 'Google Cloud Platform? (contd..)' is in red at the top. Below it is the subtitle 'Performance you can count on' in bold. A paragraph explains Google's compute infrastructure provides consistent CPU, memory, and disk performance, with responses from the network and edge cache to users worldwide. A bulleted list follows:

- ✓ CPU, Memory and Disk
- ✓ Global Network
- ✓ Transparent maintenance

At the bottom, there are logos for IIT Kharagpur and NPTEL, followed by the text 'NPTEL ONLINE CERTIFICATION COURSES'. On the right side of the slide, there is a small circular video player showing a person speaking.

See also claim that or we can have a good performance or return on our investment or hiring these Google cloud platform. So, Google's compute infrastructure give you consistent CPU a memory and disk performance, so that is guarantee. So, that they can network and age case response is rapidly to your users across the world. So, if you are a business user or a you have intern that user across the world, so they also get the advantages of Google's scale up and type of services right. So, it can be in terms of CPU memory disk, it can be in terms of the network global network. It can be or maybe in transparent maintenance of the infrastructure or you know application.

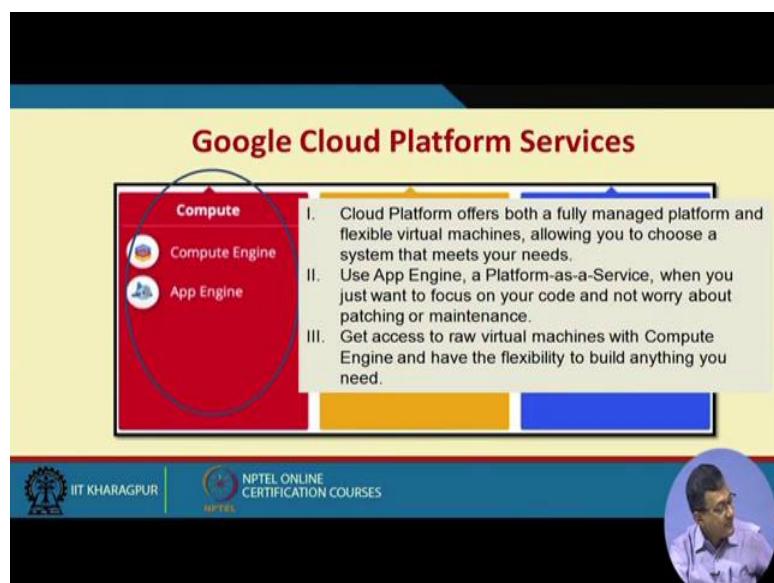
(Refer Slide Time: 10:09)



The slide has a black header and footer. The main content area is yellow. The title 'Google Cloud Platform? (contd..)' is in red at the top. Below it is the subtitle 'Get the support you need' in bold. A paragraph states that with a worldwide community of users, partner ecosystem and premium support packages, Google provides a full range of resources to help you get started and grow. At the bottom, there are logos for IIT Kharagpur and NPTEL, followed by the text 'NPTEL ONLINE CERTIFICATION COURSES'. On the right side of the slide, there is a small circular video player showing a person speaking.

And also you get support you need with a worldwide community users, partner ecosystem and premium support packages, Google provides full range of resources to help you to get start and go so that means, the Google the overall support come into play wherever you launch in the global cloud platform. So, these are more what we say means advantages Google claim they provide and it is if these are available to a customer or a user, so it is the headache of maintenance, manageability etcetera are is reduced to a large extent and you pay for the services you use and type of things.

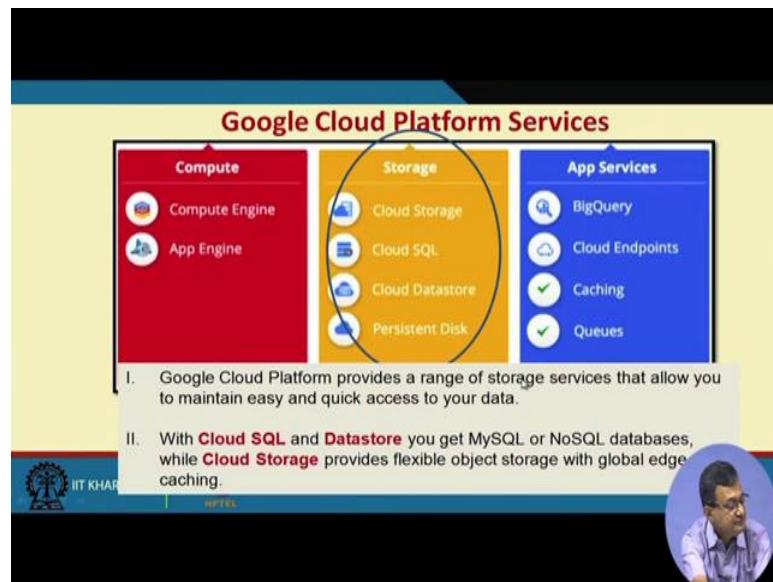
(Refer Slide Time: 11:02)



So, if we look at the Google cloud platform, so compute services one is compute engine, another is a app engine, one is primarily IaaS types of things and another is the PaaS type of stuff. So, cloud platform offers both a full manage platform with flexibility flexible virtual machine allow you to choose a system that you needs to that you needs, so that when a Google cloud platform it has a both a manage platform and a flexible machines, so that means, you can have a flexibility configure VM. Use app engine or past type of services when you just want to focus on your code and not worry about the patching on maintenance etcetera not about the infrastructure, so you require a development platform.

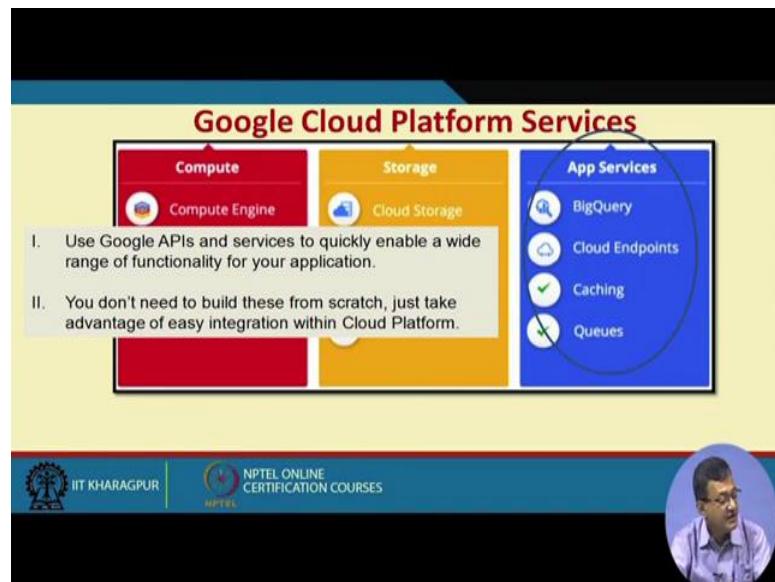
And get access to raw virtual machine with compute engine and have flexibility to build anything with this, so other than you can have your own VM and then you have flexibility of loading apps or other type of things in that virtual machine, and have a IaaS type of services out of it.

(Refer Slide Time: 12:24)



Secondly, another vertical is storage. So, it provides when variety of sense one is cloud storage, cloud SQL, cloud data store persistent disk. So, these are the different things which are offered by TCP or the Google cloud platform. It provides a range of services that allow you to maintain and easy quick access to your data right. So, it is a range of services with cloud SQL and data store you get mySQL and noSQL databases while cloud storage provides flexible object storage with globalized things, for with data store and SQL or SQL and you get the mySQL thing and another type noSQL type of databases. So, this is more of a base related service or cloud storage provide object storage with globalized thing that means, you can basically store and access across the globe.

(Refer Slide Time: 13:30)



And finally there are several app services which is like big query, cloud endpoints, catching and queues. So, these are app services with Google APIs and services to quickly you can use Google API services to quickly enable a wide range of functionality for your application right. So, you can build your application using this app services, one need not build these from scratch just take advantage of easy integration of Google platform, so that that is the app engine tries do. So, it is a bunch of APIs which can be leverage for your own developing your own application.

So, this is a quick brief overview of what are the different modules and vertical of Google is. So, if you, want more details you can basically go to their website and check that more how things works these are says to shows that these are the different verticals. But as I mentioned at the beginning of the thing our objective is to show some example cases on some commercial and open source cloud, so that you can try yourself and see that how things works, and what are the nitty-gritty of the things right.

(Refer Slide Time: 15:08)

**Google Cloud Platform Services – from User end!**

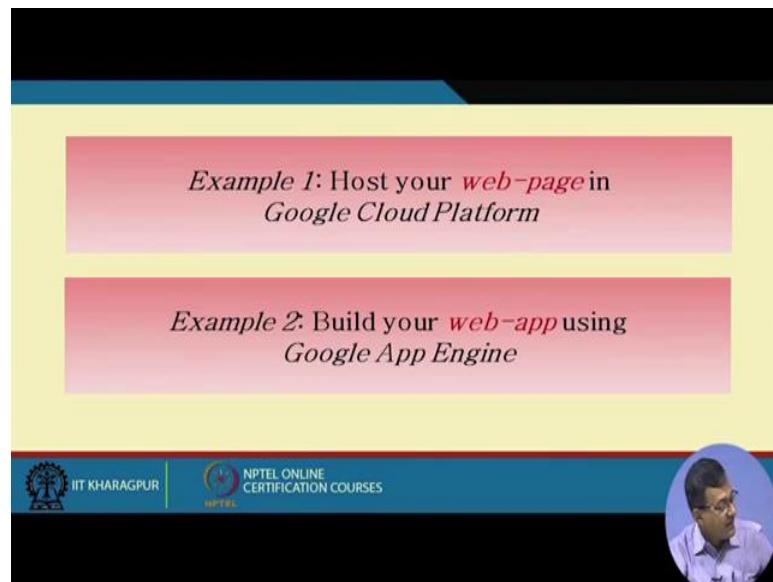
- Consider to migrate your web application to Google Cloud Platform for better performance using **GoogleAppEngine**.
- Your application should go wherever your users go: Scale your application using **GoogleCloudEndpoints**.
- Integrate Google's services into your Application using **GoogleAPIs**.

 IIT KHARAGPUR |  NPTEL ONLINE CERTIFICATION COURSES

So, like here also we will, so two applications to example cases now so before that. So, we if you look at the whole thing like it is a cloud platform services from the user end considered to migrate your application to Google cloud platform or for better performance using Google app engine where you can migrate your application web application to cloud platform or your applications should go wherever your users grow like your scale applications using Google cloud end points. And integrate Google services into your application using APIs right.

So, if you can integrate Google services in your applications which may be running locally or in cloud and using Google API. So, Google app engine, Google end points, Google API or some other things which plays important role.

(Refer Slide Time: 16:02)



And as we were discussing we will, so two example scenario with these Google cloud platform or GCP which to demonstrate that how a typical cloud works right, you can try yourself and have more complex example. These are two simple example, one is to host your web page in the Google cloud platform that is already you are having a web page design locally and you want to upload host is the web pages into Google cloud platform, sorry Google cloud platform or GCP, and use their storages for that. The second one is building web application using Google app engine. So, we will use these Google app engine services to build a web application. So, two simple applications, but I believe that it will help you to specially those who are new to this cloud thing to help you in developing small application and hands on experience with the commercial cloud fine.

So, we will we will continue with the application, then for this with me Shreya will join. She will show you how a Google like a Google webpage can be hosted or a webpage can be build using Google app engine.

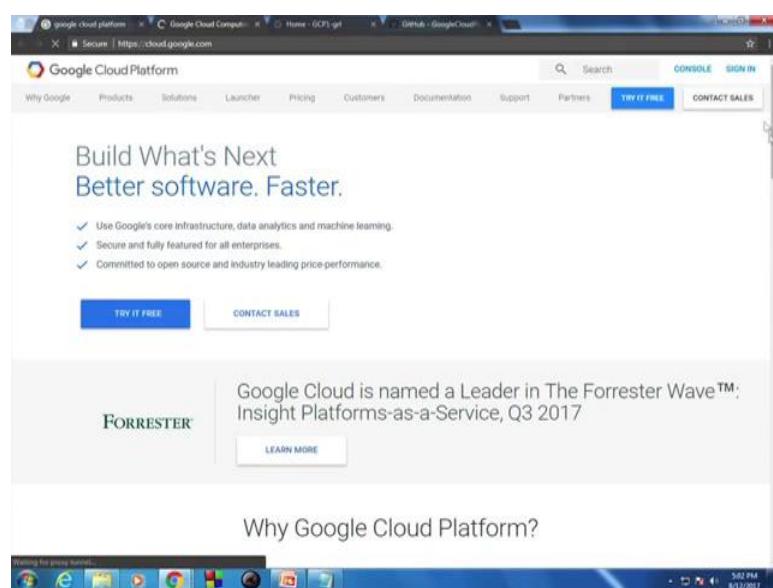
**Cloud Computing**  
**Prof. Soumya Kanti Ghosh**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**

**Lecture - 20**  
**Demo on Google Cloud Platform ( GCP )**

Hello. So as we are discussing about Google cloud platform now we show you two example scenario, one for hosting a web app in a Google cloud platform another building a app in a web app in a Google plat platform. And with me Shreya is there, Shreya will demonstrate the thing on a hands on. So, it will be easy for many of you to just do the same exercise on yourself and have a feel of how things work, right. I will now hand over to Shreya at. So, that she can continue with initially hosting a web app in the Google cloud platform right. So, I will give you to give it to her.

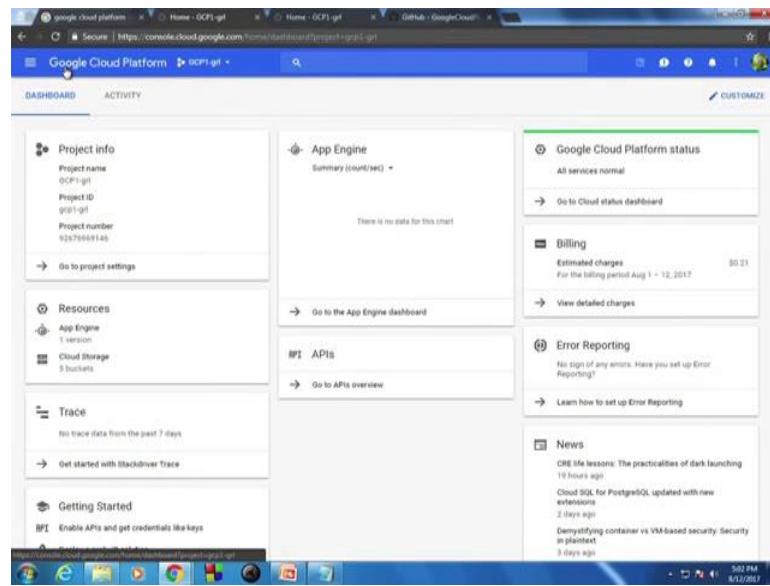
Thank you, sir. So, first we will go to the Google cloud platform console.

(Refer Slide Time: 01:18)



So, we need to login in the Google cloud platform to in order to host our web page. So, we will go to the console. And after login in the GCB account.

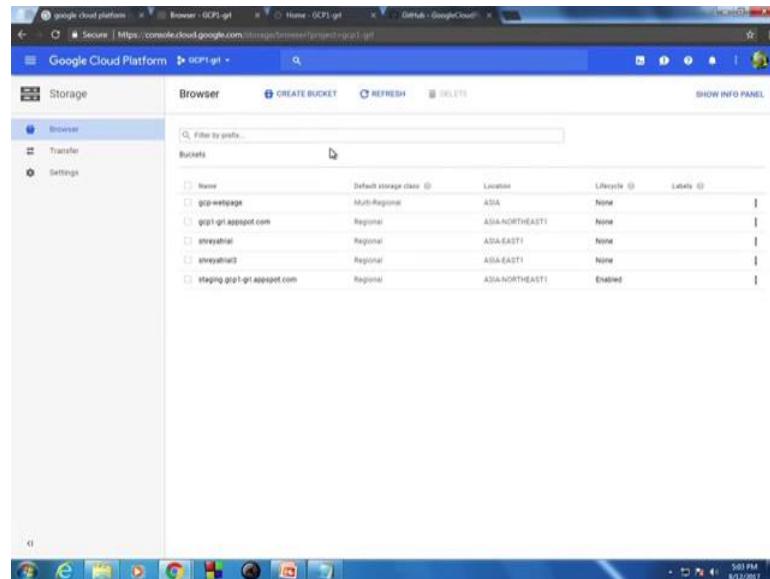
(Refer Slide Time: 01:29)



We will here we can show the project information if we have already created any projects. And all other resource information like app engine information or computing in engine information. And in order to host a static web page or website we need to create we need to configure the Google storage bucket. So, we will go to the storage option.

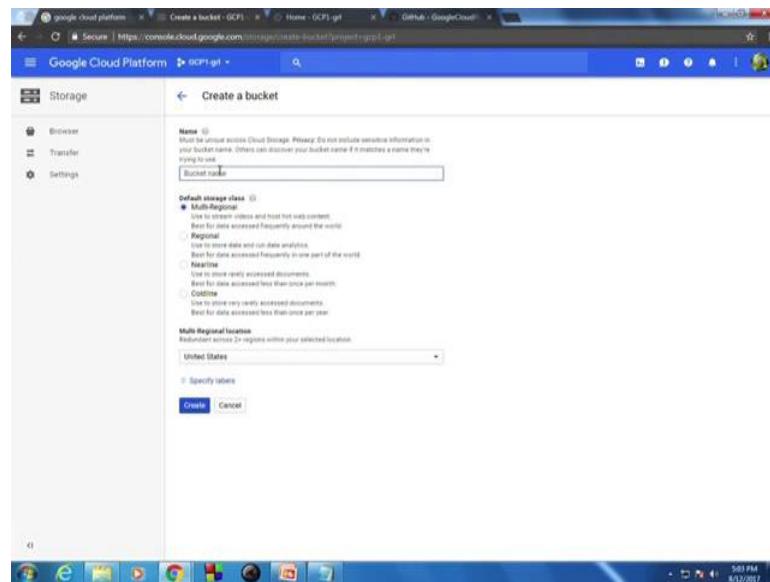
So, initially login to the console then create storage bucket right.

(Refer Slide Time: 02:07)



So, in the storage bucket under the browser tab we will see a some options like create bucket. So, I will now create a new storage bucket, so giving a name like NPTEL.

(Refer Slide Time: 02:18)



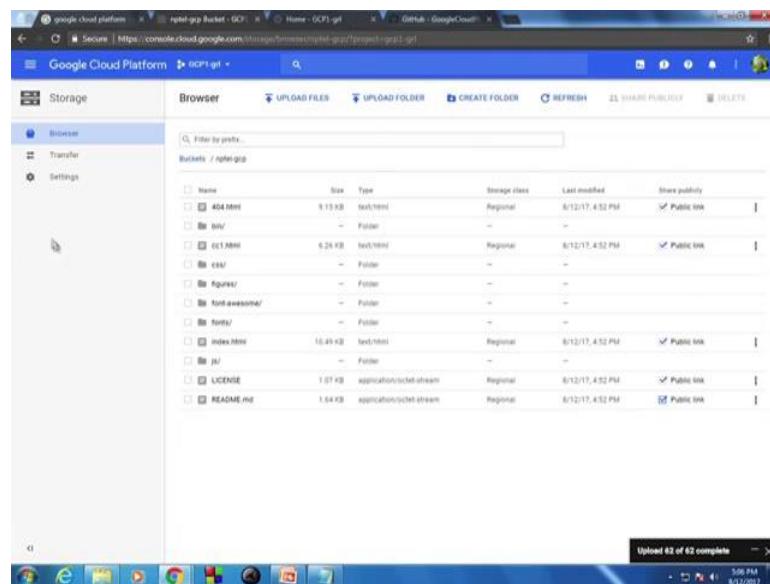
NPTL or Web page.

TCP web page.

Now, in NPTL you see something anyway you give.

Right.

(Refer Slide Time: 02:49)



Then we will choose the storage class that is where this particular web app will reside. So, I am choosing here regional, and Asia East1 and creating the bucket. So, the bucket has been created, but there are no objects in the bucket. So now, we will upload the files or the content of the websites in this particular buckets sir.

So, you are having locally already in the desktop itself go next. So, you are already having that locally the files created which you want to host on Google, GCB.

So the individual files have been uploaded no I will upload the folders.

So, you have created a locally a site and then you are uploading now in the GCP. So, you want to host it on Google cloud platform.

Yes.

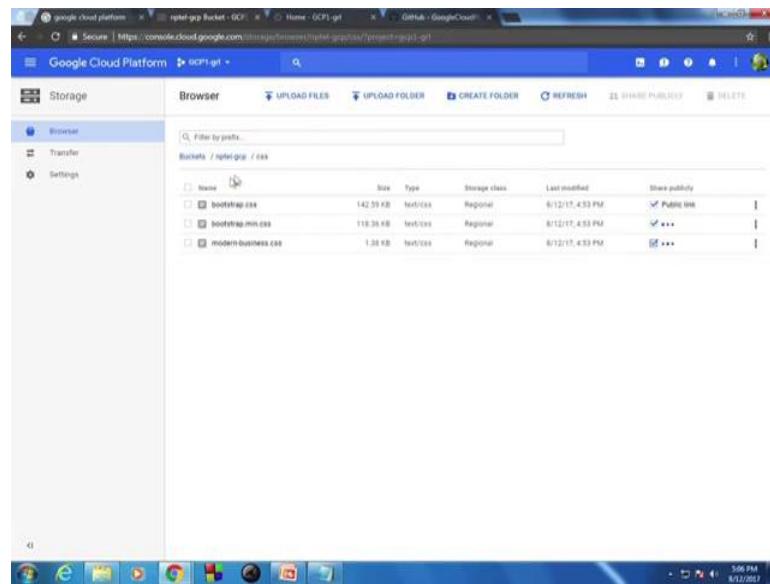
It is the first objective example is to host the thing into the Google cloud platform.

So, to host any website or any WebPages; there must be 2 files one is the 404 and not 4 0 4 not found file. And other is the main HTML file of the home page.

That index dot HTML.

Index dot HTML. So, all the files have been uploaded. Now we need to check whether the all the files are shared publicly or not. So, we should check the all the links here. And not only these files we need to check all the files inside this inside the folders also.

(Refer Slide Time: 05:06)



So, all the files are checked now. Now the web contents has been uploaded in the GCP.  
So, we need to go to the homepage of our website.

(Refer Slide Time: 06:19)



So, just click the public link. So, from the URL you can see that this is the project name or the bucket name what we have created. And this is the homepage HTML file.

(Refer Slide Time: 06:33)

The screenshot shows a web browser window with the URL <https://storage.googleapis.com/npTEL-gr1/index.html>. The page has a dark blue header with the text "Hi there!" and navigation links for "Home" and "Summary". Below the header is a large, stylized graphic of a globe with a network of lines and dots, representing cloud computing. Overlaid on the globe are three icons: a house, a lightning bolt, and a shopping cart. The text "Ubiquitous, On-demand access to configurable computing resources" is displayed in the center of the globe. The main content area is titled "Welcome to Cloud Computing NPTEL Course!". It contains three columns: "About this Course", "Course PRE-REQUISITES & Suggested Reading", and "Course Instructor & Certification". The "About this Course" section describes the course's purpose and prerequisites. The "Course PRE-REQUISITES & Suggested Reading" section lists books like "Cloud Computing: Principles and Paradigms" and "Enterprise Cloud Computing - Technology". The "Course Instructor & Certification" section provides details about the professor, exam schedule, and certification.

And this particular side has been upload has been hosted from the storage dot googleapis dot com.

(Refer Slide Time: 06:43)

The screenshot shows a web browser window with the URL <https://storage.googleapis.com/npTEL-gr1/1.html>. The page has a dark blue header with the text "Hi there!" and navigation links for "Home" and "Summary". Below the header is a large, stylized graphic of a globe with a network of lines and dots, representing cloud computing. Overlaid on the globe are various icons representing different cloud computing services like mobile phones, tablets, and social media. The main content area is titled "Summary of the Lectures". It features two main sections: "Cloud Computing - Overview" and "Cloud Computing - Economics". The "Cloud Computing - Overview" section includes a diagram showing "Cloud computing" at the center, surrounded by "Off-line access", "Online storage", "Platforms", "Shared calendars", "Online office", "Online resources", "3rd Party integration", "Online collaboration", and "Outsource processes". The "Cloud Computing - Economics" section includes a diagram of a cloud connected to various devices like a laptop, smartphone, and tablet, with icons representing data storage and processing.

So, you can navigate to any other web page like cc1, HTML and all. Also the external links can external web pages can we linked from this website as well.

(Refer Slide Time: 06:53)

The screenshot shows the NPTEL Cloud Computing course page. At the top, there's a navigation bar with links for Announcements, Course, Forum, and Mentor. Below the header, there's a section titled 'Cloud Computing' with a sub-section 'ABOUT THE COURSE'. It describes cloud computing as a service-oriented platform and its various benefits. To the right of the text is a video thumbnail of Prof. Soumya K. Ghosh speaking at a desk. Below the video, there's a section titled 'INTENDED AUDIENCE' listing CSE, ECE, and EE. Under 'PRE-REQUISITES', it lists Computer Architecture and Organization, Networking, and IT industries. A 'COURSE INSTRUCTOR' section shows a small thumbnail of the professor. The bottom of the page has a Windows taskbar with icons for various applications.

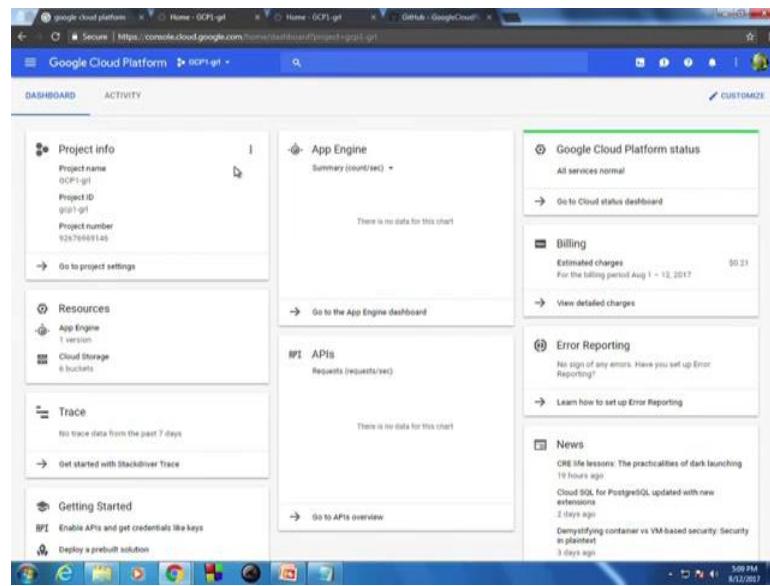
So, this is a hosting part.

(Refer Slide Time: 07:05)

The screenshot shows the Cloud Computing 2017 course page. At the top, there's a 'Hi there!' greeting and a 'Home' link. Below the header, there's a section titled 'Course Layout' which lists the course weeks: Week 1 (Introduction to Cloud Computing), Week 2 (Cloud Computing Architecture), Week 3 (Service Management in Cloud Computing), Week 4 (Data Management in Cloud Computing), Week 5 (Resource Management in Cloud), Week 6 (Cloud Security), Week 7 (Types of Clouds and Commercial Clouds, Cloud Simulator), and Week 8 (Research trend in Cloud Computing, Fog Computing). To the right of the layout list is a central graphic titled 'What is Cloud Computing' surrounded by five bubbles: 'Multi-tenant solution provided by vendor', 'Automated backups, uptime, SLA, maintenance', 'Automated upgrades', 'Elastic, pay as you go – scale up or down', and 'Modern web based integration'. At the bottom of the page, there's a 'Welcome to the Course' message and a copyright notice: 'Copyright © Shreyas2017'. The bottom of the page also features a Windows taskbar.

Hosting part: so hosting is pretty straight forward. So, you need to have a login and followed by that locally create the file, upload the file.

(Refer Slide Time: 07:28)



And that is that is it that is hosted. So, it is primarily using the storage of the use of Google that is mostly the storage services, and you have to enable that public accessibility to the files wherever you are required to you. So, the next example what we will be showing is building app within that Google app engine right.

So, using Google app engine services, ok.

Sir, so.

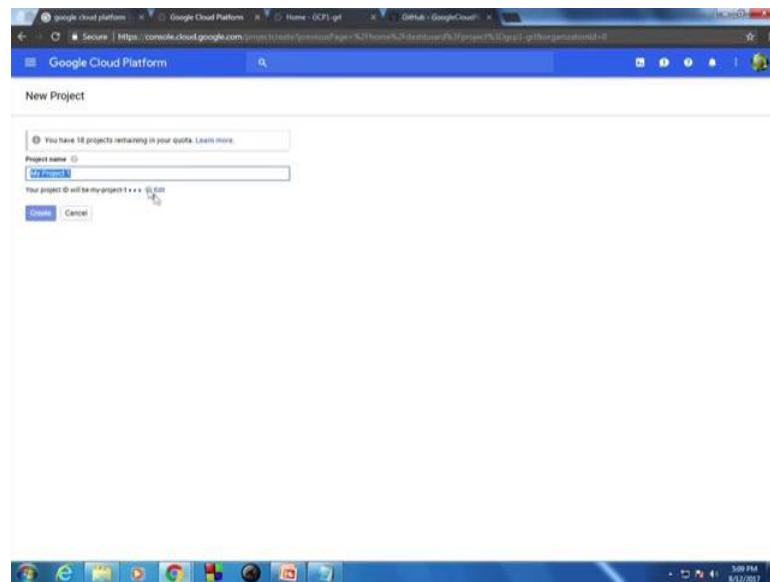
So, what this app will do?

This is a simple web application that will just print a message in the web page.

Ok.

So, build a python app we need to create a new project from the Google cloud perform dashboard. So, I create a new project here.

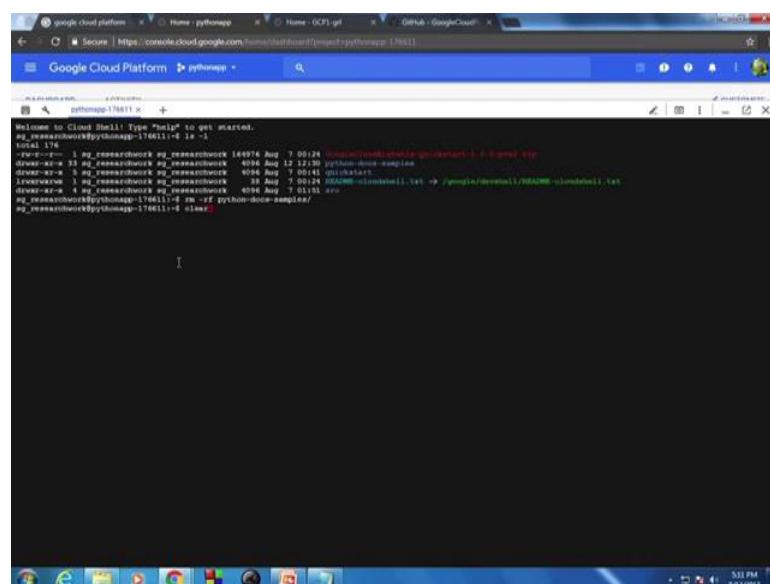
(Refer Slide Time: 08:08)



Python app.

So, here we can see the one globally unique identifier will be created. So now, the python app have the project has been created. So, I will go to the project. So, information about the project will be listed here, now I will activate tar Google cloud shell.

(Refer Slide Time: 09:10)

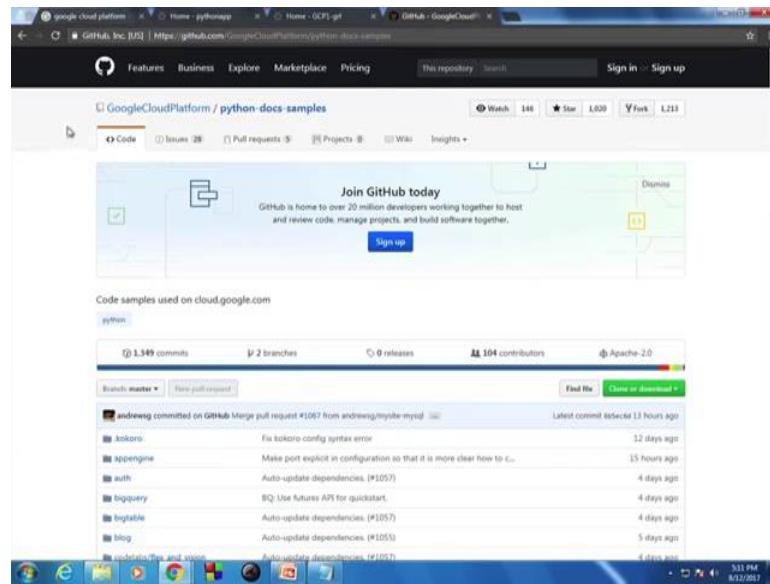


So, basically it works like any terminal in a Linux machine. So, command prompt has come. So, we can execute any command here.

Executing.

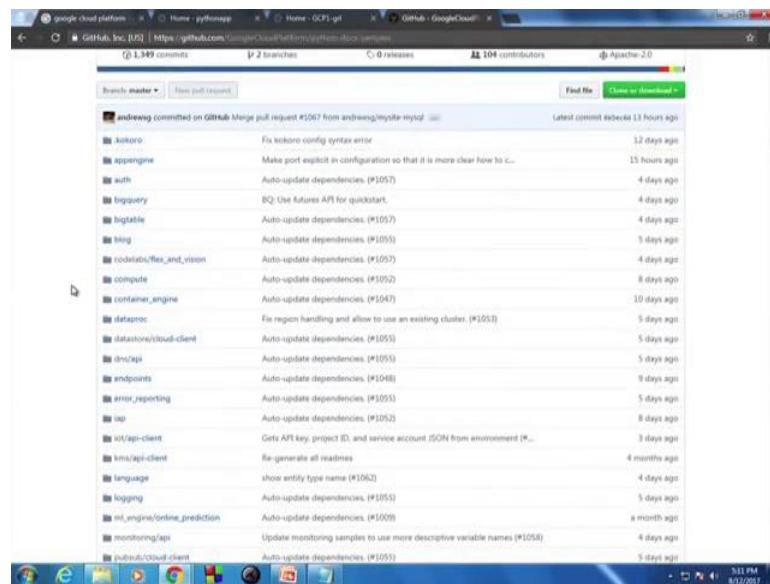
Yeah.

(Refer Slide Time: 10:04)



So, now I will clone or download the Google one example application, from this GIT hub repository. So, here you can see in here are a number of application has been listed.

(Refer Slide Time: 10:14)

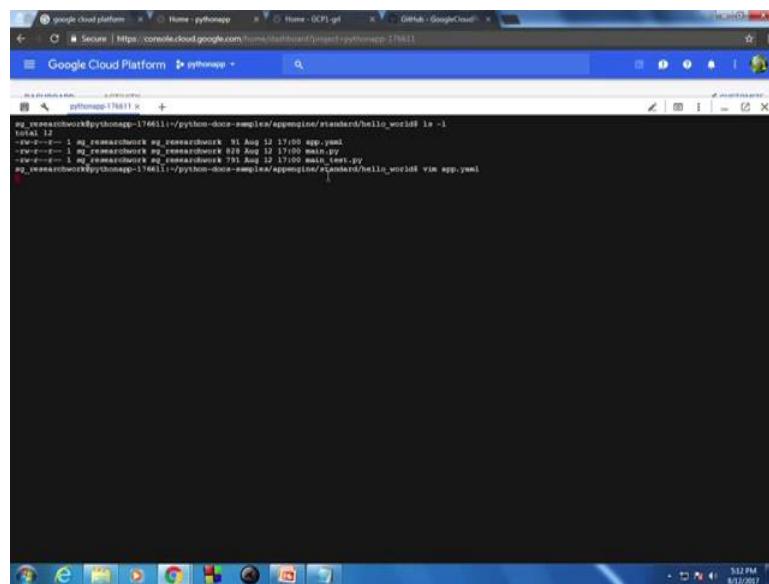


So, we can just download one or fetch the information, and then can create on your own. So, I will face all the application in my Google cloud shell machine. So, the cloning has

been completed. So that means, the all the all the contents from this repository has been downloaded in my local Google cloud shell machine.

So, I will navigate to the directory. So, here 2 main files are there one is the appear mail that is the configuration file of the application.

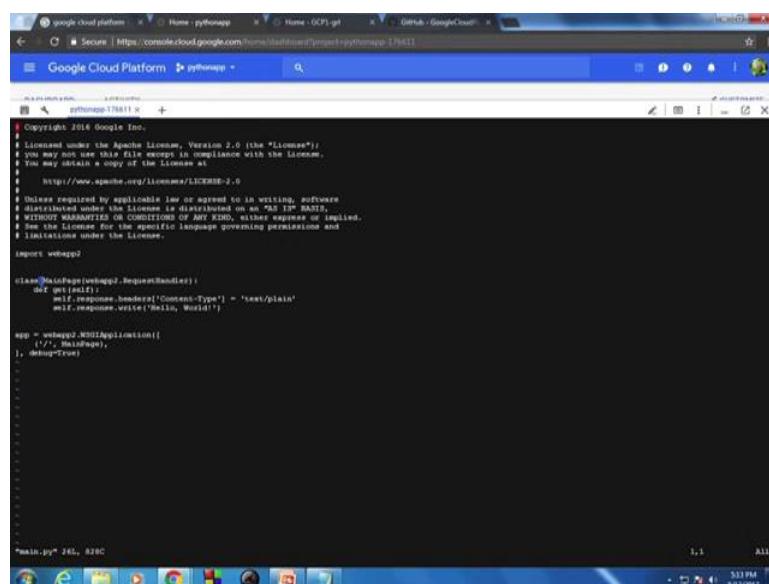
(Refer Slide Time: 11:12)



```
ls -l
total 12
drwxr-xr-x 1 ag_researchwork ag_researchwork 91 Aug 32 17:00 app.yaml
drwxr-xr-x 1 ag_researchwork ag_researchwork 620 Aug 32 17:00 main.py
drwxr-xr-x 1 ag_researchwork ag_researchwork 4096 Aug 32 17:00 static
ag_researchwork@pythonapp-176611:~/python-docs-samples/appengine/standard/hello_world$ vim app.yaml
```

Another is the main python file. So, if you open the files. So, this is the main configuration files it tells that the run time required for this application is python 2.7 environment.

(Refer Slide Time: 11:28)



```
Copyright 2014 Google Inc.

Licensed under the Apache License, Version 2.0 (the "License");
you may not use this file except in compliance with the License.
You may obtain a copy of the License at

 http://www.apache.org/licenses/LICENSE-2.0

Unless required by applicable law or agreed to in writing, software
distributed under the License is distributed on an "AS IS" BASIS,
WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
See the License for the specific language governing permissions and
limitations under the license.

import webapp2

class MainPage(webapp2.RequestHandler):
 def get(self):
 self.response.headers['Content-Type'] = 'text/plain'
 self.response.write('Hello, World!')

app = webapp2.WSGIApplication([
 ('/', MainPage),
], debug=True)
```

And this is a traits application that mean this particular application can handle a number of simultaneous request from any URL's. And any URL that matches with this regular expression can be will be handled by the main app file. And in the main dot py file it is a develop on the flux web development frame. So, in the class file you can see a simple message hello world has been written.

(Refer Slide Time: 12:34)

The screenshot shows a Windows desktop environment. In the foreground, a terminal window titled "pythonapp-17661" displays log output from a Python application. The logs show various INFO and DEBUG messages related to the application's startup, configuration, and search indexing. One notable message is "INFO 2017-08-12 11:30:15,418 app\_server.py[118] Starting module 'default' running at: http://0.0.0.0:8080". Another message indicates "INFO 2017-08-12 11:30:15,418 app\_server.py[118] Starting admin server at: http://0.0.0.0:8080". The logs also mention "INFO 2017-08-12 11:30:15,418 app\_server.py[118] Starting web server at: http://0.0.0.0:49897". The application appears to be a Google App Engine project named "pythonapp-17661".

In the background, a web browser window titled "Google Cloud Platform > pythonapp" shows the URL <https://console.cloud.google.com/home/dashboard/projects/pythonapp-17661>. The page displays the "Hello, world!" message, indicating that the application is running correctly.

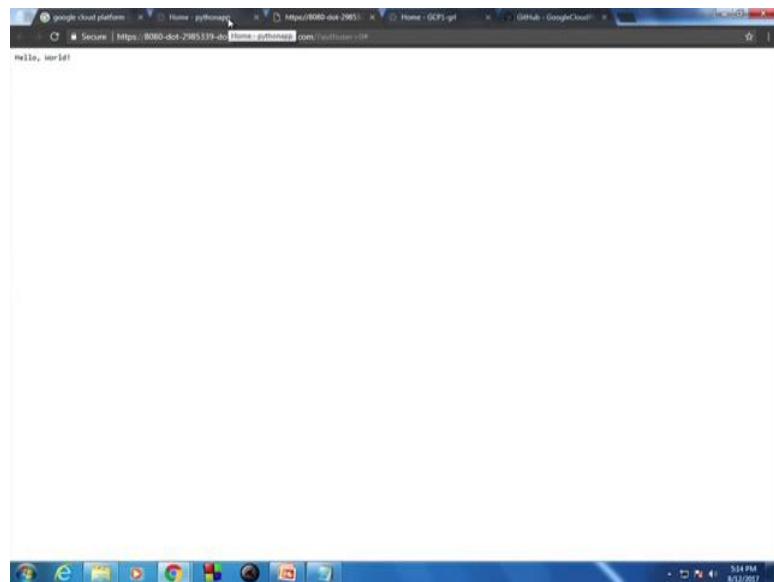
So, that you can change also, right. Hello, hello.

Yes that can change.

Yeah, change it.

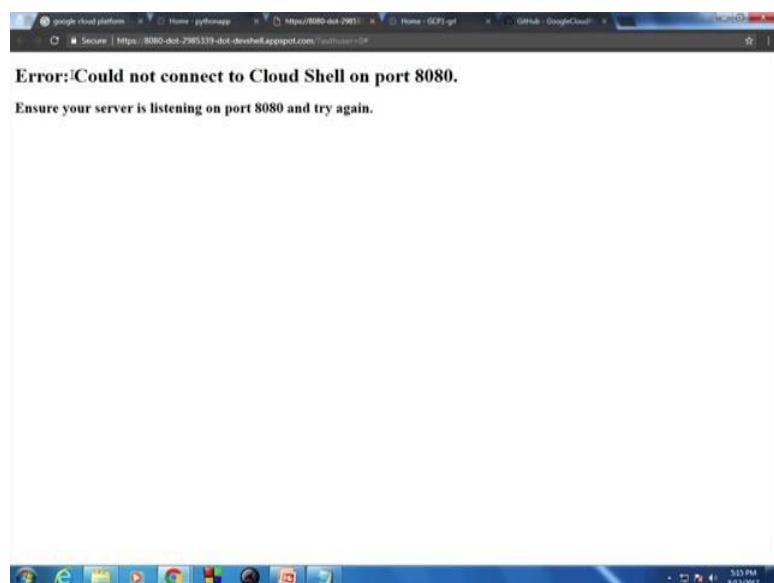
So now, I will start the development server. So, it is it the development server has been started in this particular link. So, if you go to the web preview. You can see that is the hello word message has been printed here. And at any time you can also shut down this development server.

(Refer Slide Time: 13:23)



Then it will show some connection error because the cloud shell has shut down the particular server there. So now, I will change the application.

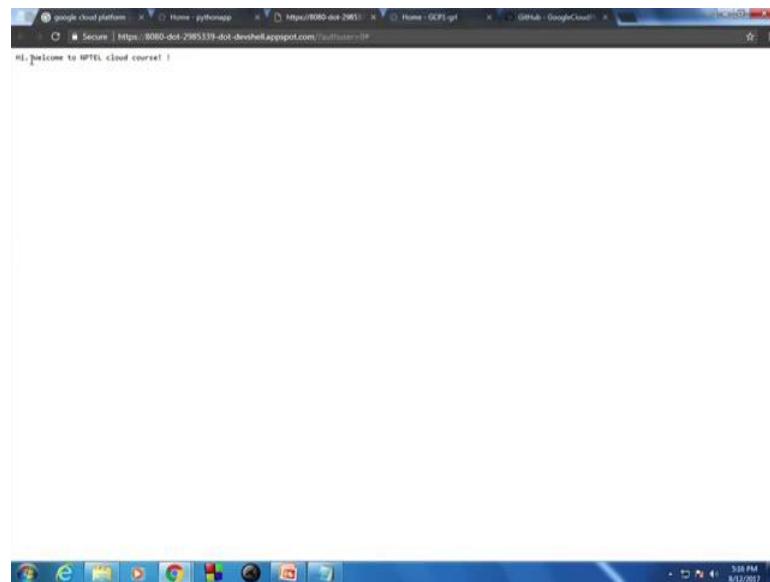
(Refer Slide Time: 13:46)



(Refer Time: 13:56)

Now again we need to start the development server here and we just check it.

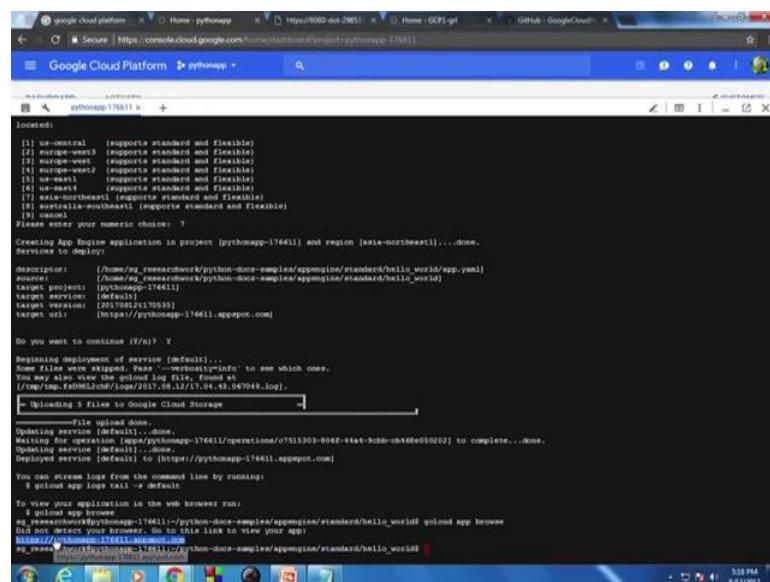
(Refer Slide Time: 14:48)



(Refer Time: 14:49)

So, the message has been changed here. So, till now we are developing in a local development server, but the files are not yet uploaded in the Google app engine. So now, we will deploy once we have done with the deployment and modification and all. We need to deploy the application in the Google app engine. So, we will run the command gcloud app deploy. And then it will then ask for the location where I want my app engine will be located. So, I am giving it 7.

(Refer Slide Time: 16:20)



So now the files are being uploaded in the Google cloud storage, and now that deployment has been completed. So, we can view our application by using gcloud app brows and it will give me the URL. So, from the URL you can see that this is the unique identifier of our project. And the particular application has been launched from Google app engine. And we can exit normally and after the project has been deployed. We also can shut down the project.

So, shut down it will basically close that, close.

Tell if we open the same project later?

Yeah, So, it is basically clearing of the thing.

So, as you see that we what Shreya demonstrated to simple example scenario in google cloud platform. So, the example other rather than the example the procedure is more important so that you can try out from simple to complex things. So, one of the example was locally developed a web app is see uploaded in the Google cloud platform and which can be accessible rather from it other from anywhere and the next app is basically next what she did is build a Google a web app using Google app engine.

So, you can build your own app on the things. So, this is related to web app you can develop other applications it is with those. And there are if you look at there are several services which are provided by Google for that matter any commands in cloud or any cloud if you use there are the different services which can one can leverage on to develop different webs. So, so that over all what we tried to show that how a thing works so what I try you can try this out and see that how this sort of clouds cloud works, and what is what is operational aspects of the cloud from the user perspective.

Thank you.

**Cloud Computing**  
**Prof. Soumya Kanti Ghosh**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**

**Lecture - 21**  
**SLA - Tutorial**

Hi. Today we will have a small tutorial on SLA it is in continuation with whatever we discussed on SLA. So, SLA as we understand for so service level agreement, and it is important for any cloud service provider and cloud service consumer to have an agreement to execute this either consume or provide this service. This as we understand that when we are using cloud computing we are basically leveraging on a third party services which are the service providers, and you the consumer things are hosted. So, it is both way to we should sign up or we should have a SLA.

So, unfortunately there is as such there is no standard or there is a rather I should I should say that standard state evolving to have a standard SLA across the different services that, one of the major reason, maybe there are different type of services which are provided by different providers they are at different category of consumer we who require services at different scale.

So, never the less there are some broad guideline by which these SLA's are governed. So, whenever you take a service from any service provider let it be commercial any of the commercial cloud like say microsirisio or Google cloud, or IBM cloud or Amazon cloud or any cloud. So, you need to agree on some of the some agreement. So, what we have seen in our SLA talk that how this agreements are means how this agreements can be formed or what will be the basic underlining infrastructure for that.

So, for that my matter we have talked about SLO, KPI's and so on so forth, which allows us to build up this SLA. So, in some cases this is some of the metrics which will be there in some cases it is policy driven right. Like where your data through reside what should be the backup policy and so on and so forth, are more policy driven where as some of the agreements are more on parameter based, like what is the see up time or CPU uses or disc uses these are some of the things which are metrics.

So, what we will do to we will try to look at one or 2 problem, before that we will see that how different parameters are considered in different type of SLA's right. In different type of commercial life, we have taken this from again from internet resources primarily from commercial providers like Azure, IBM, Google and Amazon and others.

(Refer Slide Time: 03:37)

**What is Service Level Agreement?**

- A formal contract between a Service Provider (SP) and a Service Consumer (SC)
- SLA: foundation of the consumer's trust in the provider
- Purpose : to define a formal basis for performance and availability the SP guarantees to deliver
- SLA contains Service Level Objectives (SLOs)
  - Objectively measurable conditions for the service
  - SLA & SLO: basis of selection of cloud provider

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, it the idea is to say that what how they frame they are things in a in a way. So, before looking at those things those stuff we just see as we as we have discussed earlier this slide, like it is a formal agreement between contract between the provider and consumer. Foundation of consumer trust on provider, and sometimes visa-vise that how the providers wants to have this consumer on the things purpose to define a formal basis for performance and availability of service providing provider guarantee is to deliver. And as we have talked about SLO's like objectively miserable condition for services and SLA, SLO basic for the selection cloud provider this way seen just I kept the one slide. So that things will be there.

(Refer Slide Time: 04:21)

**Problem-1**

Cloud SLA: Suppose a cloud guarantees service availability for 99% of time. Let a third party application runs in the cloud for 12 hours/day. At the end of one month, it was found that total outage is 10.75 hrs.

Find out whether the provider has violated the initial availability guarantee.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, we will discuss this two problem and then try to look at some of the aspects of how somehow this commercials cloud another things and clouds look at, the how they define. So, like a let us have a simple problem like suppose a cloud guarantee service availability of 99 percent of up time right. Late third party application runs in the cloud for 12 hours a day at the end of one month it was found that there is a outage often 0.75 hours. Find out whether the provider has violated the initial availability guarantee right; so very straight forward.

So, it gives in the SLA as 99 percent of time, late third party run say cloud for 12 hours a day end of a month it is 10.75 and you want to find out whether the provider has violated the initial availability guarantee right. So, if we look at.

(Refer Slide Time: 05:33)

Total time for which the application to run (in a month)  
 $= 12 \times 30 = 360$  hrs.

Outage time = 10.75 hrs.  
Therefore, Service duration =  $(360 - 10.75) = 349.25$  hrs.  
% availability =  $(1 - 10.75/349.25) \times 100 = 96.92\%$ .  
Initial Service guarantee = 99%.  
Final Service availability < Initial Service guarantee  
 $\Rightarrow$  CSP has violated the SLA.

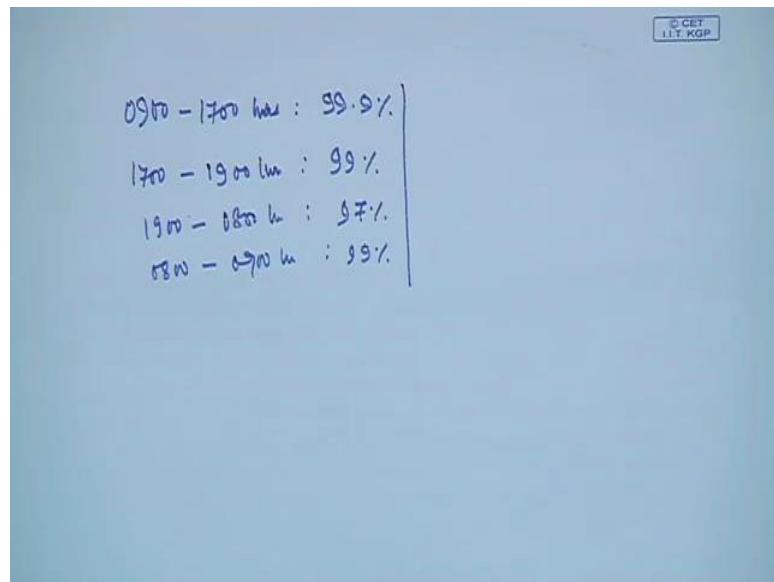
So, that is a problem. So, the total time for which the application to run in a month is equal to 12 cross 33, 360 hours right.

Now, what we say outage time is 10.75 hours, this has been declared this has been given. So, therefore, service duration equal to 360 minus 10.75 hours. So, percentage availability equal to 1 minus 10.75 by 349.25 into 100 which is 96.92, right. So, this much percentage availability will be there. This you can you can straight forward calculate. So, what it as it was their, initial service guarantee was 99 percent? So, has hence as final service guarantee. So, final service availability is less than initial service guarantee. So, what we can conclude that the service cloud service provider CSP if we say has violated the SLA ok.

So, it is a very straight forward simple arithmetic right, but what we see that if we can somehow measure this type of things, I can I can as per as the availability is concerned, we can basically calculate weather this SLA violations are there. If there is a set of SLA's which need to be looked into for everything every component, we can have this sort of simple calculation and or in some cases it may be little complex, when you want to do some statistical analysis to find out something. And then you can say that this is weather the SLA violation is there or SLA has been honored or not right. So, this way we can calculate this whether this any SLA is satisfied or not right ok.

So, this is this is pretty straight forward, but in reality it may not be that straight forward, what I can have different type of availability at different point of time, for that for that matter I can say that say if I considered a commercial say for example, a banking organization.

(Refer Slide Time: 09:59)



So, the a SLA's I can say that during my peak hours like I say 9 to 1700 hours, I require a availability of 99.9 percent. Whereas, in a off peak hours like 700 to say I can divide them into different scale, I can say that 700 to 1700 to this hours I can have 99 percent whereas, 1900 to next day 0800 hours I can still bring down to say 97 percent, and 0800 to 0900 hours I can say it is something again 99 percent. Now what I mean to say this availability requirement me also vary over time right, based on your business requirements right.

So, based on your requirement things will be like a institute like us I can I say, that if my lab if our labs are running between 2 to 5 or say morning in the morning say 2 to 6 and morning since an 8 to 12. So, during those lab hour I require a high percentage of availability. However, during the off peak hours or evening hours I may I may require much reuse thing. Because more you guarantee the services more you pay for it right. So, that is require so, there are there may be more complex calculation to look at, right.

(Refer Slide Time: 11:46)

**Problem-2**

Consider a scenario where a company X wants to use a cloud service from a provider P. The service level agreement (SLA) guarantees negotiated between the two parties prior to initiating business are as follows:

- Availability guarantee: 99.95% time over the service period
- Service period: 30 days
- Maximum service hours per day: 12 hours
- Cost: \$50 per day

Service credits are awarded to customers if availability guarantees are not satisfied. Monthly connectivity uptime service level are given as:

Monthly Uptime Percentage	Service Credit
<99.9%	10%
<99%	25%

However, in reality it was found that over the service period, the cloud service suffered five outages of durations: 5 hrs, 30 mins, 1 hr 30 mins, 15 mins, and 2 hrs 25 mins, each on different days, due to which normal service guarantees were violated. If SLA negotiations are honored, compute the effective cost payable towards buying the cloud service.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, similarly we can look at another problem, which is a little bit extension of the other of the previous one. So, we consider a scenario where a company x, a service provider x sorry a company x, want to use cloud service from a provider p. So, there is a company x which wants to use a provider p like say IIT Kharagpur have to correct once more cloud service from a external one or any anything. The service level agreement guarantees negotiation negotiated between the 2 parties prior to initiating the business are as follows.

So, before that the service level guarantees are like this like availability guarantee is 99.95 percent time over the service period. So, the service period it should be 99.95 percent time. Availability period is 30 days maximum service hour per day is 12 hours, and cost say 50 dollar a day right. So, this is this is the type of agreement or type of requirement and that formal requirement which has been agreed upon with within the service provider and the service provider. So, availability 99.95 percent, service period 30 day maximum service hours per day is 12 hours, and cost is 50 dollar a day right. So, serve with credits are awarded to the consumer if availability guarantees are not satisfied right.

So, there is another part like if you if the provider fails to provide fails to provide service at the guaranteed level for which it has been agreed upon and the consumer is charging, then thus there has to pay the penalty right. So, penalty can be in terms of money detents,

or the penalty can be in terms of giving some extra compute hour or data or whatever way like

So, in this case availability can set monthly connectivity up time service level given as like a monthly up time percentage is less than 99.95 percent, but more than 99 percent greater than equal to 99 percent then the service get ride is 10 percent right. Whereas, if it is less than 99 percent then the service credit should be 25 percent right; however, in reality it was found that over the service period the cloud server support 5 outages right. During for these following durations, like one is 5 hours 30 minutes, one is 1 hour 30 minutes, one is 15 minutes, 2 hours 25 minutes each on different day. So, this due to which normal service guarantees where violated right; if unless if SLA negotiation are honored we need to compute the effective cost payable towards buying this cloud services right. So, this is this we need to check that how much, effectively need to be paid by the consumer to for this cloud services right.

So, that is fine. So, again just to quickly repeat. So, there are some of the guarantees are there availability 99.95 percent service period 30 days, maximum things 12 hours 50 dollar, and there are some penalty for not providing the services less than 99.95 percent, but greater than equal to 99 percent is 10 percent and less than 99 percent is to 25 percent. And there are 5 outages 5 hours 30 minutes, 1 hour 30 minutes, 15 minutes and 2 hour 25 minutes. I mean to find out that the effective cost payable towards the buying the cloud services. So, this we have to work on again not a difficult problem, but it gives us a idea the how things works.

(Refer Slide Time: 15:56)

© CET  
I.I.T KGP

Service period duration = 30 days  
 Service duration per day = 12 hrs.  
 Total service uptime =  $(12 \times 30)$  hrs = 360 hrs.  
 Unit: \$50 per day.  
 Total cost (at the time of service negotiation) =  $\$(50 \times 30) =$   
\$1500  
 Total downtime =  $(5 \text{ hrs} + 30 \text{ mins} + 1 \text{ hr } 30 \text{ mins} + 15 \text{ mins} + 2 \text{ hrs } 25 \text{ mins})$   
= 9 hrs 25 mins.  
 Service Availability =  $1 - \frac{\text{downtime}}{\text{uptime}} = \left(1 - \frac{9 \text{ hrs. } 25 \text{ mins.}}{360 \text{ hrs.}}\right) \times 100$   
= 97.385%.

So, service period duration is 30 days right, 12 hours. So, total so, there for we have total so much hours, or 360 hours cost what we have seen 50 US dollar per day. So, total cost so this is at the time of at the time of service negotiation is dollar 50 cross 30 or this fine. So, these are the facts what we have given 30 days service duration per day is 12 hours as we are using, total service up time is expected this one 50 dollar this 50 per day is the cost and total cost at the time of service negotiation is, 15 to 30, 50, 1500 dollar, that is the thing.

Now, total service total service down time is 5 hours plus 30 minutes plus, 1 hour 30 minute plus 15 minute, plus 2 hours 25 minute right. So, if you add up it is 9 hours 25 minutes. So, this is the total outage time or the total down time for the things right. So, we can we can say service availability equal to 1 minus, this we have seen previously also and this is the standard thing 1 minus downtime by uptime equal to what we can say 1 minus 9 hours 25 minutes by 360 hours 100 so much percentage, so this fine. So, this was our total expected out time and this is the outage or the downtime, so 1 minus down time by so and so forth. And so, what we have the 97.385 percent. So, this is fine.

We calculate the service availability as 97.385 percent, so as per this data available.

(Refer Slide Time: 20:19)

So, what we see monthly up time percentage is 97.385 as we have calculated which is less than 99 percent right. Not only 99.5 percent, but like that 99 percent. So, service credit available, due to that whatever we whatever during that service negotiation or SLA things are there 25 percent of total cost. So, it is total cost as we have calculated 1500. So, it is dollar 375; so effective cost payable towards by buying the service equal to dollar 1500, minus dollar 375, equal to dollar 1125. So, this is the effective cost.

(Refer Slide Time: 22:09)

Monthly uptime % = 97.385% < 99%.

Service Credit available = 25% of total cost  
=  $\left(\frac{25}{100} * \$1500\right) = \$375$

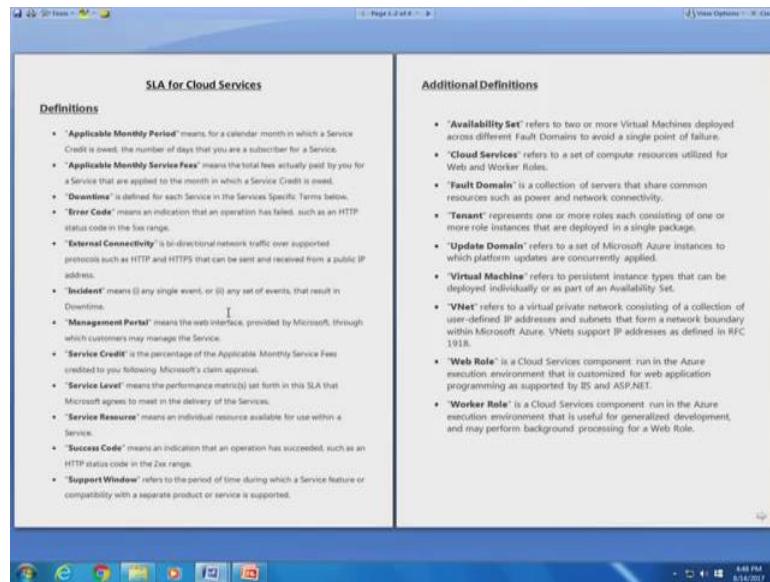
Effective cost payable towards buying the cloud service  
=  $(\$1500 - \$375)$   
= \\$1125

So, what we see that based on the outage. So, it is the if you look at the problem 2 it is the extension of the means little bit extension of the first problem. But this are in reality things happens right. So, we need to measure this things log this things, and calculate the things accordingly so right. This is with respect to the up time and there can be with respect to means it may not be total down, but you are availability band width may be may availability in the network may be slow and type of things.

It depends some lot of other aspects it is not that always straight forward, but there are lot of other complex consideration in doing so. So what we tried in this 2 problem. So, show that the how a SLA guarantees can be calculated or looked into an type of a means how it can be calculated and see that whether the violation of SLA or not right. So, this I believe this will give you a broad means board idea or a things of the.

So now what will see that in like what are the different base practices or what are the different components like one what we have calculated is the up time. So, what are the different components of a cell which are considered by primarily the commercial cloud, right?

(Refer Slide Time: 23:59)



So, that just couple of things will see right. So, so SLA for cloud service provider already we have seen, but just to. So some of the aspects which we would like to highlight this like in case of commercial cloud what things that there is applicable monthly period right, which means for a calendar month in which the service credit is odd the number of days you are a subscriber of a service right.

So, it is applicable service period similarly, applicable monthly service fees right. So, the concept of downtime like as we have seen services in the service specific terms below. Error code means the indication that the operation has failed such that http status code is 5xx or something right. So, that is services should have a error code otherwise you will it is difficult to try pinpoint we services has failed this type of thing.

External connectivity is a bidirectional network traffic support like protocol for http or https can received a public IP and so on and so forth, where they are external connectivity incident means any single or a set of incident that result in down time. Management portal means that the wave interface provided by this is basically meant for Microsoft azure. So, through which the consumer may manage the services like that

management portal. Service credit if there is a failure that how much credit will be given that we have seen in this problem. Service level means the performance mistakes says set forth in the SLA, and in case of myself azure, it agrees to meet the delivery of services service resource.

Success code like as we have seen failure code we have a success code like in a http we know that 2xx is the success code. Support windows refer to the period of time which during which the service features on compatibility with the separate product services is supported. So, there is a support window where the where the things will be there. Along with that there are some additional definitions like availability set refers to 2 or more virtual machine deployed across different fall domains. So, that it will not go for down time at the same time at the same period to avoid single point of failure. Cloud services referrers to the compute resources utilized for web and web role and worker role. Fault domain is a collection of servers that share common resources which are power and network connectivity.

Tenant, represent one or more roles that is one or more role instances that it deployed in a single packages this we have seen like it can be a worker role it can be a web role type of thing. Update domain, refers to a set of in this case Microsoft azure in instances which platform update are concurrently applied, virtual machine this we know. VNet this is a virtual private network and this also is known that web role and worker role. So, these are some additional definitions which will be utilized for service level for SLA calculations, right.

(Refer Slide Time: 27:36)

**Monthly Uptime Calculation and Service Levels for Cloud Services**

- "Maximum Available Minutes" is the total accumulated minutes during a billing month for all Internet facing roles that have two or more instances deployed in different Update Domains.
- "Downtime" is the total accumulated minutes that are part of Maximum Available Minutes that have no External Connectivity.
- "Monthly Uptime Percentage" for Cloud Services is calculated as Maximum Available Minutes less Downtime divided by Maximum Available Minutes in a billing month
  - o Monthly Uptime % = (Maximum Available Minutes-Downtime) / Maximum Available Minutes
- The following Service Levels and Service Credits are applicable to Customer's use of Cloud Services:

MONTHLY UPTIME PERCENTAGE	SERVICE CREDIT

So, similarly as we have seen here, in our example here if you can see the monthly up time calculation and service level for cloud services using those definitions say monthly available minutes is the total accumulated minutes during a billing period for all interfacing roles and 2 or more instances deployed in different update domain similarly down time is that how much time and up time percentage is, the maximum available minutes minus down time by maximum available minutes.

(Refer Slide Time: 28:16)

**Monthly Uptime Calculation and Service Levels for Cloud Services**

- "Downtime" is the total accumulated minutes that are part of Maximum Available Minutes that have no External Connectivity.
- "Monthly Uptime Percentage" for Cloud Services is calculated as Maximum Available Minutes less Downtime divided by Maximum Available Minutes in a billing month
  - o Monthly Uptime % = (Maximum Available Minutes - Downtime) / Maximum Available Minutes
- The following Service Levels and Service Credits are applicable to Customer's use of Cloud Services:

MONTHLY UPTIME PERCENTAGE	SERVICE CREDIT
< 99.95%	10%
< 99%	25%

That we have calculated here right. So, that is the thing and there are can be service script credit rules as we have seen right, 99.95 percent in these exactly the same type of valves we have used.

(Refer Slide Time: 28:30)

**Monthly Uptime Calculation and Service Levels for Virtual Machines**

- "Maximum Available Minutes" is the total accumulated minutes during a billing month for all Internet facing Virtual Machines that have two or more instances deployed in the same Availability Set
- "Downtime" is the total accumulated minutes that are part of Maximum Available Minutes that have no External Connectivity.
- "Monthly Uptime Percentage" for Virtual Machines is calculated as Maximum Available Minutes less Downtime divided by Maximum Available Minutes in a billing month for a given Microsoft Azure subscription.
  - Monthly Uptime % = (Maximum Available Minutes-Downtime) / Maximum Available Minutes
- The following Service Levels and Service Credits are applicable to Customer's use of Virtual Machines:

MONTHLY UPTIME PERCENTAGE	SERVICE CREDIT
< 99.95%	10%

Similarly, this is for calculation and service develop cloud services. Similarly we can have for the VMS right. So, VMS like I want to have infrastructure service and the virtual machines are allocated similarly. So, maximum available minutes is the total accumulated time is billing period and so on and so forth.

(Refer Slide Time: 28:47)

Available Minutes in a billing month for a given Microsoft Azure subscription.

- Monthly Uptime % = (Maximum Available Minutes-Downtime) / Maximum Available Minutes
- The following Service Levels and Service Credits are applicable to Customer's use of Virtual Machines:

MONTHLY UPTIME PERCENTAGE	SERVICE CREDIT
< 99.95%	10%
< 99%	25%

Down time is similarly we can calculate. And we can have several separate same type of service credit. So, it can be at what we mean to say it can be a different type of level it can be at a IaaS level, it can be a PaaS levels any SaaS level. I can have as a storage level of the if there is storage downtime or access ability problem so on and so forth.

So, this need to be clearly specified, now when we want to do this type of thing; so what are the different best practices or rules we need to follow right. So, that way let us see some of the best practices what are follow.

(Refer Slide Time: 29:31)

**SLA Best Practices**

The Cloud Standards Customer Council provides cloud consumers with the seven steps they should take when evaluating cloud SLAs (published by the Cloud Standards Customer Council in April 2012)

- **Identify the cloud actors**
  - As per NIST reference architecture:
    - Cloud consumer
    - Cloud provider
    - Cloud carrier
    - Cloud broker
    - Cloud auditor
- **Evaluate business-level policies**
  - Data policy, SLA guarantees, list of services not covered, excess usage, payment and penalty methods, subcontracted services, licensed software and industry specific standards.
- **Understand SaaS, PaaS and IaaS**
  - To understand what SaaS, PaaS, and IaaS are about and which type of cloud it is running on (private/public or hybrid).
  - Terms and conditions in the SLA depend on the complexity of control variables that the provider gives to the consumers.
- **Metrics**
  - To identify what metrics should be used to achieve performance objectives. Some examples of availability and response time metrics are:
    - Metric name in SLA
    - Constraints
    - Method and frequency of collection
- **Security**
  - Consider key security requirements for cloud SLAs, including:
    - Asset sensitivity
    - Legal/regulatory requirements
    - Cloud providers' security capabilities
- **Identify service management requirements**
  - What should be monitored and reported (for example, load performance, application performance), and what should be metered.
  - How rapid provisioning should be (speed, testing, demand flexibility) and how resource change should be managed.
- **Prepare for and manage service failure**
  - Determine what remedies should be provided (for example, service credits) and what are the liability limitations.
  - How the disaster recovery plan will work when needed.
  - An exit clause should be part of every cloud SLA.
  - The consumer or provider wants to terminate the SLA.

So, the cloud standard custom council right, provides cloud consumer with 7 steps they are they should take when they evaluating the SLA's like it is also provided in their document of April 20, 12 right. So, identify the cloud actors. Who are the actor according to NIST architecture? So, these are the actors, consumer, producer or the provider, carrier, broker and auditor. So, these are the 5 actor which are there as far as NIST. So, we need to evaluate the business level policies right. This is important what will be the data policies SLA guarantee least of services, not covered under this excess uses payment penalty sub contract services, license software, industry specific standards and these are different aspects of the things.

So, what while we are discussing about simple SLA's in actually the things are more complex like what should be the data policy is how which are covered which are not covered, if there is a sub contract of services what should be that policies, whether you

are using license software licensing mechanisms and so on and so forth, right. Because most of the cases when we try to use these we may be using different license software, and those licensing cost etcetera come into play not only that licensing period, and so and other things come into play.

Then we need to understand that which level operations we are looking at, SaaS PaaS or IaaS, because the different type of things are the different type of services have different type of requirement. In some cases SaaS is much easier to control maybe or measure, but we need to look at that which type of services we are leveraging on, whether we are having multiple this sort of services. So, to we need to understand what SaaS, PaaS, IaaS are about and which type of cloud it is running whether is a public private or hybrid. Terms and condition in a SLA depends on the complexity of control variables that are provide that the provider gives to the consumer or the service consumers. So now consumer need to calculate the availability etcetera. So, for that the controlled control variables are provided by the like I say it gives me the CPU up time etcetera or different uses time or hard that disc uses parameters. So, these are the different control parameters provided.

Now, more the complexity of this parameter depends on which level of operations you are doing and where you are running the things, like is a IaaS, space or SaaS or whether it is a hybrid or your public or private. So, the other things are one is that metric what we are discussing about, to identify what metrics should be used to achieve performance objectives right. Some examples availability at availability response metrics are like metric name in the in SLA like availability and other type of things, there may be other constraints whether and frequency of collection of these data is also important. The aspect the next aspect is the security like, consider key security parameters for cloud including asset sensitivity, legal regulatory requirements like I a I may say that the that datas would reside within these particular geographically boundary or within this type of things. Cloud provider security capabilities what is the capability of the cloud provider to provide that.

Then we have service management requirement, to need to identify the service management requirement. So, what should be monitored and reported for example, load performance application performance or what should be metered right. What you are need to be bill meter. How rapid provisioning should be like speed, testing, demand

flexibility and how resource charged should be managed right. So, how is the provisioning, what need to be monitor and reported and meter type of things need to be looked into.

And then prepare for and manage for the failure right. There is another important exit. Determined what remedy should provided like for example, service credits and what are the liability limitation. So, how much service credits to provide and what are my liabilities on the provider the provider point of view, and in order to that what the consumer are signing of. How the disaster recovery plan will work when needed. So, how the disaster recovery plan will work when it is needed, and exit clause should be a part of every cloud SLA right. In either the consumer or the provider wants to terminate the relationship. So, SLA what it is there it is a agreement. So, what should be the exit clause suppose the consumer at some point of time to exit or the providers says that I am not able to provide that thing. So, that should be in the thing.

So, what we says that this are some of the essential best practices, or some best practices we should keep in mind when formulating the SLA etcetera. Like identify the cloud actors, evaluate business level policies, understand this type of services, what are the different matrixes security capability of the; and requirement security requirement of the consumer, and the capability of the provider. Service management requirements, and how to manage failure or what should be the remedies for failure.

So, what we tried this is a what we send a extension of the SLA already we have discussed. So, what we try to give that the there are different aspect to the things and try do in this thing which we have also seen to simple SLA related problem, how it can be how this type of SLA's are calculated though the problems are very simple and straight forward, but it gives us a idea that how you can apprise the approach this type of things. So, will let us conclude here for this SLA tutorial.

Thank you.

**Cloud Computing**  
**Prof. Soumya Kanti Ghosh**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**

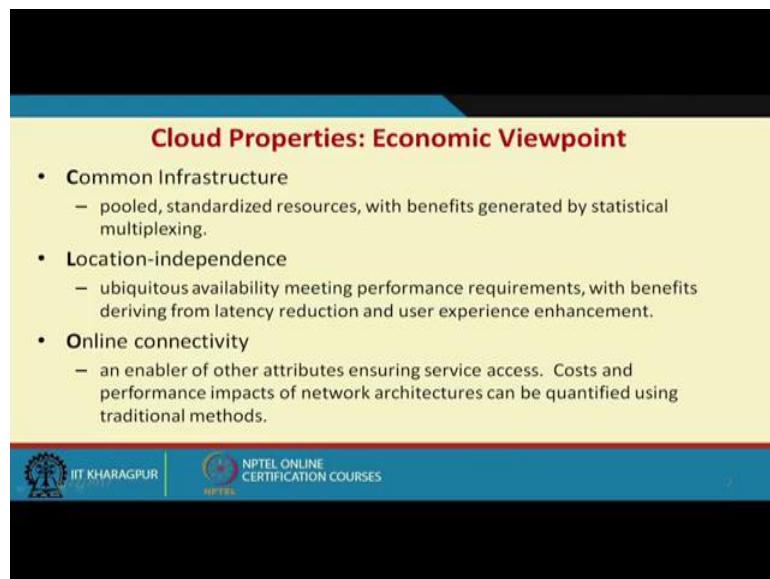
**Lecture - 22**  
**Economics Tutorial**

Hello. We will continue our discussion on Cloud Computing. Today, we will do some problem or some exercise on cloud economic, so that whether it is economical to use cloud or when it is economical to use cloud.

We have already discussed this particular topic in our previous lecture on cloud economic but we will see that we will look at some of the problems, some troy problem so that it makes our concept little more clear.

So, I just have couple of slide recap of over the earlier slide so that it will be easy for ask to get tune to the problem.

(Refer Slide Time: 01:11)



**Cloud Properties: Economic Viewpoint**

- Common Infrastructure
  - pooled, standardized resources, with benefits generated by statistical multiplexing.
- Location-independence
  - ubiquitous availability meeting performance requirements, with benefits deriving from latency reduction and user experience enhancement.
- Online connectivity
  - an enabler of other attributes ensuring service access. Costs and performance impacts of network architectures can be quantified using traditional methods.

**IIT KHARAGPUR** | **NPTEL ONLINE CERTIFICATION COURSES**

So, as you have discussed in cloud property if you from the economic point of view, it is a common infrastructure like pooled, standardized resource with benefits generated by statistical multiplexing. So, we are multiplexing the demands so that is from a pooled resources, we can do; location independent, online connectivity; that is broad network access.

(Refer Slide Time: 01:36)

**Cloud Properties: Economic Viewpoint (contd...)**

- Utility pricing
  - usage-sensitive or pay-per-use pricing, with benefits applying in environments with variable demand levels.
- on-Demand Resources
  - scalable, elastic resources provisioned and de-provisioned without delay or costs associated with change.

 IIT KHARAGPUR |  NPTEL ONLINE CERTIFICATION COURSES

Utility pricing like you uses; sensitive or pay-per-use or pay as you go model; these benefits applying environments with variable demand levels session so on and so forth and on demand resources; so, it is a scalable, elastic resources provisioned de-provisioned without delay or cost in the associated with change.

So, overall it looks like that it is a win-win situation that whatever we do is all positive things, but whether it is a company or a organization needs to look into that whether it is a beneficial to deploy cloud depending on its type of things. As we discussed, suppose you require system a some high end system for a short period of time; then it may be beneficial to go for cloud.

Suppose the requirement of the system is longer period of time; daily you are using 12 hours or so; or may be more than that and then it may not be always economical to go for cloud type of things; so, in economic point of view that we will try to look at.

(Refer Slide Time: 02:45)

D(t)	demand for resources $0 < t < T$
P	$\max(D(t))$ : Peak Demand
A	Avg ( $D(t)$ ) : Average Demand
B	Baseline (owned) unit cost [ $B_T$ : Total Baseline Cost]
C	Cloud unit cost [ $C_T$ : Total Cloud Cost]
U ( $=C/B$ )	Utility Premium [For rental car example, $U=4.5$ ]

$C_T = \int_0^T U \times B \times D(t) dt = A \times U \times B \times T$

$B_T = P \times B \times T$

- Because the baseline should handle peak demand

When is cloud cheaper than owning?

$C_T < B_T \Rightarrow A \times U \times B \times T < P \times B \times T$

$\Rightarrow U < \frac{P}{A}$

- When utility premium is less than ratio of peak demand to Average demand

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

We have also seen these slides where we tried to look at that overall costing model where cloud cost is like; here the variables were that the  $D(t)$  demand  $D$   $t$  is a demand is function of time; within a time period 0 to  $t$ ;  $t$  is the max of  $D(t)$ , that is a peak demand,  $A$  is the average demand,  $B$  is the baseline that is the unit cost for owning the things,  $C$  is the cloud unit square; the total cost is denoted by  $C_T$ ; total baseline  $B_T$ .

So,  $U$  or  $C$  by  $B$  is the utility premium; like the example we have shown in our earlier slide is utility premium is 4.55, so; that means, it is all beneficial to use clouds so what is the utility. Similar, so we can calculate we have shown that we can calculate that  $0$  to  $t$   $U$   $B$   $D$   $t$  as a multiplication of these units. Similarly,  $B$   $t$  that is the total baseline cost is  $P$  into  $B$  into  $t$  because the baseline should be handle peak time.

So, peak demand rather; so, baseline whenever we do a baseline costing of baseline design or what whenever you own or something; then we usually look for a peak demand. Like suppose I want to big make system for our say M Tech lab; so, what you have to say that if the number of students which can be admitted say  $n$ ; so, number of system we think is; it should be at least  $n$ ; if not more to have some redundancy into the things, if there is a failure then recovery means; we can replace the systems.

But in actual things that the number of student's strength may not be; for a particular year of admission you may not attend that thing. Usually it may be less than that and so you have redundancy resources which are underutilized or not utilized.

So, when cloud is cheaper? If this cost of  $C_T$  is less than  $B_T$  or utility is  $P$  by  $A$  is less than  $P$  by  $A$  or here if you have the cost of cloud; by this what should be the utility premium. When utility premium is less than the ratio of the peak demand to the average demand then we can say it is cheaper.

(Refer Slide Time: 05:07)

**Utility Pricing in Real World**

- In practice demands are often highly spiky
  - News stories, marketing promotions, product launches, Internet flash floods, Tax season, Christmas shopping, etc.
- Often a hybrid model is the best
  - You own a car for daily commute, and rent a car when traveling or when you need a van to move
  - Key factor is again the ratio of peak to average demand
  - But we should also consider other costs
    - Network cost (both fixed costs and usage costs)
    - Interoperability overhead
    - Consider Reliability, accessibility

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Now, utility pricing in the real world is not that simple; if there are lots of other things like in practice demands of highly spikes are there. Like there are maybe news stories or marketing promotion and so on and so forth; of an hybrid model is based. You own a car for daily use; for whenever you are going out or traveling or for a longer distance, you use those rented things. Similarly, you have your own system but when you require some; when you are thinking there is some peak demand; you go for this type of things.

So, key factor is again the ratio to peak to average diamond, so that is the key factor. But we should also consider other cost which are not considered here; like network cost, interoperability over head; if two data are talking to each other then interoperability over head. Then consider reliability accessibility and so on and so forth.

So, those are other things which need to be accounted for when we calculate this cost.

(Refer Slide Time: 06:10)

**Value of on-Demand Services**

- Simple Problem: When owning your resources, you will pay a penalty whenever your resources do not match the instantaneous demand
  - Either pay for unused resources, or suffer the penalty of missing service delivery

D(t) – Instantaneous Demand at time t  
R(t) – Resources at time t

Penalty Cost  $\alpha \int |D(t) - R(t)| dt$

- If demand is flat, penalty = 0
- If demand is linear periodic provisioning is acceptable

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, we also have seen this is a slide again; where there is on demand services, value of on demand services. When owning your own resource, you pay penalty whenever your resource do not match with the instantaneous demand. Either you penalty for unused resource or suffer for penalty for missing the service delivery; it can be there. So, if  $D_T$  is the instantaneous demand and  $R_T$  is the instantaneous resource; then the penalty cost is  $D_T$  minus  $R_T$  integral of that. Like  $D_T$  minus  $R_T$  mode integral of that over the time period 0 to capital T.

So, that is the thing if demand is flat; penalty is 0. If the demand is linear; periodic provisioning is acceptance. So, if the demand is linear periodic provisioning is accepted.

(Refer Slide Time: 07:00)

### Penalty Costs for Exponential Demand

- Penalty cost  $\propto \int |D(t) - R(t)| dt$
- If demand is exponential ( $D(t)=e^t$ ), any fixed provisioning interval ( $t_p$ ) according to the current demands will fall exponentially behind
- $R(t) = e^{t-t_p}$
- $D(t) - R(t) = e^t - e^{t-t_p} = e^t(1 - e^{-t_p}) = k_1 e^t$
- Penalty cost  $\propto c.k_1 e^t$

**IIT Kharagpur** | **NPTEL ONLINE CERTIFICATION COURSES**



But most of the case; there is a lag between when the demand rises and the provisioning is done. And this sometimes creates a; it may create a big challenge because if the growth of demand is exponential, then chasing that demand with resource provisioning maybe extremely difficult. Like here we have seen that if the demand is exponential;  $D_t$  is some form of  $e$  to the power  $t$ , any fixed provisioning interval  $t_p$ ; like it whenever you look for that extra provisioning, it will look for some provisioning interval  $t_p$ .

So, around any according to the current demand will fall exponentially behind it; we have seen. So,  $R(t)=e^{t-t_p}$ ; so, you have a lag between that. So,  $D(t)-R(t)$ ; if we calculate something constant into the  $e$  to the power  $t$ ; so that means, the penalty is something  $C$  into  $k_1$ ;  $e$  to the power  $t$ ; that means, it also grows exponentially. Like, if you see the picture, so that due to this provisioning delay that it unserved demand goes on increasing. So, it is extremely difficult to chase this sort of situations.

So, these are all practical consideration; so, it is not that just doing a simple calculation, but there are different other consideration. For that, you need to have a predictive model that whether you can say that this time the demand will increase like; these are true for our other commercial things also, we do predict that during some seasonal things whether if the demand will increase on that time and have store accordingly. Here also you need to have some predictive model to do that if at all required.

(Refer Slide Time: 08:51)

**Assignment 1**

Consider the peak computing demand for an organization is 120 units. The demand as a function of time can be expressed as:

$$D(t) = \begin{cases} 50 \sin(t), & 0 \leq t < \pi/2 \\ 20 \sin(t), & \pi/2 \leq t < \pi \end{cases}$$

The resource provisioned by the cloud to satisfy current demand at time  $t$  is given as:

$$R(t) = D(t) + \delta \cdot \left( \frac{dD(t)}{dt} \right)$$

where,  $\delta$  is the delay in provisioning the extra computing recourse on demand.

The cost to provision unit cloud resource for unit time is 0.9 units.

Calculate the penalty.

[Assume the delay in provisioning is  $\pi/12$  time units and minimum demand is 0]

(Penalty: Either pay for unused resource or missing service delivery)

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Now, we will try to look at one or two problems; which will allow us to have little more clarity on the whole thing. So, these are very toy problems; simple problem, but by doing so it will give us some sort of a clarity or confidence into the things.

So, what it says; consider a peak computing demand of an organization is 120 units, something, some 120 units. The demand is a function of time that can be expressed as this  $D_T$  equal to  $50 \sin t$  from  $t=0$  to  $\pi/2$ . Whereas,  $20 \sin t$  where  $\pi/2$  to  $\pi$ ; this is the same functional model is given. The resource provisioning by cloud to satisfy current demand  $t$  is  $R(t) = D(t) + \delta \left( \frac{dD(t)}{dt} \right)$ .

So, this is the resource provisioning; where  $\delta$  is a delay in provisioning the extra computing resource in time; this we consider. Now, the cost of provision unit cloud resource for unit time is 0.9 units; so, we need to calculate the penalty. So, our assumption is that the delay in provisioning is  $\pi/12$  time units and minimum demand is 0; there is no demand is the minimum case. Penalty either pay for the unused resource or you miss some service delivery. So, this is the consideration; so, we need to calculate the penalty.

So, let us just have a very straight forward working and see that how things work. So, if I have your  $R$  of 0 to  $\pi$  integral 0 to  $\pi$  from that given equation you can do.

(Refer Slide Time: 10:50)

The image shows a handwritten derivation of two integrals. On the left, the integral  $R[0, \pi] = \int_0^\pi D(t) dt + S \int_0^\pi \frac{dD(t)}{dt} dt$  is given. A bracket groups the first term and the second term with its derivative. An arrow points from this bracket to the right side of the equation. On the right, the first term is evaluated as  $S \left[ \int_0^{\pi/2} 50 \cos t dt + \int_{\pi/2}^\pi 20 \cos t dt \right]$ . The second term is evaluated as  $S \left[ 50 \sin t \Big|_0^{\pi/2} + 20 \sin t \Big|_{\pi/2}^\pi \right]$ . Both terms simplify to  $S [50(1-0) + 20(0-1)] = S \times 30 = \pi/12 \times 30$ , resulting in  $= 70$ .

So, if we look at this portion; so, it is del of  $D_T$  30; so, on other sense this we can have; so if you go on calculating; so, it is 50 minus cos of t 0 to  $\pi/2$  plus 20 minus cos of t  $\pi/2$  by pi. So, if you go calculate; so, it will come as 70.

So, that what we have finally, if you look at.

(Refer Slide Time: 14:05)

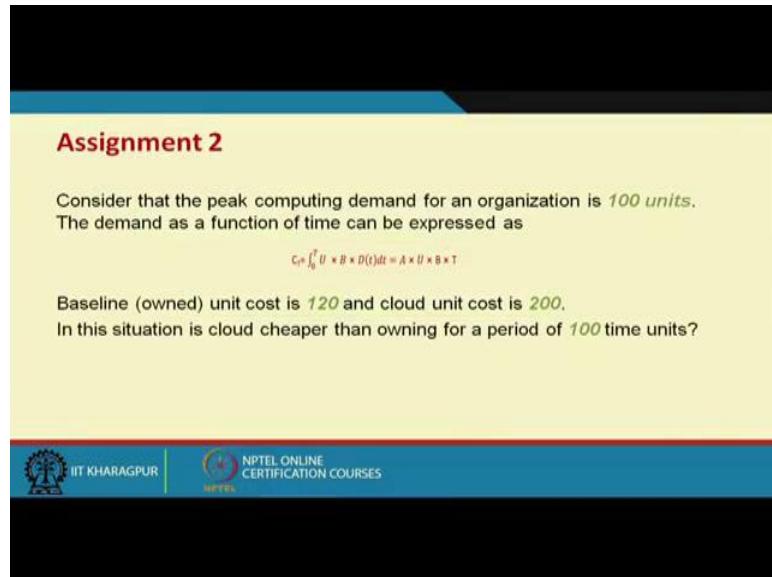
The image shows a handwritten derivation. It starts with  $R = I_0 + 30 \pi/12$  and notes a 'penalty' of  $\int_0^\pi |D(t) - R(t)| dt$ . Below this,  $D = I_0$  is written. Then,  $|R - D| = \frac{30\pi}{12} = \underline{7.86}$  is shown with a checkmark. There is a small blue scribble at the bottom left.

So, R equal to 70 plus 30 pi by 12 and your D value is 70; so, R minus D mode is 30 pi by 12; equal to 7.86. So, this is the value what we answer we will get; so, if you actually

this corresponds; if you remember as we are talking about penalty; penalty equal to is proportional to  $0; 2 \pi$  demand minus mode of  $D_T$ . So, we get these as the value of the penalty because in this case; what we say that is your resource availability that provisioned is greater than your; the demand; so, we get a positive value.

So, this is again from the equation we calculate; again I should say this all this calculation path, we have considered the other factors; like network provisioning or other type of maintenance cost and all those things, we considered; those we are not considered.

(Refer Slide Time: 15:55)



The slide has a black header and footer. The main content area has a blue bar at the top and a yellow background below it. The title "Assignment 2" is in red. The text asks to consider peak computing demand of 100 units and provides a formula  $C_T = \int_0^T U * B * D(t)dt = A * U * B * T$ . It then asks if baseline unit cost of 120 and cloud unit cost of 200 make cloud cheaper over 100 time units. Logos for IIT Kharagpur and NPTEL are at the bottom.

**Assignment 2**

Consider that the peak computing demand for an organization is **100 units**.  
The demand as a function of time can be expressed as

$$C_T = \int_0^T U * B * D(t)dt = A * U * B * T$$

Baseline (owned) unit cost is **120** and cloud unit cost is **200**.  
In this situation is cloud cheaper than owning for a period of **100** time units?

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, rather if we look at another problem like; consider the peak computing demand of an organization is 100 units. The demand is a function that can be expressed and that can be expressed as  $C_T$  equal to  $0$  to  $t$   $U$  into  $B$  into  $D(t)$  and like this. So, this is the function what we have discussed.

(Refer Slide Time: 16:24)

$$C_T = \int_0^T U \times B \times D(t) dt = A \times U \times B \times T$$

Demand function:  $\underline{D(t) = 50(1 + e^{-t})}$

Baseline (normal) unit cost = 120

Cloud unit cost = 200

So, what we say; the cost of cloud  $C_T$  equal to 0 to  $t$ ;  $U$  cross  $B$  into; this we have the already seen. So, rather here the this is the total cloud cost; this we have already seen. Actually, if we; please note the demand function, which is there is it is not mentioned here; there is some typo. So, demand expressed as a time can be expressed as  $50(1 + e^{-t})$ .

So, let me reframe the question again; consider a peak computing demand of an organization is 100 units. So, demand as a function of time can be expressed as this, so if I say the demand  $D_T$  is expressed as this. So, this is not there in the question because there is a some missing terms. So, demand you considered  $1 + e^{-t}$ ; so, this is a; demand is a function of time.

Now, what we additionally give baseline unit cost is 120; so, base line that you means in unit cost is 120 and cloud unit cost 200. So, what we need to do in this situation is cloud cheaper than owing for a period of 100 unit times. So, you want to calculate if the cloud is cheaper then owing it into 100 unit times. So, what you need to add here? This you need to calculate that whether if 100 unit time; if it is a utilizing is there; so, it is again straight forward.

(Refer Slide Time: 19:00)

Total Baseline Cost :  
 $B_T = P \times B \times T = 100 \times 120 \times 100 = 1200000$

Total Cloud Cost :  
 $C_T = \int_0^{100} C \times D(t) dt = \int_0^{100} 200 \times 50 (1 + e^{-t}) dt$   
 $= [10000 (t)]_0^{100} + [-e^{-t}]_0^{100}$   
 $\approx 1010000$

Utility Premium =  $\frac{C_T}{B_T} = \frac{1010000}{1200000} = 0.84 < 1$

Cloud will be cheaper.

So, if I see total baseline cost  $B_T$  equal to  $P$  into  $B$  into  $t$  equal to 100 into 120 into this much; total cloud cost  $C$  of  $t$  equal to 0 to 100;  $C D$  of  $t$ ;  $D_T$  equal to integral of 0 to 100; 200; 50, 1 minus 1 plus minus this into  $D_T$ .

So, if you do a integration; so, if you do it, it approximately count as 1010 so much. If you just do that calculation. So here what we get 120; 1, 2, 3, 4, 5; 1, 2, 3, 4, 5 fine. So, what we do utility; so, utility premium as if from our this; so 0.84 or utility premium  $U$ , what we get is  $U$  is less than 1. So; that means, at least in this case cloud will be cheaper.

So, I can have simple calculation to find out that what are the different; for different scenarios, what are the different cases are there.

(Refer Slide Time: 21:45)

**Assignment 3**

A company X needs to support a spike in demand when it becomes popular, followed potentially by a reduction once some of the visitors turn away. The company has two options to satisfy the requirements which are given in the following table.

Expenditures	In-house server (INR)	Cloud server
Purchase cost	6,00,000	-
Number of CPU cores	12	8
Cost/hour (over three year span)	-	42
Efficiency	40%	80%
Power and cooling (cost/hour)	22	-
Management cost (cost/hour)	6	1

- Calculate the price of a core-hour on in-house server and cloud server.
- Find the cost/effective-hour for both the options.
- Calculate the ratio of the total cost/effective-hour for in-house to cloud deployment.
- If the efficiency of in-house server is increased to 70%, which deployment will have now better total cost/effective-hour?

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Let us look at another problem; here what says a company X needs to support a spike in demand when it come when it become demand when it becomes popular followed by potentially a reduction once some of the visitors turns away; that means, when there can be spike and fall in the demand; this sort of scenarios.

The company has two option to satisfy the requirement which are given in a table. So, what it says purchase cost of in house is something and cloud server in a way there is no purchase cost. Number of CPU cores for in house server is 12 whereas the cloud server is 8. Cost per hour over 3 year span for the cloud sever as you are renting; it is something 42 units per hour cost.

Efficiency of the in house server because it is underutilized, so it is 40 percent whereas, the cloud server is 80 percent; because you are using it when the demand is there. Power and cooling requirement in case of in house server is 22; whereas, for the cloud server is nil and management cost per hour is 6; in case of in house server and cloud server it is 1.

So, again let me just repeat purchase cost is 6 lakh for in house and nil for cloud. Number of CPU core in house is 12, cloud 8; cost per hour in a 3 year span that in or other; means for cloud is 42 units whereas, for in house there is no cost hour; you can after on purchase the thing. If you since is 40 percent because most of the things are there is 60 means not that utilized. And whereas cloud sever; it is 80 percent power and

cooling is 22 and nil and management cost per hour in case of a cloud is much more than is 6 and it is 1.

So, what you need to calculate with this data; calculate the price of a core hour on in house and cloud sever so that is straight forward.

(Refer Slide Time: 24:07)

Cost / hour for in-house server (in yr) =

$$\frac{6,00,000}{3 \times 365 \times 24} \approx 22.83 \text{ INR}$$

(a) Cost / hour for in-house Server =  $\frac{22.83}{12} = 1.90 \text{ INR}$

Cost / hour for Cloud ... =  $\frac{42}{8} = 5.25 \text{ INR}$

(b) Cost / Effective hr for in-house =  $\frac{22.83}{40/100} = 57.075$   
... Cloud =  $\frac{42}{80/100} = 52.5$

So, what we have cost per hour for in house sever considering 3 years is divided by 3 cross 365 cross 24. So, if you calculate it comes around 22.83 INR; if it is Indian currency or units of cost.

Now, if we have a cost hour for in house; server it is 22.83 by 12 because the first question was core hour; for in house server. So, core hour for in house server is this one is 1.90 unit or if we consider a new thing. So, core hour for cloud server is 42 by 8; it directly comes from the question 5.25 INR. Now an question B; find the cost per effective hour for the both the option.

So, cost per effective hour for house is 22.83 and as it is 40 percent efficient; so, 40 by 100, so it is 57.075; so, it is INR or something; whatever the unit. Effective hour for cloud equal to it is 80 percentage efficiency; so, 5; so, this is the effective.

And now calculate the ratio of total cost per effective hour for in house to cloud deployment.

(Refer Slide Time: 27:15)

© CBT I.I.T. KGP

(c) Total cost / effective hr for in house =  $57.075 + 22 + 6$   
 $= 85.075 \text{ INR}$

Cloud =  $52.5 + 1 = 53.5 \text{ INR}$

Ratio  $\left[ \frac{\text{Inhouse}}{\text{Cloud}} \right] = \frac{85.075}{53.5} = 1.59$

(d) Modified effective cost / effective hr for in house =  $\frac{22.83}{70/100} = 32.61$

(e) Total cost / effective hr for in house =  $\frac{(32.61 + 22 + 6)}{70/100} = 60.61 \text{ INR}$

(f)  $\frac{22.83}{90/100} = 25.37 \rightarrow 25.37 + 22 + 6 = 53.37 \text{ INR}$

So now need to calculate the ratio of total cost per effective hour for in house and cloud deployment. So, total cost per effective hour in house is 57.075; if you have calculated earlier that is for the in house sever plus 22 because that is a power and cooling plus 6 for maintenance management cost and then we have 87.075. Same thing for cloud equal to 52.5 and it has only one management; so, it is 53.5 and if it is INR fine or whatever the unit.

So, ratio of effective hour of in house by cloud equal to 85.075 by 53.5; so 1.59 and finally, if the efficiency of in house sever units is to 70 percent; which deployment have will become better in the total cost effective hour. So, it has efficium efficiencies initially now 40; if you increase to 70 percent; so, which will be better. So, modified cost per plus effective hour for in house will be 22.83 by 70 percent now 100; so, we get 32.61.

So, total cost plus effective hour for in house; everything is for in house total cost per effective hour; still it is for in house, still it is 32.61 plus 22 plus 6; equal to 60.61. But if we see that the cloud it is still 53.25, now if it is instead of this is 70 percent increase. If it is a 90 percent in house; then if you do the same calculation, we will effectively what we will get instead of this value. So, 22.83 by 90 by 100; it is 25.37; so, with this total cost for 90 percent will become 25.37 plus 22 plus 6; 53.37 INR.

So, this is for if it is 90 percent; so, if you see 90 percent; then it is better than this cloud. So, it is if you look at the problem; so, if in your efficiency is less. So, as we discussed

the overall utilization of the resource of the in house is highly underutilized, so you lose on those things. So, effectively we get a less it means we get a more benefit if we take from cloud.

However if your efficiency increases or in other sense that your infrastructure or your in house infrastructure; if it is heavily utilized like a we have seen up to 90 percent, then it may be better then it will be performance wise better than cloud. This is a very synthetic for example, to show that if the efficiency if you are more utilization are there then in this case is the in house will be better than cloud type or cloud purchasing a cloud.

So, it all depends on overall; what is your demand? What sort of demand is there? How long duration things are there? So, we need to take a call that whether you go for a cloud type of; economically beneficial to go to cloud service provider or have in house infrastructure.

So, this three is small problems on economic model of cloud shows that how by simple calculation we can find out and though in actual practice these are more complex with considering other for parameters as we have discussed.

So, hope this will help you in clearing or having a better understanding of this economy behind this cloud.

Thank you.

**Cloud Computing**  
**Prof. Soumya Kanti Ghosh**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**

**Lecture - 24**  
**MapReduce – Tutorial**

Hi. Today we will discuss some tutorial on MapReduce. Already we have discussed on MapReduce. So, today we will try to solve one or two problems or rather try to see how one or two; how we can decompose a problem into a MapReduce problem and how to work on it.

So, if you remember that a MapReduce paradigm is used for processing huge volume of data where paralysation is possible. And primarily rather developed by Google and later on used in various fields. So, what we will do; we will initially couple of slides we will have a quick recap before we take up one or two problems related to this MapReduce framework.

(Refer Slide Time: 01:15)

**Introduction**

- MapReduce: programming model developed at Google
- Objective:
  - Implement large scale search
  - Text processing on massively scalable web data stored using BigTable and GFS distributed file system
- Designed for processing and generating large volumes of data via massively parallel computations, utilizing tens of thousands of processors at a time
- Fault tolerant: ensure progress of computation even if processors and networks fail
- Example:
  - Hadoop: open source implementation of MapReduce (developed at Yahoo!)
  - Available on pre-packaged AMIs on Amazon EC2 cloud platform

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, as we discussed already; so, it is a programming model developed at Google. Implement large scale search primarily; the basic objective was to implement large scale search. Text processing on massively scalable web data using data stored using Big Table and GFS distributed file system; Google file system and big data; Big Table.

So, this was a objective of Google which it started with; so, design for processing and generating large volume of data by massively parallel computation; utilizing tens and thousands of processor at a time. So, there are a huge number of processors to the tune of tens and thousands. So, whether if there is an inherent parallelism inside the things whether I can exploit it into the thing, so it is designed to be fault tolerant; that means, ensure progress of the computation; even if the processor or network fails. So, it should be fault tolerant to the extent that even there is a failure in the processors or the network; it to still the working should go on.

So, that was the basic assumption or basic; let us say precondition of doing this or that was basically these; what they was taken up. So, there are several things like Hadoop, open source implementation of MapReduce; incidentally it was developed by Yahoo; available on pre packaged AMIs on Amazon EC2 and so and so forth.

(Refer Slide Time: 02:45)

## MapReduce Model

- Parallel programming abstraction
- Used by many different parallel applications which carry out large-scale computation involving thousands of processors
- Leverages a common underlying fault-tolerant implementation
- Two phases of MapReduce:
  - Map operation
  - Reduce operation
- A configurable number of M 'mapper' processors and R 'reducer' processors are assigned to work on the problem
- The computation is coordinated by a single master process

 IIT KHARAGPUR |  NPTEL ONLINE CERTIFICATION COURSES



So, if we look at apart from its history; so, it is a parallel programming abstraction used by many different parallel applications; which carry out large scale computation involving thousands of processors. It is again as we have doing the underlining fault tolerant implementation; both on the data side and on processor and network side. So, everything is fault tolerant; divided into two phases; one is a map phase. So, mapping the; so, the given a problem I divide into two phases. So, it is mapped into a intermediate

result and reduced to a; again, the reduce function or reduce phase; reduce those intermediate result to the actual results.

So, in doing so; what we have seen in our earlier lecture or earlier discussion on MapReduce that we can have parallel efficiency in; we can achieve substantial parallel efficiency, when we are dealing with a large volume of data and there is inherent parallelism into the process. So, what we look at there are M number of mapper processor and R number of reducer processors; which are assigned work based on the problem.

So, there is a master controller, M mapper processor and say R number of reducer cell processors which work on that. And in some cases this mapper and reducer; processor can share at the same physical infrastructure. So; that means, sometimes we acting as a mapper and at a later stage acting as a reducer type of things. So, it is a our capability of the developer or it is that how you devise these mapping and map and reduce functions. Implementation also based on that whatever that language; the developer working on, maybe there are lot of; means people are working on python, it can be on c plus plus and other type of coding things.

So, that that coding part is based on that what sort of problem and what is the environment you are working on. But primarily a philosophy there are M number of mapper and a set of reducer; you have intermediate results into the thing.

(Refer Slide Time: 05:09)

The slide has a blue header bar. The main content area has a yellow background. At the top, the title 'MapReduce Model Contd...' is centered in red. Below the title, there is a bulleted list describing the Map phase:

- Map phase:
  - Each mapper reads approximately  $1/M$  of the input from the global file system, using locations given by the master
  - Map operation consists of transforming one set of key-value pairs to another:  
Map:  $(k_1, v_1) \rightarrow [(k_2, v_2)]$ .
  - Each mapper writes computation results in one file per reducer
  - Files are sorted by a key and stored to the local file system
  - The master keeps track of the location of these files

At the bottom of the slide, there are two logos: IIT Kharagpur and NPTEL. To the right of the NPTEL logo is a circular portrait of a man.

So, as we discussed earlier that each map is each mapper reads  $1/M$  th of the input from the global file system, using locations given by the master. So, master controller; the controller of the master node says that these are the chunks you need to read.

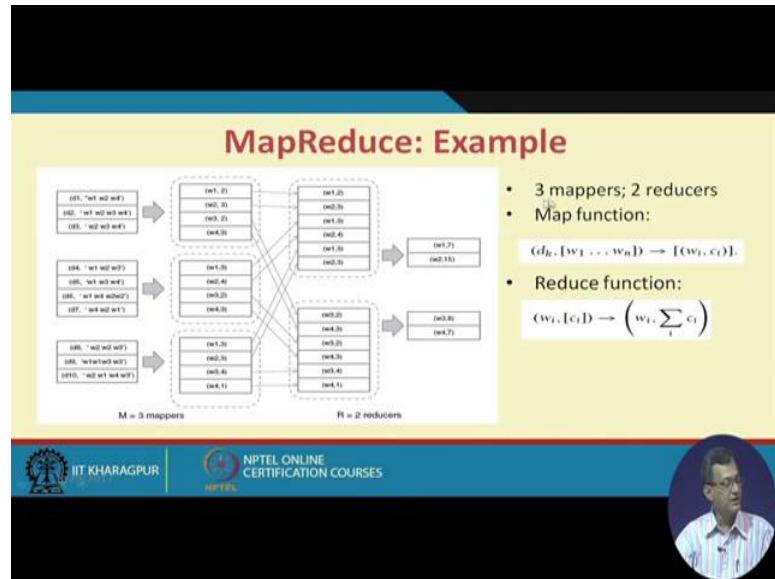
So, map function consists of a transformation from one key value pair to another key value pair. Like here I have a  $k_1, v_1$  mapped to  $k_2, v_2$ ; each mapper writes computation results of one file per reducer. So, it is prepared typically if there are  $R$  reducers. So, it prepares the result for one file per reducer, if there is one reducer; so one file that it creates a file for the user.

Files are stored in a key and sorted by a key and stored in a local file system. So, that is a local file systems where the output of the mapper are stored. The master keeps track of the location of this file, the master has the tracking of the things. On the reducer phase or the reduce phase; the master informs a reducer where the partial computation because the mapper has done a partial computation of the whole process; have been stored on the local file system for respective mappers.

So, if there are  $M$  mappers for the respective mappers where the files are stored for that particular reducer. Reducer make remote procedure call; request the mapper to face the files. Each reducer groups the results of the maps tape using the same key and performs a function; some function  $f$  and list of values corresponding to this value. That means, if I have as you as  $k_2, v_2$  then I map it to some  $k_2$  and function of that  $v_2$ .

So, if the function may be as simple as averaging; so, maybe frequency or the count or some complex functions of doing some more operations. So, results are written back to the Google file system, so the Google file system takes care of them.

(Refer Slide Time: 07:18)



So, MapReduce example; so, there are 3 mappers; 2 reducers, map function is in this case as we; if you remember or if you look at our previous lecture and discussion. So, there is a huge volume of word and what we want to do a word count. So, the every mapper has a chunk of the data things like this mapper has D 1, D 2, D 3 and this is D 4, D 7, D 8 etcetera.

So, every mapper does a partial count of the word like and for w 1, w 2, w 3 and so and so forth. And there are two reducers, so it creates file for both the reducer and so the reducer one is responsible for w1 and w ; whether the users two is for w3 and w4. And we do a word count on the thing, so there is a mapping function where this is done and there is a reducing function, where it is basically the function is for summation of this count for every word; w1. So, that is dividing this is what count problem into a MapReduce problem and last talk or last lecture we have shown that this can give parallel efficiency in the system.

(Refer Slide Time: 08:43)

**Problem-1**

In a MapReduce framework consider the HDFS block size is 64 MB. We have 3 files of size 64K, 65Mb and 127Mb. How many blocks will be created by Hadoop framework?

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES



So, we now look at a couple of problems; so, this is not exactly MapReduce problem, just to go for Hadoop file system or GFS; Google file system. So, if the block size is 64 MB if you remember these file systems are larger chunk block size than never natural file system. And another thing was that there is a three replica of every instance of the data. So, there is a three replica where which allows you to have a fault tolerant mode; so based on that there are read write operations done.

(Refer Slide Time: 09:30)

HDFS Block size = 64MB

3 files : 64Kb, 65Mb, 127Mb

64Kb => 1	Replicas (3)	5x3 = 15 blocks
65Mb => 2		
127Mb => $\frac{2}{5}$		

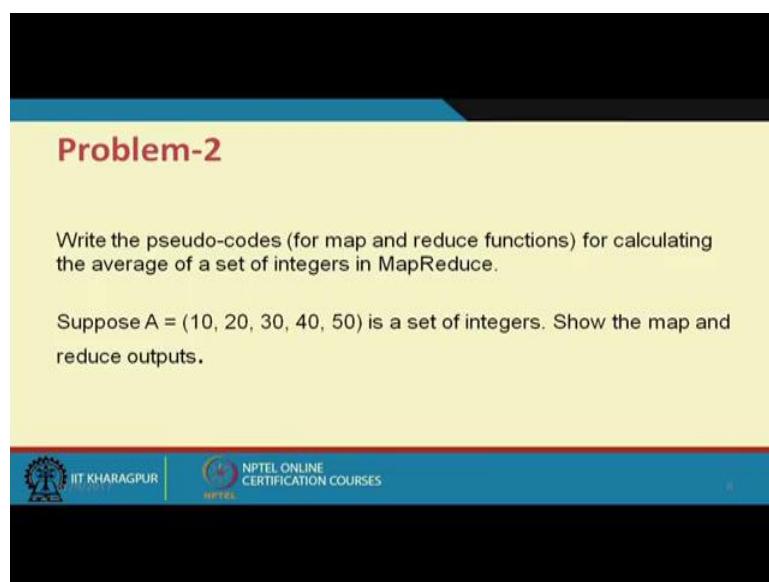


So, in this particular thing; if we say if there are; if the HDFS block size is 64 MB, then we want to find out; if there are three files of 64 K, 65 MB, Kb; MB and 127 MB.

So, how many blocks will be created by the HDFS framework. So, if for the 64 kb; how many block one will be created and 65 MB, we have 2; because upto 64 MB; 1 and 127 MB also 2. So, total 5, but in reality as there are replicas like you have replicas; so, there is typically 3 replica, so effective block size will be 5 into 3 equal to; these blocks.

So, very straightforward; nothing, no complexity in it, so if I have different type of thing. So, we can calculate this straightforward; so, again nothing to do with; immediately nothing to do with MapReduce, but nevertheless; the data is stored in either HDFS or if it is a open source or in a GFS, if it is Google file system and it need to be this data size or the storage need to be budgeted, when you are working with large data set that how much storage you require, how much storage you require to work on this type of data sets.

(Refer Slide Time: 11:29)



The slide has a yellow background with a black header and footer. The title 'Problem-2' is in red at the top left. The text asks for pseudo-code to calculate the average of integers in MapReduce, and provides an example with set A = {10, 20, 30, 40, 50}. The footer contains the IIT Kharagpur logo and the NPTEL online certification courses logo.

**Problem-2**

Write the pseudo-codes (for map and reduce functions) for calculating the average of a set of integers in MapReduce.

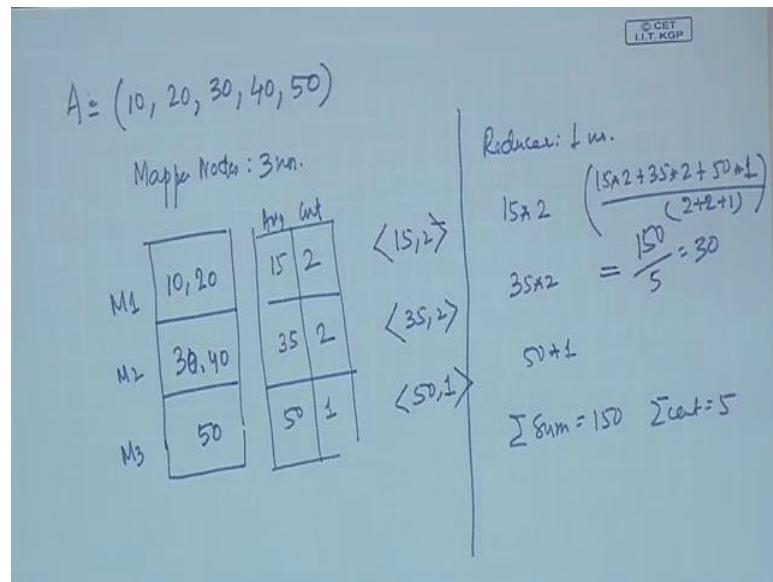
Suppose  $A = \{10, 20, 30, 40, 50\}$  is a set of integers. Show the map and reduce outputs.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Now, let us see one problem on very again very straightforward problem on MapReduce framework. So, whether we can have this MapReduce framework; though again we may not very much appreciate directly because of this simplicity of the problem. But to understand the MapReduce framework, it may be good. So you want to write the pseudo codes or codes in any language that; where there are; what we want to do? Calculate the average of a set of integers in MapReduce. So, a set of integers being pumped into the

system, it may be a direct input from the keyboard or something and so we want to find out the average of the set of integer. In other sense, in this typical case I have set of integer A as; 10, 20, 30, 40, 50.

(Refer Slide Time: 12:22)



So, set of integers are there so I want to make a average. So, in other sense we want to basically sum it up and divide by the cardinality. So, totally divided by 5. So in this case what we do? The master node say we consider there are three mapper. So, there are mapper nodes we considered as 3 numbers and a reducer say 1 number.

So, what do in a master node what it does for this mapper; it divides into say M1, M2, M3; 3 mapper node. So, a portion of this data say it gives 10, 20 to the first one; 30, 40 and 50. So, each mapper does a partial counting of the things; it does a averaging of these two things. So, it is something which comes up as; I can say average and count. So, first one is 15, 2; then this is 35 cardinalities 2 and this is 50; 1.

In other sense; in the temporary local file system is store 15; 2, 35; 2. What basically the output of the mapper; it does by the combiner. So, this is the map functions wants to achieve.

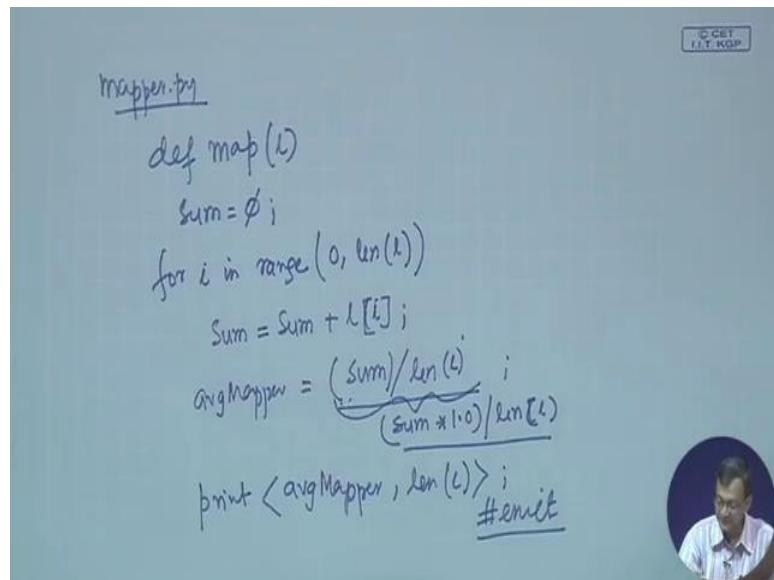
On the reduce; reduce is primarily is there is a one reducer it takes all the things and it does a averaging of the things or more this averaging of the whole thing. So, in other

sense what it does say 15 star 2; here 35 star 2 and 50 star 1. In other sense is sigma sum is 150; sigma count is 5.

So, it says 150 by 5 are 30; what it is basically 15 star 2 plus 35 star 2; so, that the exactly does. So, the problem is pretty straightforward or simple you may not find that; what is the big deal here? If the number of things is pretty high coming as a stream, and then I can basically do a parallel things; these are parallely processed and this is reduced by the one particular reducer.

So, if we write the code for it; so, you can use any language to do that. Here we are using say python or python type language, it is the language is does not matter; the representation you can do use any pseudo code and type of things.

(Refer Slide Time: 16:21)



The image shows a handwritten Python script titled "mapper.py". The code defines a function "map" that takes a list "l" as input. It initializes a variable "sum" to 0. Then, it iterates through the list from index 0 to len(l)-1, adding each element to "sum". After the loop, it calculates the average "avgMapper" as (sum / len(l)). Finally, it prints the pair <avgMapper, len(l)> followed by "#emit". A small circular portrait of a man is visible in the bottom right corner of the slide.

```
mapper.py
def map(l)
 sum = 0;
 for i in range(0, len(l))
 sum = sum + l[i];
 avgMapper = (sum / len(l));
 print <avgMapper, len(l)>; #emit
```

So, you have that mapper function or say mapper dot py; so, it is something python type. So, we are not much giving importance to the syntax; rather we are more giving importance to the concept.

So, def map; so l is the least; so, what we do? We initialize a sum equal to 0 for i in range 0 to length of l, sum equal to sum plus. So, every mapper does this; every mapper what it does; it takes that chunk of data which is being allocated to it for by the master node and then in our cases how many data is there?

The mapper 1 has 2 data; mapper 2 has 2 data mapper 3; M 3 has 1 data; so, it is 2; 2 each. So, for every mapper we done this; so, we that average sum by length of 1 in this case 2; it should be like this sum by length of 2. Or in other more strictly speaking, we should; to have a floating point difference, floating point divide because this is a; otherwise there may be integer divisions.

So, we can ideally give some star 1.0 divided by length of 1. So, length of 1; in this case 2 and then we output this; let us use print function, output this to the local file system. So, there can be this command wise; it may be different if you are using different programming paradigm lengths; so, the mapper basically emit these data.

So, it is stored in the local file system; so, what we are doing, for every mapper we are reading a list of data which is being assigned by its master node, making that some initializing the sum to value, then we add what we are doing for i; for in the loop. So, calculating the sum making a average out of it rather this is nothing, but to make it float division. And then it is emitting or dumping that value into the local file systems; which the reducer will read it.

So, this is the mapper portion of the thing; so, if we look at the reducer portion what we have def reduce. So, what it reads; so, whatever the mapper has dumped in the things. So, if you look at it is giving these average value and the lengths of the thing.

(Refer Slide Time: 20:14)

Reducer by

```

def reduce (avg, l)
 sum = count = 0;
 for i in range (0, len(l)):
 sum = sum + avg[i] + l[i]
 count = count + l[i]
 average = sum / len(l)
 print <average>;

```



Here also we read that particular thing sum equal to; so, here for i in range of 0; length of 1. So, it is what it is doing if we look at our previous thing; so, what it is doing, it is basically trying to calculate this sort of values. So, count also equal to count plus length of i; so, finally what we get average. Again sum to make it float; yes multiply this or you can basically typecast also count and then print average it.

So, what it is reducer is doing taking this local file system as there are only one reducer. So, it takes all the values and for all that data; it goes on summing up that output from the each mapper, in this case there are 3. So, it is coming to be 15 into 2; plus 35 into 2 plus 50 into 1; divided by count, which is here 2 plus 2 plus 1 is 5. So, and it calculates the average value and then it again writes the average value to the Google file system or Hadoop file system based on the whatever the requirement is this.

So, here this is that again though this is may be a straightforward simple thing, but we see that I can divide a problem; as there are inherent parallelism like there are; I could have like in order to do a averaging I have taken a chunk of data and that we try to solve it using in MapReduce framework.

So, if there is a huge volume of data then the mapper; that the master node divides accordingly and do the partial computation and the reducer read it from and do the final computation. So, this is again a simple example of a MapReduce framework; so, next we see another problem.

(Refer Slide Time: 23:41)

**Problem-3**

Compute total and average *Salary* of organization XYZ and group by *Gender* (male or female) using MapReduce. The input is as follows

*Name, Gender, Salary*

John, M, 10,000  
Martha, F, 15,000

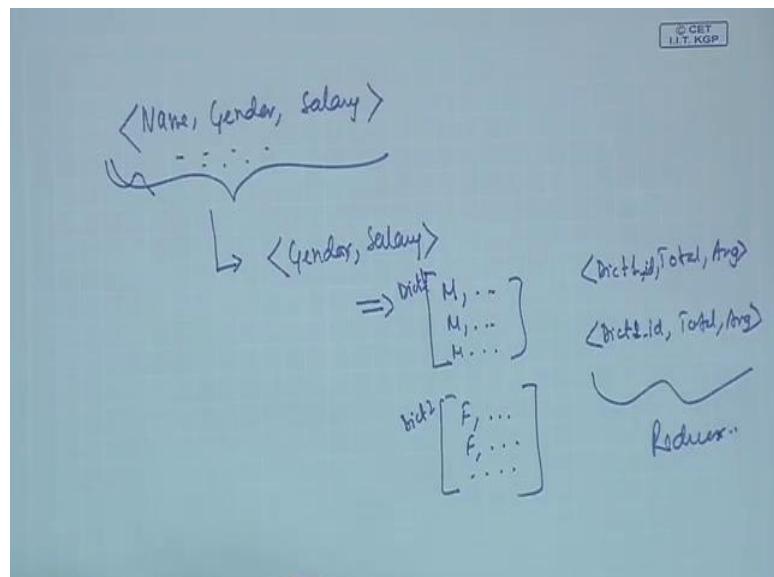
----

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, what it says I want to compute the total and average salary of organization XYZ; some organization grouped by the gender using MapReduce. So, input is name, gender and the salary of the thing; in this case say name is John, gender is M or male and salary is something 10000 unit or maybe 10000 dollar or something. And the next one is Martha, gender is F and salary is something 15000.

So, what we want to do? We want to find out that male wise, like gender wise; in this case male and female that what is the your total and average salary; anyway that total divided by the cardinality will be the average salary of the things; so, the output will be like in that form. So, what we try to look at whether we can employ MapReduce problem to look into this particular problem. So, let us look at it.

(Refer Slide Time: 24:53)



So, what we are having? We are having this tuple like this; name gender and salary. So, this is the tuple; so, what we want to do in the map phase? So, want to if the input data whatever the input data say it is there; there are different set of the input data set. So, what we want to do? We want to extract only because we are not bothered about the gender M; bothered about the name of the person or that is not required in the; so, in the salary.

So, we want to calculate from there these two thing say M the salary; so, respective and so, they want to do type of this type of things or I can have a key value pair. Key is this; male or female and value is the salary of the things or we can say that we have two sort

of a dictionary structure which is having a key value pair and then having say; I say this is Dict 1, Dict 2 and these two type of key value pair and for every 1, I can have Dict 1 then maybe total an average for other one also Dict 2; id maybe.

So, so id in this case is this particular male or female and then having total or average salary. So, compute total average for the two separate say we consider there is a dictionary structure and the reducer basically does this thing. So, want to see that how to realize this, so what let us look at the problem.

So, I have a set of name, gender and salary; want to extract that gender and the salary from every topple. So, if I have multiple mapper; so, I extract those things and dump as a particular two type of dictionary type. One is that two type of thing; one is that what we have this with M and f and the at the reducer part, we calculate the total and average salary of the things or any of the thing.

(Refer Slide Time: 28:01)

```
Mapper.py
import sys
for line in sys.stdin:
 line = line.strip()
 line = line.split(" ")
 name = line[0]
 gender = line[1]
 sal = line[2]
 print '%s,%s,%s' % (name, gender, sal)
```

So, again we look at as a mapper dot py or mapper dot some python type code. So, again I am just want to again repeat it; that you can do with any coding language which is suitable for this and whatever we are doing may not be; there may be some syntactical problem with the actual python thing, but it does not matter the conceptually you want to show that things works and then actual syntactical syntax need to be followed, if you are really want to implement this.

So, this is the thing for a line in sys dot std in; what we are doing. So, what we are considering that it is separated by a comma. So, we are split is; sorry it should be comma. So, I now separate name equal to line 0, salary represents Sal is; line 2 while generating the; or emitting the mapper phase, the data into the local file system. So, we keep print comma percentage d.

Then, so what we do? Gender and salary; in other sense this syntax you need to check up. In other sense, what we do we dump basically gender in the salary into the thing or M and the salary portion; like as you see we want to generate this M and the salary portion and another thing is that either M or f and the salary portion.

(Refer Slide Time: 31:22)

```

Reducer.py
import sys
dictOrg = {}

for line in sys.stdin:
 line = line.strip()
 gender = line[0]
 Sal = line[1]
 if gender in dictOrg:
 dictOrg[gender].append(int(Sal))
 else:
 dictOrg[gender] = []
 dictOrg[gender].append(int(Sal))

print('M.S, M.d, M.t, F.S, F.d, F.t')

```

So, in the reducer phase; what we do that import; so, define that or call this dict org that is dictionary class for line in sys in what we do. So, what it is reading? It is reading basically that the gender or that key value pair with the gender and the value of the things or in other sense the gender and the salary values. So, we do not have that name into the things because this particular query does not require the name of the thing; line of 1.

So, if it is already existing; that means, once you have read then dict org; so, what is our objective? So basically sum up the salaries by adding go on adding on the salary values for the same gender type. So, already if there; that means; so, what that now my existing the reducer dictionary counting, a key value pair. So, if the key is already that gender is

there or male or female, then I go on adding those things whenever I get. If it is not there if it is else; that means, this is basically the initialization thing dict org.

So, initialize with a blank thing; so first time when it is coming. So, in first time it is coming; that means, there is a blank thing. So, if it is blank then it basically initialize; then append the salary; that means, it is initialized with the salary of the dict org dot keys; salary average equal to sum of dict org; gender divided by length of dict org gender. So, it is summing up divided by things straightforward and total salary equal to only sum of dict org gender. And then we basically write back the other thing from to the Google or GFA or the HDFS file system. If we want to separate it by a comma or tab as the case may be; again maybe D if it is a integer or based on that if it is a float and all those things.

So, we have this as gender total sal and salary avg. So, that is the what we do at the final reduce surface. So, if we try to just quickly have a look, so what we are doing in the mapping function; we have three thing like name, gender and salary. Our objective is to the mapping functions; so, see this all the map are we like find out this individually this whether it is a; which gender M and find keep this salaries along with that g M or f and salary and the reducer will basically; so, that it exactly that gender and salary and the reducer will basically extract that intermediate result and calculate the average and the total.

So, here that that operation is there; so, this is a typical python type, I am not strictly telling python because there may be some syntactical issue. But you can implement in anything, the idea is that I divide the problem into smaller parallel things by the mapper and then in the second phase; the reducer put it to another key value pair. So, key value from the input set; to a set of key value pair, reducer takes that key value pair and put a function; in this case average or total of the things, to another set of key value pair and the finally, it goes to the HDFS pair or GFS file system ok.

So, what we tried to look at in today's thing that; this MapReduce functions say simple problems; how we can put it into map and reduce things. That this number of mappers available; allow is basically availability of the resource and the how the master nodes divide it and the number of reducers also based on the term. What type of functional

things you want to do and so they master node is there, it divides into M number of mapper and a number of reducer.

The problem is the functionality of the problem is divided in such a way so that it can be executed in two phases, and we can have a parallel implementation of this sort of paradigm.

Thank you.

**Cloud Computing**  
**Prof. Soumya Kanti Ghosh**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**

**Lecture - 24**  
**Resource Management - I**

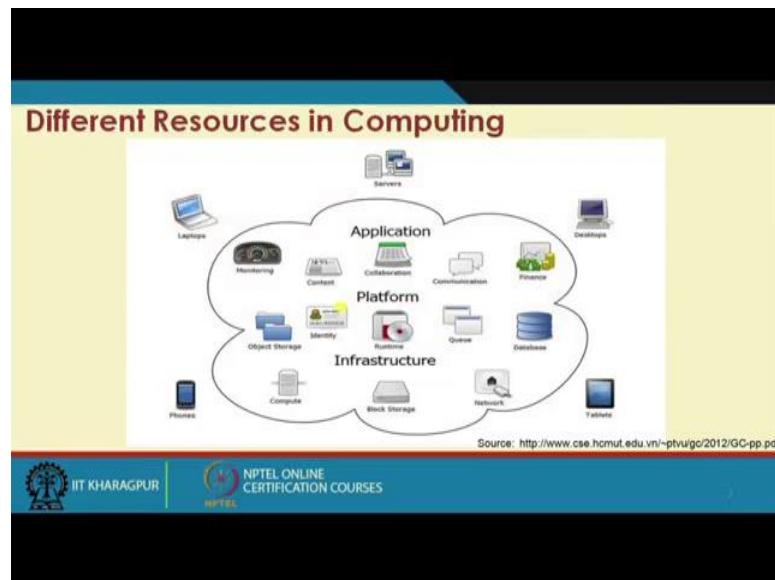
Hello. Today we will discuss one of the important aspects of a Cloud Computing or any type of this sort of service oriented mechanisms like cloud or distributed system or agreed or any type of systems that is the Resource Management. So, what we are thinking of a cloud? We are thinking that cloud is a infinite resource pool; along with that we are basically leveraging these resources to multiple users.

So, from the customer point of view or the user point of view or the consumer point of view; that it is a something a scalable service and a consumer can get as many as resources based on something of the concept of pay as you go or metered services. So, from the provider end; it is important to properly utilize the resource in both terms that it can serve the consumer in a better way and in other sense also, which is limited or what we say finite amount of resource; it can serve as many as consumer or the customer.

The thing is that from the providers there is a business angle also; he need to optimize the or it need to maximize his profit without compromising on the quality of services or the SLA violations. As we have seen, if there are violation of SLA's; it has a long implication, there has to give penalty and so and so forth. There is another aspect of the whole thing; even if a particular provider has say money or can afford number of resources, but there is a limitation or there is a overall environmental aspects of it; how much of your carbon footprint, how much energy resource you are having.

So, there is a restriction or there is a obligation on that mind of. So, keeping all these thing into mind, this resource management plays a important role in the thing. So, what we will do in a couple of lectures; we will look at that what are the different aspects of a resource management in the context of cloud computing.

(Refer Slide Time: 03:02)



So, if we look at different type of resources in a typical cloud computing things. So, what we have as we have mentioned; we have infrastructure, so basic infrastructure where is a compute resource, storage resource, networking resource or a related resource on the things. There is a platform which is above the infrastructure or here if I say that infrastructure as a service is the major thing, then we have a platform as a service.

And then at the end we have the application as the resource. So, the resource can manifest in a different way; it can be the infrastructure. Though primarily when we talk about resource or resource management, we usually fall back to infrastructure as a service, but considering the overall cloud or overall operation of the cloud; so, basically at least we are having infrastructure as a service, platform as a service or software as a service, also we have infrastructure as a resource management thing or platform as a resource management thing and application but though prominently a infrastructure plays a bigger role.

So, these are the different type of resources and the customer basically hook into the cloud with different type of heterogeneous systems. It can be servers, it can be laptops; maybe a smart phones, tablets, desktop and so and so forth. Number of cases, a cloud; we can say that a cloud or a service can take the cloud service or service providers thing to leverage on other things. I provide service taking some service from a service provider and so and so forth.

So, overall this management and optimization of these overall operations is a very tricky one. So, what we try to look at today is that; what are the different aspects of the things at all; whether with the same type of a hardware another things available, is it possible to manage the resource in a efficient way. That is our basic call for this particular resource management.

(Refer Slide Time: 05:27)

**Resources types**

- **Physical resource**
  - Computer, disk, database, network, scientific instruments.
- **Logical resource**
  - Execution, monitoring, communicate application .

Source: <http://www.cse.iitm.ac.in/~ptvu/gc/2012/GC-pp.pdf>

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, as we have seen resource type can be physical resources like as we see that computer, disk, memory databases, network, scientific instrument. So, database is again it is a soft resource so what we say that integrated this; means that hardware systems meant for the databases and also. There are other logical resources like what we say; there are applications like communication applications and other applications; executions of some processes which execute on the CPU's.

Monitoring and management of the things, so these are monitoring tools, management tools etcetera; these are different other resources. Some resources are directly utilized; like if I say that CPU and a hard disk and other things; some are basically meant to manage those resources.

(Refer Slide Time: 06:26)

**Resources Management**

- The term **resource management** refers to the operations used to control how capabilities provided by Cloud resources and services can be made available to other entities, whether users, applications, services in an *efficient* manner.

Source: <http://www.cse.iitm.ac.in/~p...pdf>

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES



So, if you look at in a broad term like; what we are trying or what if we say the resource management; what we want to do with a resource management, the term resource management refers to the operations used to control how capabilities provided by the Cloud resources and services; there is a typo can be made available to other entities, whether users, applications, services in an efficient way.

So, what we want to do that whatever the resources and services; that means, hard and soft resources are available with the cloud; how it can be made available to the external entities like; it may be a user, it may be an application, it may be a service; in an efficient way. So, there is quote unquote the term efficient is tricky, but efficient means it can be efficient in maximizing the profit of the ISP, it can be efficient in energy optimization or in efficient in giving or respecting SLA's and providing best quality of services; nevertheless the resource management plays a role in all those things.

(Refer Slide Time: 07:44)

**Resources Management**

- The term **resource management** refers to the operations used to control how capabilities provided by Cloud resources and services can be made available to other entities, whether users, applications, services in an *efficient* manner.

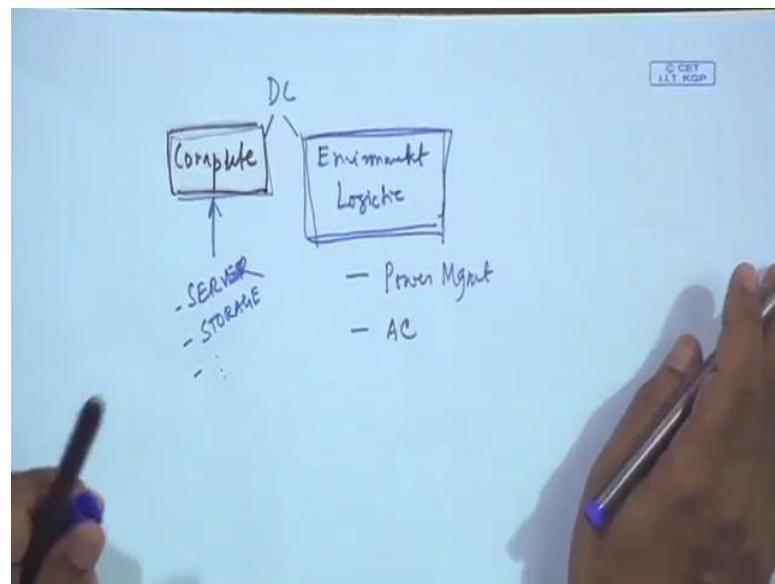
Source: <http://www.cse.iitm.ac.in/~ptvugc/2012/GC-pp.pdf>

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Now, if we look at data center a power consumption. So, what we are referring at? That is a ISP or a internet is or a CSP or a cloud service provider. So, they are having data centers; like major providers like IBM, Google, yahoo or; so, there are several things like Microsoft, Amazon and so and so forth. They have a huge data centers across the world, as we have seen.

So, based on the things that these resources are outsourced from this datacenter to be more specific; so, what is the power thing? So, there is some report; it may not be the very recent report; it is much more higher than the whatever we are talking about. So, it is estimated that servers consumed around 0.5 percent of the world's total electricity usage; so, it is a huge amount. Now, if it is not only the server; if you look at a data center.

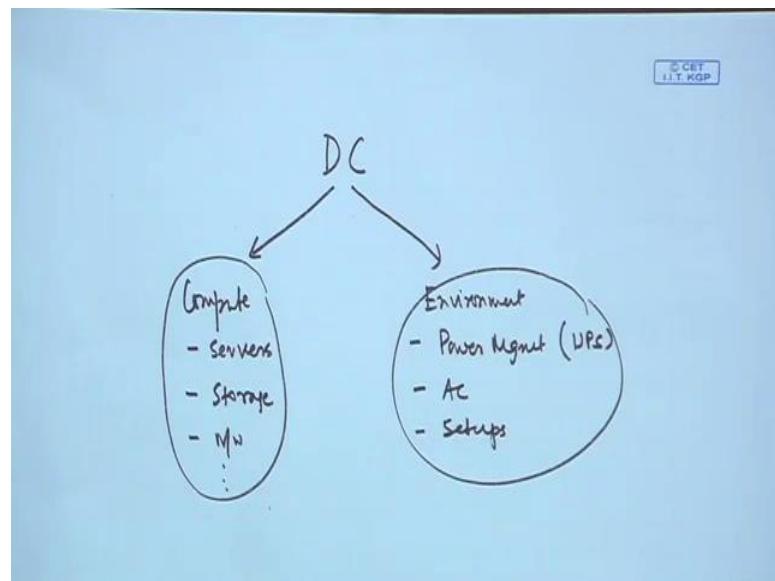
(Refer Slide Time: 08:56)



So, data center is; if you look at it, there are two major component; one is that compute infrastructure, another is what we say overall environmental infrastructure or data center other logistics; what we say; Environment or Logistics of the data center.

So, what we want to do? These basically gives you say servers or say storage and maybe adjoining thing; like network etcetera. These are providing primary looking for your power management or power supply or air conditioning and other related stuff which is related to these data center things.

(Refer Slide Time: 10:26)



So, if I have a DC; so, it has two component; incidentally this two have; if not equally power hungry things. So, if you look at the amount of power; so, if we make it. So, we have one; if we have DC; one component is towards compute, one component is towards maintaining the overall environment of the data center.

So, here what we having a servers, storage network and other other, whereas, here primarily your power management; maybe UPS; sort of power, then AC and other overall logistics arrangement or other means; different type of setups and it is said that to have the house the data centers. Now, there is a power requirement out here; there is a overall power requirement out here. So, if you look at; and not only that this is not only power, there is a power cum space, power cum space.

So, if you look at the overall things are; if I say the overall consumption here is x. So, it is also somewhat towards x. So, if it is not only the servers, but the consumption by the other type of; other logistic units are equally high. So, when we talk about data centers where it is 0.5 percent of the world total uses.

(Refer Slide Time: 12:00)

**Data Center Power Consumption**

- Currently it is estimated that servers consume 0.5% of the world's total electricity usage.
- Server energy demand doubles every 5-6 years.
- This results in large amounts of CO<sub>2</sub> produced by burning fossil fuels.
- Need to reduce the energy used with minimal performance impact.

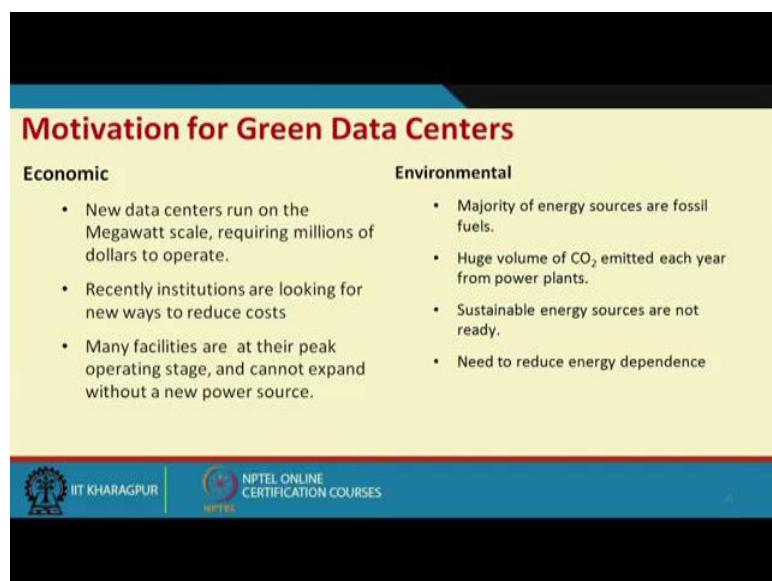
Ref: Efficient Resource Management for Cloud Computing Environments, by Andrew J. Younge, Gregor von Laszewski, Lizhe Wang, Sonia Lopez-Alarcon, Warren Carithers,

IT Kharagpur | NPTEL ONLINE CERTIFICATION COURSES

So, if you look at; if you basically make provision for other things, it may go above 1 percent of the things over power. And this is some data which is reported in some places, but it may be much higher.

And it is also saying these servers' demands when some report says that energy demand for the servers are increasing every 5 to 6 hour; a 6 years, maybe becoming double. There is a large volume of carbon dioxide footprint, which because of this fuel; burning of this fossil fuels. Need to reduce energy use for the minimal a performance, so what is our goal that we need to reduce this energy utilization with minimal effect or the compromising this performance. So, that should not be practically or that should not be very minimal impact on the performance; so, that is one of the goal.

(Refer Slide Time: 13:05)



The slide has a blue header bar with the title 'Motivation for Green Data Centers' in red. Below the header, there are two columns: 'Economic' on the left and 'Environmental' on the right, each containing a bulleted list of reasons. At the bottom, there is a footer bar with the IIT Kharagpur logo and the text 'NPTEL ONLINE CERTIFICATION COURSES'.

Economic	Environmental
<ul style="list-style-type: none"><li>New data centers run on the Megawatt scale, requiring millions of dollars to operate.</li><li>Recently institutions are looking for new ways to reduce costs</li><li>Many facilities are at their peak operating stage, and cannot expand without a new power source.</li></ul>	<ul style="list-style-type: none"><li>Majority of energy sources are fossil fuels.</li><li>Huge volume of CO<sub>2</sub> emitted each year from power plants.</li><li>Sustainable energy sources are not ready.</li><li>Need to reduce energy dependence</li></ul>

So, as you are discussing; so, the overall motivation if we say that one to make quote, unquote green data centers; which will take very less energy; perform at the highest level. So, there are different aspects, so one is this economic aspect; the new data centers run on megawatt scale; require millions of dollars or millions of money to operate.

Recently institutional looking for new ways to reduce different type of new ways to reduce cost, many facilities are there. Peak operating stage cannot explained without a new power source in some cases; like if I have; even that is why I was talking that even if I am having a fund available as a provider, but I do not have any resource to do that.

Suppose we have a data centers at IIT Kharagpur and we are taking power from the state electricity board. And even tomorrow, we say that we want to scale nothing to double and we are ready to pay, but the state electricity board may not have that type of surplus

power to supply. So, it is not only that whether I am having money or having more space etcetera, but whether that can be supplied.

There is a definitely environmental aspects and there are different type of legal obligations towards environment comes into play. Majority of energy sources are fossil fuels still to date. Huge volume of carbon dioxide emitted each year from power plants; which creates a environmental hazards.

Sustainable energy sources are not; till not ready to handle such huge requirement of the data centers. Need to reduce energy dependency; so, we need to have some mechanism to reduce the energy dependency. So, this is the two very very broad aspects of the thing.

(Refer Slide Time: 14:57)

**Green Computing ?**

- Advanced scheduling schemas to reduce energy consumption.
  - Power aware
  - Thermal aware
- Performance/Watt is not following Moore's law.
- Data center designs to reduce Power Usage Effectiveness.
  - Cooling systems
  - Rack design

IT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, whether we are talking about green computing in this respect definitely. So, what we can do? What cloud providers tries to look at; is that whether this advanced scheduling schemes to reduce energy consumption is possible so, that; which can be power aware or an ender or thermal aware; both power aware and thermal aware, this sort of things.

Performance per watt is not follows the Moore's law; like our famous Moore's law. But it is unlikely that performance per watt is not following the Moore's law. So, data center design so that whether we can redesign or basically design the data center to reduce power uses effectiveness. So, that we can have effectively use the power like it can be cooling system or the rack design or the placement of the racks and type of things that is

more of the data center design; as we are talking just a couple of minutes back. That one is that having the server etcetera compute part of it, another is that; the data center design aspects of it that how efficiently you can design.

And if those who are visited some sort of data center; you know that there are different typical arrangement of the cooling, it is not conventional cooling of the may not be the whole server room or the space; it is a separate enclosure made on the data on the major server racks and the server racks are cooled specifically on a very confined area so that the energy required for this cooling purpose can be minimized.

(Refer Slide Time: 16:44)

**Research Directions**

How to conserve energy within a Cloud environment.

- Schedule VMs to conserve energy.
- Management of both VMs and underlying infrastructure.
- Minimize operating inefficiencies for non-essential tasks.
- Optimize data center design.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

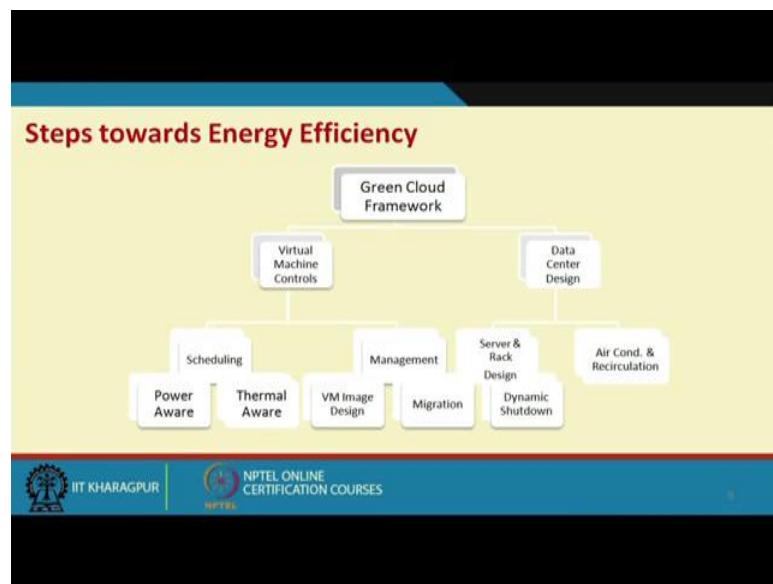
So, there can be different type of research direction and some of you; who are interested maybe doing some research on the cloud computing. So, there is a important this resource management looking at different aspects is a extremely what we say hot topic to do research, at different parts of the world people are working on this; that how to optimize this resource, how to have a optimal resource utilization for; without compromising services or without compromising the performance of the cloud.

So, how to conserve energy within a cloud environment? So one may be scheduled VMs to conserve energy, whether we can have proper VM scheduling on the conserve energy. So, management of both VMs and underlining infrastructure, minimize operating inefficiencies for non-essential tasks; so, that whether we can minimize that operational efficiencies or non essential task. Like if you take a VM with a particular flavor and lot

of resource; a lot of other packages etcetera running, but practically you do not require all those things.

Whether we can have a tailor made VM or tailor made OS; which can be run on the thing. Optimize data center design; the other aspects. Now these are; so, one is aspect of scheduling the VM; one aspect of is management; another aspect is design, design of the data centers and so and so forth.

(Refer Slide Time: 18:32)



So, these are three broad aspects; we will try to have a quick look on these aspect so that we can have a feel of it. So, if you look at this overall green cloud framework; so one side is towards VM controls, how these VM scheduling management, other side is data center design.

So, this is more on the compute side of it; this is more of the data center infrastructure side of it. So, here the infrastructure is the hard data center infrastructure; the other come. So, if you look at the VM controls; so, one part is scheduling the view. So, which can be power aware or thermal aware; looking at the things. Other is that VM management, which is VM image designing; whatever the VM image having, whether we can have a efficient design of this VM image with basically only those packages or those services which are required are uploaded there.

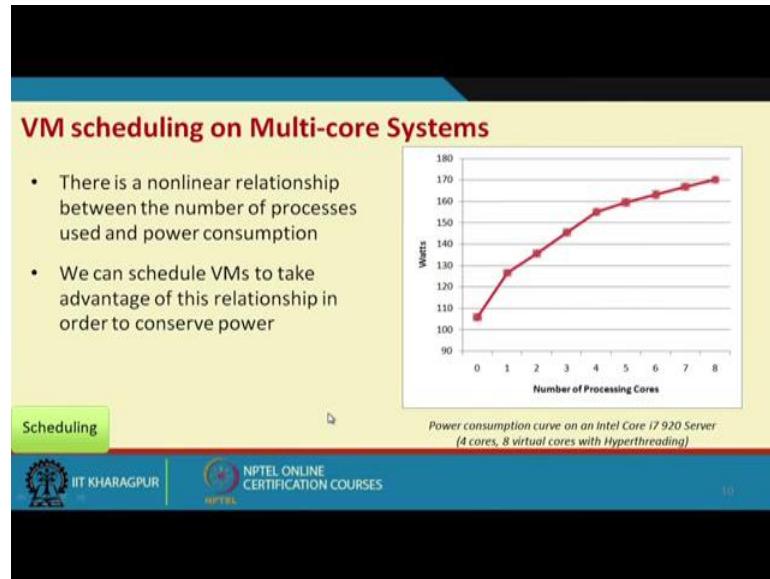
So, other is that whether you can look for VM migrations; so, what we are having actually if you look at it. So, what we are having a underlining hardware and there VMs are there. Suppose there are typically say 16 blade server; so, every blade server is running say 8 VMs. So, 16 into 8 is the total number of VMs are there, so now out of the 16 into 8 or say 10 VM.

So, 16 into 10 VMs are running or say 10 maybe on the higher side, so 16 into 5 say 80 VMs are running. Now 80 VMs; if there are at any point of sign; say 10 or 20 VMs are active, whether it is efficient to distribute over the whole VM servers available or I can concentrate the VMs into couple of servers.

So, because one is that; whether the server, VM is running on a server or not; it will go on consuming a base energy. Whenever I run a VM; it may consume some incremental energy. So based on that; it requires some sort of a simple mathematics to work out that whether it is efficient; so, one the data center design side, one is that server rack design is a important aspects and there is separately if you look at do not the cloud computing part; this is another research on the building of data center or infrastructure type of things. We are conditioning recirculation is a type of another aspects of the things.

So, there are if you look at that management side; there is another aspect of dynamics shutdown if the things are not utilized; so, whether I can dynamically shutdown there. So, there are a lot of aspects and if you see these are all have lot of research potential especially this part of the things on the compute side.

(Refer Slide Time: 21:40)



So, this is a typical example scenario if you see there is a non-linear relationship between the number of processor used and the power consumption. So, it is not like that that the number of processor used in the power consumption is a linear thing. So, if the number of processor goes on high, the power consumption is basically somewhat going towards some saturation; if not saturation the increment is not linear.

So, whether we can exploit this one; whether we can schedule VMs to take advantage of this relationship in order to conserve power. So that means, as I was talking that I want to concentrate my VMs into the underutilized server so that the VMs running on the number of servers are reduced. So, that effectively I can put this idle servers into sub hibernate or mode or low power consumption mode so that the overall power consumption of my data center reduces.

(Refer Slide Time: 22:50)

**Power-aware Scheduling**

- Schedule as many VMs at once on a multi-core node.
  - Greedy scheduling algorithm
  - Keep track of cores on a given node
  - Match VM requirements with node capacity

**Algorithm 1 Power based scheduling of VMs**

```
FOR i = 1 TO i ≤ |pool| DO
 pei = num cores in pooli
END FOR

WHILE (true)
 FOR i = 1 TO i ≤ |queue| DO
 vmi = queuei
 FOR j = 1 TO j ≤ |pool| DO
 IF pej ≥ 1 THEN
 IF check capacity vmi on pej THEN
 schedule vmi on pej
 pej - 1
 END IF
 END IF
 END FOR
 END FOR
 wait for interval t
END WHILE
```

Scheduling

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, that schedule as many VMs once on a multi core node; greedy scheduling algorithms. We can do; there is a snap shot of a greedy scheduling algorithm, you can basically try out and see that how things work. Keep track of the core on a given node. So, we have to see that at a way for it typically in a node that how this codes are being busy or how much loading is there. Match VM requirements with the node capacity; if it is there then whether we can migrate the thing.

(Refer Slide Time: 23:26)



A simple very vanilla type example, so if I have 4 nodes; so, that it is in a idle condition or when a no load condition they consume 105 watt and when it is fully loaded with 8 VMs, it consumes say 170 watt. So, if I have some 8 VMs; then if we run on a 1 node, the overall consumption is 170 plus; 105 star 3.

(Refer Slide Time: 23:57)

$$170 + 105 \times 3 = 485 \text{ watts}$$
$$138 \times 4 = 552 \text{ watts}$$

So, that is the overall consumption 4 watts in this case, this typically looking at the things. If I had it been it is distributed in thing; so, with this sort of 2 VM; let us say that is 138. So, I have 4; so, in this case 552 watts; so, that exactly what want to see that in doing so, we can basically reduce the things.

But there is a little catch in it; now in order to achieve from this stage to this stage, I need to migrate this; migration also have some cost. Migration has some cost, not only that here we are taking all VMs to be the same category. So, VM can be of different categories; so, it may not be able to put all the VMs into a 1 node or 2 nodes etcetera. So, we need to check and monitor that how much loading is there, how much free VMs are there so that we can basically migrate the VMs.

So, it is not like straightforward of multiplying that to wattage; it is also this migration has some cost on doing that.

(Refer Slide Time: 25:18)

**VM Management**

- Monitor Cloud usage and load.
- When load decreases:
  - Live migrate VMs to more utilized nodes.
  - Shutdown unused nodes.
- When load increases:
  - Use WOL to start up waiting nodes.
  - Schedule new VMs to new nodes.

Management

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, what we need to do? Monitor cloud usage and load. When load decreases; live migrate VMs to more utilized nodes; unused node. Shutdown unused node so, that I can basically what we are doing? We are compacting those things into the server things, so that is the live migration. Now, number of these your hypervisor or say number of VMM do support this type of live migration, though it maybe commercially costly but you can do a live migration and it has lot of implication other way also.

Like if a server is having problem, then you live migrate to the other servers and so on, but that is those are things are possible. So, when load increases; so, we can have some sort of wakeup call; wake up on lan sort of things to start the waiting node. So, in the load increases; I basically give a wakeup call to the nodes. So, which are sleeping to wake up, schedule new VMs to the new nodes. So, there are technologies available and if we can be efficiently use; effectively we can have a energy efficient thing.

(Refer Slide Time: 26:31)



So, that it is node 2, it is migrated there; the VM is put into service, then the node 2 become idle, then made into hibernate or offline mode; so, effectively I am running on a node 1. So, it is this; though there is a cost of this sort of a migration, but overall if I am achieving efficiency or in terms of power consumption, then that maybe a good option to look at.

(Refer Slide Time: 27:03)

**Minimizing VM Instances**

- Virtual machines are loaded!
  - Lots of unwanted packages.
  - Unneeded services.
- Are multi-application oriented, not service oriented.
  - Clouds are based off of a Service Oriented Architecture.
- Need a custom lightweight Linux VM for service oriented science.
- Need to keep VM image as small as possible to reduce network latency.

Management

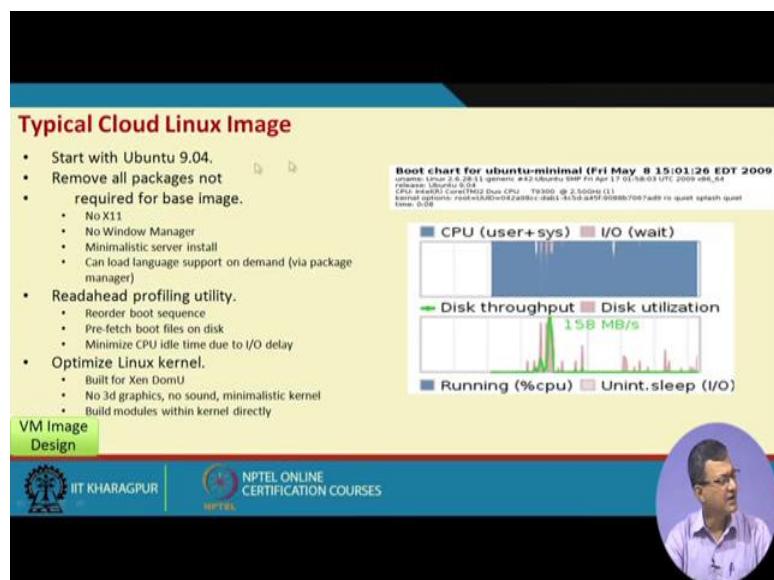
At the bottom, there are logos for IIT Kharagpur and NPTEL Online Certification Courses.

So, another aspects what we have seen that management; why is that? Whether I can minimize VM instances? VM machines are sometimes too loaded, so lots of unwanted

packages, unneeded services are there. So, you take some while VM it comes with a basic configuration; there which may not be utilized by the consumer. So, our multi application behaves on multi application oriented, not service oriented in number of cases, clouds are based on of a service oriented architecture.

So, we should have service orientation; need to customize lightweight Linux VM for service oriented science or services, so that we can need to have a customized Linux VM or the customized VM per say. Need to keep VM image as much as possible to reduce a network latency. So, if we have the VM image less then the; means while carrying of the network, the latency will be minimized. So, these are different aspects; so these are typical example scenario which starts with a Ubuntu; typically 9.04.

(Refer Slide Time: 28:10)



Remove all packages which are not required; like if I do not require X11 or windows manager etcetera etcetera. Read ahead profiling utility, reorder the boot sequence so that I can have a prefetch boot files on the disks minimize CPU idle time due to IO delays etcetera.

So, I can have a read ahead things like if I have that the steps to be followed, I can do a; a priori thing. Optimize Linux kernel; build on say in this typical case DomU. So, there is no 3D graphics, no sound so that I can have a customized Linux kernel which is primarily used for the things. So, based on the type of customer requirement or based on the services, we can basically optimize this type of stuff.

(Refer Slide Time: 29:05)

**Energy Savings**

- Reduced boot times from 38 seconds to just **8** seconds.
  - 30 seconds @ 250Watts is 2.08wh or .002kwh.
- In a small Cloud where 100 images are created every hour.
  - Saves .2kwh of operation @ 15.2c per kwh.
  - At 15.2c per kwh this saves \$262.65 every year.
- In a production Cloud where 1000 images are created every minute.
  - Saves 120kwh less every hour.
  - At 15.2c per kwh this saves over 1 million dollars every year.
- Image size from 4GB to 635MB.
  - Reduces time to perform live-migration.
  - Can do better.

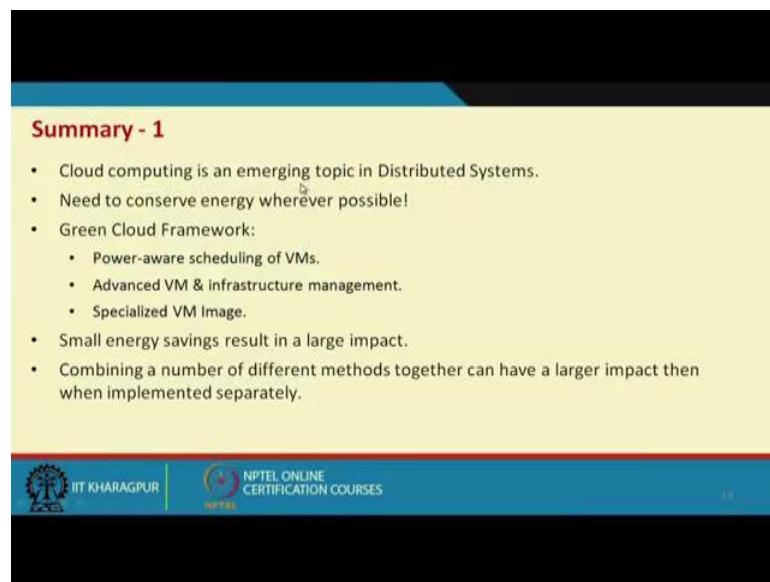
VM Image Design

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, there are different energy saving some snaps parameters like we can; if it is reduced the boot time from 38 to 8 seconds; so, effectively we can have energy savings; even if a small cloud where 100 images per hour, so much energy savings can be there.

In a production cloud where 1000 images are created every time; saves a lot of energy in other way. A image size from 4GB to something 600 plus MB; so, reduces life performance; as we are telling that the life migration also comes with a cost and we can do better. So, what it is trying to say that all those things makes its energy efficient or have a proper resource manager without compromising the actual performance. So, if we try to summarize, so it is a emerging topic definitely.

(Refer Slide Time: 30:05)



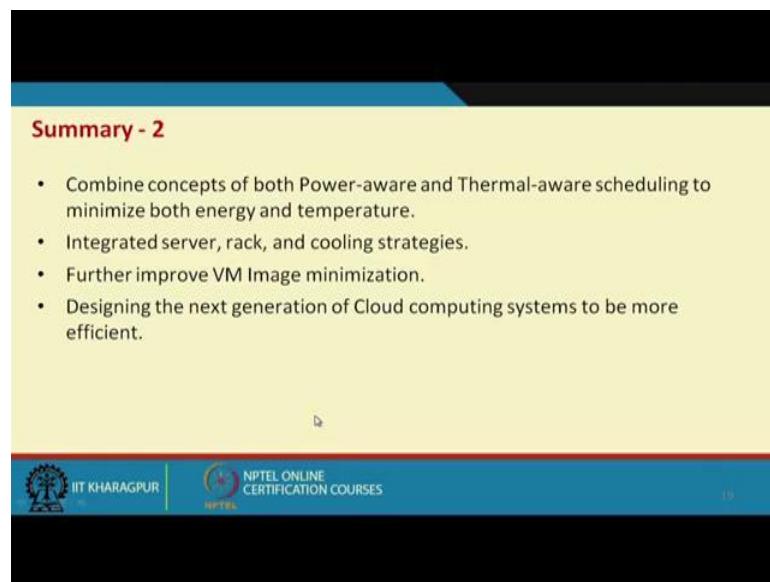
**Summary - 1**

- Cloud computing is an emerging topic in Distributed Systems.
- Need to conserve energy wherever possible!
- Green Cloud Framework:
  - Power-aware scheduling of VMs.
  - Advanced VM & infrastructure management.
  - Specialized VM Image.
- Small energy savings result in a large impact.
- Combining a number of different methods together can have a larger impact than when implemented separately.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, need to conserve energy wherever possible; so, one is that green cloud framework, power aware scheduling of VMs. Advanced VM and infrastructure management, specialized VM images or customized VMs images. So, small energy savings result in a large impact, so even it is apparently VM wise of node wise small, but considering the whole thing is a large impact. Combining a number of different methods together can have a larger impact than when implemented separately. So, it is not only piecewise having a collective or cooperative way of looking at it.

(Refer Slide Time: 30:46)



**Summary - 2**

- Combine concepts of both Power-aware and Thermal-aware scheduling to minimize both energy and temperature.
- Integrated server, rack, and cooling strategies.
- Further improve VM Image minimization.
- Designing the next generation of Cloud computing systems to be more efficient.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

And there are lot of research interests and things combined concepts of Power-aware and thermal aware scheduling to minimize both energy and temperature. Integrated server, rack, cooling strategies; these days as our server racks are coming with integrated cooling power and so and so forth.

Further improved VM image minimizes and looking at that customer needs and type of things, customer profiling. Designing next generation cloud computing system to be more efficient; both in terms of energy, thermal and type of service provided.

So, with this let us conclude today and as we understand, there is a lot of scope of doing research and for the studies in this particular way of resource management.

Thank you.

**Cloud Computing**  
**Prof. Soumya Kanti Ghosh**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**

**Lecture - 25**  
**Resource Management – II**

Hello. So, we will continue our discussion on Resource Management in Cloud. So, as we discussed last lecture or last 2 lectures on the resource management; what we have tried to look at that it plays a important role in overall cloud service, right. So, it is important not only from the service provider point of view; it is also important for the service consumer point of view, right.

So, provider want to have maximize its utilization of his resources with minimal energy cost and maximizing it profit right if you look at the from the consumer point of view. It wants to have a guarantee or a particular quality of service or support for its SLA, right. So, that the SLA is not valid. So, nevertheless this whole resource management he plays the important role for this; what we say quote unquote success of this cloud computing paradigm, right.

So, taught today; what we will try to look at some of the aspects of these resource management right we will try to look at a particular for a review paper and take up some of the aspects of resource management. So, I do not want to claim that release the all the aspects, but these are some of the important aspects where what a particular cloud computing environment or cloud computing platform. So, look at right.

(Refer Slide Time: 01:56)



So, couple of slides maybe imputation from the other. So, what we made by resources just to recap quickly. So, I have at the core infrastructure platform and application or IaaS, PaaS, SaaS and there are different kind of user across the means; user for this clouds means they can be either human user or it can be some process or machine which are indirectly consuming cloud service to the for other services. So, what we want to look at that how optimize these resources can be managed at the core.

(Refer Slide Time: 02:41)



Now as we have seen, there are 2 categories of user; physical 2 category of resources; one is physical resource, another is a logical resource. So, the physically; what is there and logically like applications monitoring and type of things; so, both plays a important role in the resource management.

(Refer Slide Time: 03:01)

**Resources Management**

- The term **resource management** refers to the operations used to control how capabilities provided by Cloud resources and services can be made available to other entities, whether users, applications, services in an *efficient* manner.

Source: <http://www.cse.iitm.ac.in/~ptvu/CloudComputing.pdf>

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

And also we seen these we have gone through these particular underlining definition of resource management it refers to operation used to control over capabilities are provided by the cloud resources and services and can be made available to other entities users applications in an efficient manner.

(Refer Slide Time: 03:25)

**Resource Management for IaaS**

- Infrastructure-as-a-Service (IaaS) is most popular cloud service
- In IaaS, cloud providers offer resources that include computers as virtual machines, raw (block) storage, firewalls, load balancers, and network devices.
- One of the major challenges in IaaS is resource management.

Source:  
<http://www.zearon.com/down/Resource%20management%20for%20Infrastructure%20as%20a%20Service%20%28IaaS%29%20in%20cloud%20computing%20A%20survey.pdf>

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Now if we look at this resource management mechanism or resource management approaches the maximum or the maximum or the major thrust is on the management of the IaaS type of resources or infrastructural resources, right other resources like platform or SaaS; though they are also management is necessary, but those are mostly dictated by the amount of underlining backbone hard resources you are having, right.

So, some of these type of techniques are applicable across the different type of services whereas, some of the things are more good to the IaaS. So, what we look at today is more about that; what are the different approaches for IaaS type of resource management, right. So, infrastructure as a service is the most considered to be most popular or seen to be most popular cloud service among these different type of services. So, in IaaS, cloud providers offers resources that include computer as virtual machines raw storage firewalls load balancer network devices and so and so forth.

So, these are the different category of things which we consider as when we talk about infrastructure as the resource and one of the major challenges in IaaS is the resource management; how to optimally manage the resource and as we have seen as energy plays a important role for overall functioning of the cloud. So, it one of the aspect is with minimal or the energy consumption how I can give service at a particular level, right.

So, that is that is very important when we talk about IaaS; when we talk about resource management. So, I want to maximize profit from the provider point of view maximize

utilization of the resources and minimal requirement of energy right and of course, on the other hand, we have; we need to satisfy these quality of services and SLA right rather there are several metrics which we will see when we discuss today, but these are the aspects we need to look at when we look at the resource management aspects.

So, we will be following or we will be taking inputs mostly from a survey paper which where the link is provided you are free to download and look at the things and there you will get lot of other corresponding paper who are interested in further research or further study on this type of resource management are welcome to look at it.

(Refer Slide Time: 06:20)

**Resource Management - Objectives**

- Scalability
- Quality of service
- Optimal utility
- Reduced overheads
- Improved throughput
- Reduced latency
- Specialized environment
- Cost effectiveness
- Simplified interface

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, if we look at the resource management, if I broad objectives or the broad goal. So, to satisfy things that my scalability of that for which that cloud is one of the properties that scalability should be preserved that I can scale up scale down and ideally infinite scalability quality of services should be preserved optimal utility reduced over its like in 1 or 2; do a resource management protocol or algorithms if the overhead is high then I loose on performance, right. So, the it should be a optimal overhead rather reduced over it and improved throughput. So, that overall throughput should be improved reduce latency. So, that it should not increase the time latency of the overall system specialized environments like whether I want to have a specialized environment in order to have say as last day we discussed as a specialized environment for rack management power and so and so forth.

Cost effectiveness that overall; it should be cost effective, it should be financially beneficial to the both provider and consumer the provider should not spend more and the consumers should not have to pay more subscription for that and it should be a simplified interface the interface should be again simple, it should not be very cumbersome interface. So, that it is or I can say that we should have a ease of use will be there. So, it is easy to use that type of environment.

So, these are our broader goal or I can say broad an objective of cloud service provisioning without compromising this whether I can have a better resource management that is the objective of the things.

(Refer Slide Time: 08:23)

The slide has a dark blue header and footer. The main content area is yellow. The title 'Resource Management – Challenges (Hardware)' is in red. The list of challenges is in black. Logos for IIT Kharagpur and NPTEL are at the bottom.

### Resource Management – Challenges (Hardware)

- CPU (central processing unit)
- Memory
- Storage
- Workstations
- Network elements
- Sensors/actuators

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Now, there are several challenges like if you look at the hardware or the bare metal or the backbone of the things the one is that CPU management, memory management, storage management, workstation, network element, sensor actuators and so and so forth. So, these are the different components which are there which need to be properly managed and it is it; these are not isolated things right like CPU memory storage these are not isolated components. So, they have a Intel linking while operations on the thing. So, you cannot have a very high power CPU in low memory and type of things then the performance will not be there.

So, if the coordination between this bare metal or the backbone resources are important. So, when we manage resources we need to take care that those are preserved, right, I

cannot make a optimal management of storage without ignoring the other component of the like network component and other things a make a faster storage my network accessibility is still slow then purpose are not solved. So, that needs to be looked into.

(Refer Slide Time: 09:35)

**Resource Management – Challenges (Logical resources)**

- Operating system
- Energy
- Network throughput/bandwidth
- Load balancing mechanisms
- Information security
- Delays
- APIs/(Applications Programming Interfaces)
- Protocols

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, there are other logical resources. So, those are what we say physical hard resources there are logical resources like operating system energy management network throughput or bandwidth load balancing mechanisms information security which is coming up in a big way or which are looked into in a big way when you are leveraging lot of things on the cloud and specially your sensitive or semi sensitive information on the cloud or privacy preserving things delays that how much delay or time delays are their application programming interface or API. So, the API is whether need to be redone on new type of API has to be there and there are various protocols.

So, these are all what we say soft resources or logical resources which plays a important role and these hard resources and soft resources are not separate to each other they are intermingled rather need to be looked into in a integrated way. So, these are the different challenges or broad objectives what you look at.

(Refer Slide Time: 10:48)

**Resource Management Aspects**

- Resource provisioning
- Resource allocation
- Resource requirement mapping
- Resource adaptation
- Resource discovery
- Resource brokering
- Resource estimation
- Resource modeling

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Now, there are different type of approaches researchers followed and or to have resource management aspects or we say the different resource management aspects one is resource provisioning right we will see what are the resource allocation right. So, I need to provision then allocate the resources resource requirement mapping. So, whether I can map the resource requirements right or sometimes whether I can do a priority mapping of the resource requirement, right, like I am forcing that this sort of resource requirement are happening is there, then I provision it accordingly right like resource requirement in different parts of the different time scale of the day are different right different time period of the day are different-different time period of the year are different maybe over the years things are different so on and so forth.

Then we have resource adaptation right how resource can be adapted resource discovery there should be a mechanism that how can I discover resource faithfully like where do I find those resources and type of things and where how do I as a user can look for the resources. So, again I repeat that use and may not be always human user it can be another process another set of processes all together.

So, for optimal and that may be a part of a larger applications. So, in order to do that; I need to discover resources and type of things. So, there should be some provisioning some cataloguing registry type of things whether resources can be there; there are resource brokering, right. So, some sort of a brokerage or agent based things where those are can

be where which acts as agent to have provide me a optimal resource; rather I should say that when a user request for a some resource then that initially it hits a agent or a broker which tries to look out that which are the which are the resource available which are less loaded nodes how VM can be allocated and so and so forth.

So, this is a important aspect that booker brokerage type of things resource estimation that is estimating that what sort of resource requirement will be there these are sometimes important when we do higher level of things like I am looking at a SaaS or PaaS label need to estimate that what sort of back backbone resources are required and there is a thing called resource modeling that how I can model the thing to the resources for considering estimation considering my present load and type of things.

So, these are different aspects and you can see that these are not all are independent aspects they have Intel linking between them also, right. So, these are different aspects and the emphasis may vary based on type of application or type of requirement you are there, right. So, in some cases the some of the aspects may be higher priority and so and so forth nevertheless these are not isolated components again they have a inter linking between the things.

(Refer Slide Time: 14:09)

Resource Management	
Type	Details
Resource provisioning	Allocation of a service provider's resources to a customer
Resource allocation	Distribution of resources economically among competing groups of people or programs
Resource adaptation	Ability or capacity of that system to adjust the resources dynamically to fulfill the requirements of the user
Resource mapping	Correspondence between resources required by the users and resources available with the provider
Resource modeling	Resource modeling is based on detailed information of transmission network elements, resources and entities participating in the network. Attributes of resource management: states, transitions, inputs and outputs within a given environment. Resource modeling helps to predict the resource requirements in subsequent time intervals
Resource estimation	A close guess of the actual resources required for an application, usually with some thought or calculation involved
Resource discovery and selection	Identification of list of authenticated resources that are available for job submission and to choose the best among them
Resource brokering	It is the negotiation of the resources through an agent to ensure that the necessary resources are available at the right time to complete the objectives
Resource scheduling	A resource schedule is a timetable of events and resources. Shared resources are available at certain times and events are planned during these times. In other words, It is determining when an activity should start or end, depending on its (1) duration, (2) predecessor activities, (3) predecessor relationships, and (4) resources allocated

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, this slide; what we are trying to look at that what are the different type of what are the different aspects and what are the different what they mean, right. So, what one

aspect is resource provisioning a location of service provided resources to a constant customers right the customer can be a user or a process, right.

So, resource allocation stands for distribution of resource economically among the computing groups of people or programs or processes resource adaptation ability or capacity of that system to adjust the resource dynamically to fulfill the requirement of the user. So, that based on the user requirement the overall; the system adjusts itself, how this resource; the available resource can be optimally used among the users, again I should say that without compromising the SLA and other quality of services and type of things.

Then we have resource mapping which says that the correspondence between the resource required by the users and the resource available with the providers. So, based on the resource available and they require requirement how we map resource modeling resource modeling is based on detailed information transmission network element resources entities participating in the network, right.

So; that means, attributes of resource managements if we look at they are different states different transitions different outputs with a given environment; so, every resource management. So, if I look at the resource management as a as a entity or a frame work. So, it goes in 2 different state it has different transition from one state to another and every state has the output type of things. So, I can have realized some sort of a state chart diagram or type of things and based on that I need to model that based on this; how this transition will go on.

So, resource estimation; so, how I can closely guess the actual resource required by application usually with some thought or calculation involved I can do some a priori I may have some a priori knowledge about the application or I can have some meta information at the application like this application may require so much memory. So, much displace. So, much threads and type of things and based on that I relocate the resource discovery and selection.

So, as we are discussing identification of list of authenticated resource that are available for the job submission and choose the best among them. So, it is always possible that you have multiple resources or multiple providers with resources are available. So, that discovering that which are the resources and which is the suitable thing and allocating

the most optimal and based about thing and resource brokering. So, negotiation of the resources through an agent ensuring; necessary ensuring that the necessary resources are available at the right time to complete the objectives.

So, I broke up because I have a requirement as a user I have a requirement as a user process, right and then I want to broker I want to negotiate with the agent that which are the things available, right, how things will be available. So, that I my objective is fulfilled and the objective may be resource wise objective the objective sometimes can be on the pricing objective also, this much cost and this much things I have to choose.

So, there is see there is a need for optimization of the whole thing, right. So, I require a brokering service for that and finally, resource scheduling, right. So, is this a scheduling is a timetable of events and resources, right. So, say our resources are available at certain times and events are planned during those times, right. So, it can be; so, I have resources I have my operation procedure. So, time I require some sort of timetable of sibling the resources that exactly a sibling problem per se.

So, I may have lot of this may have lot of components like duration ha some predecessor activities some predecessor relationship resource allocated and so on and so forth. So, there can be different component for to determine that start end and type of things.

(Refer Slide Time: 18:58)

Resource Provisioning Approaches	
Approach	Description
Market equilibrium approach using Game theory	Run time management and allocation of IaaS resources considering several criteria such as the heterogeneous distribution of resources, rational exchange behaviors of cloud users, incomplete common information and dynamic resource allocation.
Network queuing model	Presents a model based on a network of queues, where the queues represent different tiers of the application. The model sufficiently captures the behavior of tiers with significantly different performance characteristics and application lifecycles, such as, session-based workloads, concurrency limits, and caching at intermediate tiers.
Prototype provisioning	Employs the k-means clustering algorithm to automatically determine the workload mix and a queuing model to predict the server capacity for a given workload mix.
Resource (VM) provisioning	Uses virtual machines (VMs) that run on top of the Xen Hypervisor. The system provides a Simple Earliest Deadline First (SEDF) scheduler that implements weighted fair sharing of the CPU capacity among all the VMs. The share of CPU cycles for a particular VM can be changed at runtime.
Adaptive resource provisioning	Automatic bottleneck detection and resolution under dynamic resource management which has the potential to enable cloud infrastructure providers to provide SLAs for web applications that guarantee specific response time requirements while minimizing resource utilizations.
SLA oriented methods	Handling the process of dynamic provisioning to meet user SLAs in autonomic manner. Additional resources are provisioned for applications when required and are removed when they are not necessary.
Dynamic and automated framework	A dynamic and automated framework which can adapt the adaptive parameters to meet the specific accuracy goal; and then dynamically converge to near-optimal resource allocation to handle unexpected changes.
Optimal cloud resource provisioning (OCRP)	The demand and price uncertainty is considered using optimal cloud resource provisioning (OCRP) including deterministic equivalent formulation, sample average approximation, etc.

So, if we look at some of the approaches or some of the different type of aspects like as we discussed previously, resource provisioning resource allocation and try to see that what sort of what sort of approaches people are following or researchers are following into things like first one let us see at the resource provisioning approaches like. So, what we have Nash equilibrium approach for game theory.

So, using some sort of a game theoretic approach to find out that optimal uses of the resource, right; so, what it does the runtime management and allocation of the IaaS resources considering several criteria like heterogeneous distribution of resources, rational exchange behavior of the cloud users in complete common information and dynamics successive allocation and so on and so forth.

So; that means, I based on this different components like heterogeneous duration of resources or users pattern of the cloud users and type of things I want to have a game theoretic approach to look at this. So, it is; we can look at as a game where one side is that the consumer which are a; who are hungry for the resources or looking for the resources other side there is provider who are provisioning the resources and I want to find out a optimal way of allocating the resources. So, that is Nash equilibrium based using game approach.

So, there are there are research and or there are methods and approaches people are following there we have network queuing model. So, which a model based on network queue like those who have gone through network queuing model or network queue models in data networks, etcetera you can understand, it is more of a again resource provisioning mechanism queues where queues represent different tires of application.

So, the model significantly or sufficiently captures the behavior of the tires with significantly different performance characteristics and application like session based workload concurrency limits and caching the intermediate tires and like that. So, it is try to you; what we do what we are what they are doing here to try to exploit network queuing model there are approaches for prototype provisioning employs the k means clustering algorithm to automatically determine the workload mix and queuing model to predict the server capacity for a given workload mix.

So, what it is trying to do. So, it tries to cluster using say came in cluster k-means cluster to automatically determine these; what is this workload means of the different user and then to predict that what are the what how can be provisioned.

There are other resource provisioning like VM provisioning, things like user virtual machines that runs top of the Xen hypervisor. So, the system provides say some sort of a scheduler like in some work they propose a simply simple earliest deadline first scheduler that implements the weighted fair sharing of the CPU capacity among the VMs, right.

So, what it is doing it is taking the VM which run over the hypervisor and scheduling it based on the type of type of the load it is it is getting that share CPU cycles a particular VM can be changed on the runtime and so on and so forth. So, if I have requirement from more resource requirement, then I migrate from one VM to the other VMs and type of things can be done.

There are other methods and approaches like adaptive resource provisioning which tries to automatically detect the bottlenecks and residues and resolve that using dynamic resource management there are things called sea SLA oriented resource methods, handling process and dynamic provisioning to meet user SLAs in a automatic manner dynamic an automatic framework which adapt the adaptive parameters which adapt the parameters to meet the specific users or accuracy goals. So, it goes on provisioning the recruit resources based on the quality of services or the type of SLA, it has to support and also there is optimal cloud provisioning mechanisms what which tries to look at the demand and price uncertainty considering those try to optimize.

So, we see there are several approaches which can be used for this sort of resource provisioning mechanisms.

(Refer Slide Time: 24:11)

Resource Allocation Approaches	
Approach	Description
Market-oriented resource allocation	Considers the case of a single cloud provider and address the question how to best match customer demand in terms of both supply and price in order to maximize the providers revenue and customer satisfaction while minimizing energy cost. In particular, it models the problem as a constrained discrete-time optimal control problem and uses Model Predictive Control(MPC) to find its solution.
Intelligent multi-agent model	An intelligent multi-agent model based on virtualization rules for resource virtualization to automatically allocate service resources suitable for mobile devices. It infers user demands by analyzing and learning user context information.
Energy-Aware Resource allocation	Resource allocation is carried out by mimicking the behavior of ants, that the ants are likely to choose the path identified as a shortest path, which is indicated by a relatively higher density of pheromone left on the path compared to other possible paths.
Measurement-based analysis on performance	Focuses on measurement based analysis on performance impact of co-locating applications in a virtualized cloud in terms of throughput and resource sharing effectiveness, including the impact of idle instances on applications that are running concurrently on the same physical host.
Dynamic resource allocation method	Dynamic resource allocation method based on the load of VMs on IaaS, which enables users to dynamically add and/or delete one or more instances on the basis of the load and the conditions specified by the user.
Real-time resource allocation mechanism	Designed for helping small and medium sized IaaS cloud providers to better utilize their hardware resources with minimum operational cost by a well-designed underlying hardware infrastructure, an efficient resource scheduling algorithm and a set of migrating operations of VMs.
Dynamic scheduling and consolidation mechanism	Presents the architecture and algorithmic blueprints of a framework for workload co-location, which provides customers with the ability to formally express workload scheduling flexibilities using Directed Acyclic Graphs (DAGs), and optimizes the use of cloud resources to collocate client's workloads.



Similarly, if we look at the resource allocation there are again several approaches few are listed here like market oriented resource allocation. So, which are driven by the market requirement market demand on the things. So, we try to do model predictive model predictive control to find its solution of the of that particular resource allocation there are intelligent multi agent model primarily looking for resource view virtualization to automatically allocate services resource available specifically for devices which are mobile, right.

So, it is I can have a intelligent multi agent model to allocate optimal resources energy aware resource allocation. So, this allocation is energy aware. So, that I can do a optimal energy provisioning measurement based analysis on performance. So, it; allocation again based on different metrics or measurement parameters dynamic resource allocation methods real time resource allocation mechanisms like if there is a real time demand on the things how resource can be allocated. So, designed for helping small medium size IaaS cloud providers to better utilize their hardware resource with minimal operational cost by a well designed underlining hardware infrastructure, right.

So, in order to help for this especially small and medium sized IaaS cloud provider; so how it can be allocated in a real time and dynamic scheduling and consolidation mechanisms over and above I can have a dynamic scheduling and consolidation mechanisms of available resources.

(Refer Slide Time: 26:20)

Resource Mapping Approaches	
Approach	Description
Symmetric mapping pattern	Symmetric mapping pattern for the design of resource supply systems. It divides resource supply in three functions: (1) users and providers match and engage in resource supply agreements, (2) users place tasks on subscribed resource containers, and (3) providers place supplied resource containers on physical resources.
Load-aware mapping	Explores how to simplify VM image management and reduce image preparation overhead by the multicast file transferring and image caching/reusing. Load-Aware Mapping to further reduce deploying overhead and make efficient use of resources.
Minimum congestion mapping	Framework for solving a natural graph mapping problem arising in cloud computing. Applying this framework to obtain offline and online approximation algorithms for workloads given by depth-d trees and complete graphs.
Iterated local search based request partitioning	Request partitioning approach based on iterated local search is introduced that facilitates the cost-efficient and on-line splitting of user requests among eligible Cloud Service Providers (CSPs) within a networked cloud environment.
TQoP API	Designed to accept different resource usage prediction models and map QoS constraints to resources from various IaaS providers.
Impatient task mapping	Batch mapping via genetic algorithms with throughput as a fitness function that can be used to map jobs to cloud resources.
Distributed ensembles of virtual appliances (DEVA)	Requirements are inferred by observing the behavior of the system under different conditions and creating a model that can be later used to obtain approximate parameters to provide the resources.
Mapping a virtual network onto a substrate network	An effective method (using backbone mapping) for computing high quality mappings of virtual networks onto substrate networks. The computed virtual networks are constructed to have sufficient capacity to accommodate any traffic pattern allowed by user-specified traffic constraints.

There are again several approaches for resource mapping like symmetric mapping pattern that is for designing resource supply systems it divides the resource into three major functions users and providers match and engage resource supply management agreements.

So, that is users and the providers match that where the requirement are matching and do that type of matching before that users place tasks on the subscribe resource containers right. So, that it subscribe resource container place the tasks and the mapping is done or provider place supplied resource container on physical resources and type of things.

So, these are driven by container based services which is another type of; another technology which is coming up in a bigger ways that this container classes and container things. So, user can subscribe resource continent place their tasks or providers can place supplied resource container on the physical resources.

It can be a mapping of the load aware mapping. So, explore how simply VM image management and reduce image preparation over it by multicast file transferring and image caching and using. So, it is based on the load, it does a load aware mapping to reduce deploying over a rate and make efficient use of the resources. So, based on the available load it does the load availability there can be technique for iterated local search based request partitioning.

So, whether I can partition request partitioning approach follow a based on iterated local search to facilitate a cost efficient and online splitting of is your request among eligible cloud service provider. So, user may be requesting on a user requests, I can basically partition into smaller part if there is a there is a way of partitioning it, I can do a intelligent partitioning algorithm and then allocate the things into different CSPs like different cloud service providers and it that is; that means, that a large requests can be partitioned into smaller and look at it.

So, there are other approaches like distributed in ensemble of virtual applications like I have virtual applications ensemble name or mapping a virtual network of a substrate network. So, I have a underlining network and then I map a substrate network like why map a virtual network which is main for the user to this substrate network, right.

So, again this is a resource mapping from the network side of view there is a requirement from the user from the network and map it on the things.

(Refer Slide Time: 29:27)

Resource Adaptation Approaches	
Approach	Description
Reinforcement learning guided control policy	A multi-input multi-output feedback control model-based dynamic resource provisioning algorithm which adopts reinforcement learning to adjust adaptive parameters to guarantee the optimal application benefit within the time constraint
Web-service based prototype	A web-service based prototype framework, and used it for performance evaluation of various resource adaptation algorithms under different realistic settings
OnTimeMeasure service	Presents an application – adaptation case study that uses OnTimeMeasure-enabled performance intelligence in the context of dynamic resource allocation within thin-client based virtual desktop clouds to increase cloud scalability, while simultaneously delivering satisfactory user quality-of-experience
Virtual networks	Proposes virtual networks architecture as a mechanism in cloud computing that can aggregate traffic isolation, improving security and facilitating pricing, also allowing customers to act in cases where the performance is not in accordance with the contract for services
DNS-based Load Balancing	Proposes a system that contain the appropriate elements so that applications can be scaled by replicating VMs (or application containers), by reconfiguring them on the fly, and by adding load balancers in front of these replicas that can scale by themselves
Hybrid Approach	Proposes a mechanism for providing dynamic management in virtualized consolidated server environments that host multiple multi-tier applications using layered queuing models for Xen-based virtual machine environments, which is a novel optimization technique that uses a combination of bin packing and gradient search



And there are several adaptation approaches like reinforce learning guided control policy. So, that is a learning mechanism to look at the adaptation there are web service based prototypes, right. So, which can be used for resource adaptation there are several others like looking at virtual networks DNAs based load balancing and of course, we can have hybrid approaches for having this sort of load resource adaptation.

(Refer Slide Time: 30:01)

Performance Metrics for Resource Management

- Reliability
- Ease of deployment
- QoS
- Delay
- Control overhead

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, if I have this several type of techniques like as we have seen here is like resource provisioning, allocation, we discussed few of them; resource requirement, mapping adaptation and so on and so forth; how to judge that they are performance finally, what we are looking for is some matrix, right.

So, all those approaches need to be judged based on some metric like reliability ease of deployment like I have a mechanisms and it takes lot of lot of overhead to deployment quality of services should not be compromised; there should not be delay or much delay or the delays should be within the limit and control over it in order to manage these resource management things in order to control these resource management mechanisms or processes what are my control over it, right.

So, whenever we are looking for any resource management a play resource management tools or techniques we need to look at all those different aspects. So, otherwise the overall resource management may kill the basic purpose of this cloud computing paradigm, right. So, that is of uses scalability and in finite resources and type of things we may suffer.

So, we need to look at this different matrix and if you can you can see that these are the matrix may differ from different type of requirements, right. So, different user may have different or different user processes the different requirements. And where somewhere the reliability may be pretty high somewhere the quality of service somewhere some of

the applications may be delay concerned, some of the application may be accuracy concerned. And we need to take up the actual take up the resource management process resource management tools and technique based on those parameters.

Thank you.

**Cloud Computing**  
**Prof. Soumya Kanti Ghosh**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**

**Lecture - 26**  
**Cloud Security – I**

Hi. We will be continuing our discussion on Cloud Computing. Today we will be discussing on another major aspect of this cloud computing which is which we can say Cloud Security. So, we will talk we will try to have a brief overview of the security parts say, and how this security affects cloud computing. As we all understand that when we go for cloud computing whether it is the infrastructure as a service a platform as a service or software as a service or anything as a service, what we are relying on a third party service provider.

So, our application data processes are running on some third party. So, whenever it is running on third party the security becomes a issue specially that what is the availability, where my data is stored whether it is been seen or intercepted by my some other parties, and those concerns will be there and specially if this is a mission critical operations or mission critical data or some critical data like banking data, defense data even academic data related to students results and other things. This needs to be looked into in a in a various serious person.

We will see in the course of the things that one of the one of the major hindrance towards going towards this cloud is more than technology, rather more this concern about security what will be the policies what data policy etcetera and so on and so forth. So, with this we will start our thing, but before going to that I thought that it will be quick brush up of what do you mean by security, in terms of when we talk about computer security or information security or network security. So, what are the different aspects are, there it is likely that all those aspect in some form of other will be also reflected in the cloud, but the concern may be different. So, before going to the cloud security part say we will see security in general for any computing service computing and networking service ok.

(Refer Slide Time: 02:36)

The slide has a dark blue header and footer. The main content area is yellow. The title 'Security - Basic Components' is in red. The components listed are:

- Confidentiality
  - Keeping data and resources hidden
- Integrity
  - Data integrity (integrity)
  - Origin integrity (authentication)
- Availability
  - Enabling access to data and resources

At the bottom, there are logos for IIT Kharagpur and NPTEL, along with the text 'NPTEL ONLINE CERTIFICATION COURSES'.

So, if we look at security what are the 3 basic components one is the confidentiality integrity and availability right. This is what we say CIA components right. Confidentiality deals with keep keeping the data and resources hidden that you do not know that where the data is that it is confidential, integrity is that data integrity is maintained like or origin or the source integrity is mentioned, may maintained right. Like So that whatever I sent from a to b; b receives the same thing or that integrity or authentication of the source that, I am getting from the itself it is there and availability in happening access to the data and resources there is another important component right.

So, that is what we see that most of the attacks are going as denial of services where the availability is compromised. So, everything is fine, but finally, you do not have the resource at your hand. So, it is some sort of a dos or sometimes the ddos type of attacks.

(Refer Slide Time: 03:43)

**Security Attacks**

- Any action that compromises the security of information.
- Four types of attack:
  1. Interruption
  2. Interception
  3. Modification
  4. Fabrication
- Basic model:

S —————→ D  
Source                      Destination

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, any security attack on the other say that any action that compromises the security of the information, or any action which violates the CIA type of things there is basic premise right. There are lot of other components we will see.

So, if we look at there are immediately it will come up that there are typically 4 type of things maybe there, one is inter interruption, one is interception, modification, fabrication. So, this 4 components more or less encompasses or combination of this more or less or it encompasses all type of things, which are which are compromised during a attack. So, our basic model is a source sending a data to a destination, and when we talk a talk about interruptions.

(Refer Slide Time: 04:31)

The diagram illustrates two types of security attacks:

- Interruption:** A horizontal line connects source 'S' and destination 'D'. A vertical line segment extends downwards from the middle of this line, representing an interruption.
- Interception:** A horizontal line connects source 'S' and destination 'D'. A vertical line segment extends downwards from below the horizontal line, representing an interceptor 'I' listening to the message.

Below the diagrams, there is a logo for IIT Kharagpur and NPTEL Online Certification Courses.

So, that the message or the communication path is interrupted. It can be interception that goes from source to destination, but somebody else also intercept and listening to the thing.

(Refer Slide Time: 04:47)

The diagram illustrates two types of security attacks:

- Modification:** A horizontal line connects source 'S' and destination 'D'. A vertical line segment extends downwards from the middle of the line, then turns right, representing an intruder 'I' modifying the message.
- Fabrication:** A horizontal line connects source 'S' and destination 'D'. A vertical line segment extends downwards from below the horizontal line, then turns right, representing an intruder 'I' fabricating data.

Below the diagrams, there is a logo for IIT Kharagpur and NPTEL Online Certification Courses.

So, this is attack on availability this availability is blocked this is attack on confidentiality like you are sending from a to b or s to d and somebody else, that intruder I is listening to that it can be attack on modification, that this attack on integrity of that data right; so or even the origin right. Source is sending to d, but in between there is a intruder I which

intercept the message changes the message and send it to d. So, d for d it is a message coming from s and the message am has been changed to am dash, but still the for d it is it is the message which is send which has been send or which has been forwarded by sources.

So, that is a attack on integrity. So, there can be attack on authenticity right. I pretend or intrude are pretend to be the sources right. And; so it is attack on authenticity. So, I need to authenticate who is my source. So, before receiving a message I need to know that I am supposed to receive from a authenticated source is, and I am receiving those message. So, that is a attack on authenticity or what we say fabrication.

(Refer Slide Time: 06:04)

**Classes of Threats**

- Disclosure
  - Snooping
- Deception
  - Modification, spoofing, repudiation of origin, denial of receipt
- Disruption
  - Modification
- Usurpation
  - Modification, spoofing, delay, denial of service

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES



Now so, one side we see that the major security components other side that the type of attacks which can be there what a and if you look at these are this can this is true for whether it is a computer security. Or a information security or network security or cloud security right. They it may have different type of characteristics and manifestation nevertheless it has the same type of what we say same type of problems, or same type of security issues realizes.

Now, if you look at that what are the threats right. So, threats does not mean it is attacked right. So, it is like vulnerability does not mean that it is compromised, but these are the possible threats. So, classes of threats one is a threat of disclosure right. So, I have a threat of disclosure like what which is type of in the attack what we say snooping. So,

threats of deception like modifications spoofing repudiation of origin denial of receipt and type of things. So, this is a threat of deception.

So, there can be a threat of disruption that is if it is modified in the threats of disruption service, and another is a threat of usurpation, that is modification spoofing delay denial of services.

(Refer Slide Time: 07:38)

**Policies and Mechanisms**

- Policy says what is, and is not, allowed
  - This defines "security" for the site/system/etc.
- Mechanisms enforce policies
- Composition of policies
  - If policies conflict, discrepancies may create security vulnerabilities

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, these are these are different category of threats which are there. So, we have attacks which have security concerns and threats these are different components. Overall whenever a whenever a it systems or any information system whether it is organizational or it is personal or it is inter organization intra organization, whatever there are guided by policies and mechanisms very tricky issues. So, policy says what is what is not allowed right.

So, the policy says that what is allowed what is that not allowed right. So, it is it tries to do it in a fashion that which of the things can be allowed and things there can be hierarchical to a way of defining policies. There can be different way of things we are not going to that. So, this defines the security of the site systems overall information structure, a overall network access protocol and individual to group to distributed anything right. So, there is policies usually in organizations policies are made somewhat centralized. In the sense it has been formulated across the for all components of the organizations and it is something a sort of a policy making body does it.

Now, incidentally the implementation is most of the time distributed. Like I say that I have a I have this IIT Kharagpur network. So, there are several departments there as a several sub networks there are several layer 3 plus layer 3 and layer 3 plus type of switches. So, it says they policy that this way the traffic will go and etcetera etcetera, and not only that they are additionally there are also presidents etcetera. At the same time this policy need to be implemented across this different category of devices. So, the implementation is often in a distributed fashion or in different devices and type of things where the. So now, there is a big challenge the how to guarantee this implementation conform to the policy one to one right. It is nothing more or less what the policy defines.

So, there is a these are some open not exactly open problem these are very strong research problem across the world that how to how to formally say that your implementation and policy match with each other and so forth. Now so, policies says what is and what is not allowed, where I on the other hand mechanisms enforce policies right. So, I have mechanism to enforce policies. So, composition of policies if the policy is conflict discrepancies may create security vulnerabilities.

So, there is another things, if it is if when we compose policies. So, if I have several policy and composition of the policies if there are conflicts, discrepancies may arise and then the there can be security vulnerabilities. Like I can say that one policy says that this traffic can be allowed or another policy said that this category of traffic should be denied, and you see there is a overlap that which can be either allowed or denied you need to decide right. This mainly happens because number of cases this implement is in a distributed way. And if sometimes there are there are class in the local versus global policies etcetera.

So, these need to be addressed. These need to be first of all I defined and this need to be a address and this becomes a very critical thing when there is a organization is pretty large to look at An individual policies and verify and all those things.

(Refer Slide Time: 11:08)

**Goals of Security**

- **Prevention**
  - Prevent attackers from violating security policy
- **Detection**
  - Detect attackers' violation of security policy
- **Recovery**
  - Stop attack, assess and repair damage
  - Continue to function correctly even if attack succeeds

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, looking all those things, so what we have seems there are security models or security objectives. There are attack models there are threats, and I there are policies and mechanisms. So, these are a different component which looks at a different way of the things right.

Now, I need to bring them together, and have what is my security goal. So, one of the major security goal is prevention, prevent attackers from violating security policies. So, that should be there. So, attacker if I have the security policies if it is restricts that thing, attackers should be should not be able to violate the security policies. Detection detect attackers violation of the security policy. So, detection that the when the security policies are being violated by the attackers need to be detected right.

So, detection I can have we then we will try to the earlier the detection is, more strong your security perimeter right. So, we need to detect as early as possible. Because if the attack has gone on go on into the place then what we what we are left with is more of the post mortem of the what had happened. And basically learn to look at the other things. There are a issues of another issue is recovering. If attack and if compromised if down to some extent or fully or partially, then how to recover from this thing, right?

So, what will be my recovery mechanisms mechanism from these types of things, like stop attack, assess and repair damage, continue to function correctly even if the attack succeeds. And there are different type of things people that there are in best practices.

We have in critical system as redundancy system, there are logging mechanisms to recover and other things never the less, we need to recover from the thing to a stage where we were like pre attack stays type of things you know on all doing. So, we incurred cost right. All this comes with a cost.

(Refer Slide Time: 13:21)

**Trust and Assumptions**

- Underlie all aspects of security
- Policies
  - Unambiguously partition system states
  - Correctly capture security requirements
- Mechanisms
  - Assumed to enforce policy
  - Support mechanisms work correctly

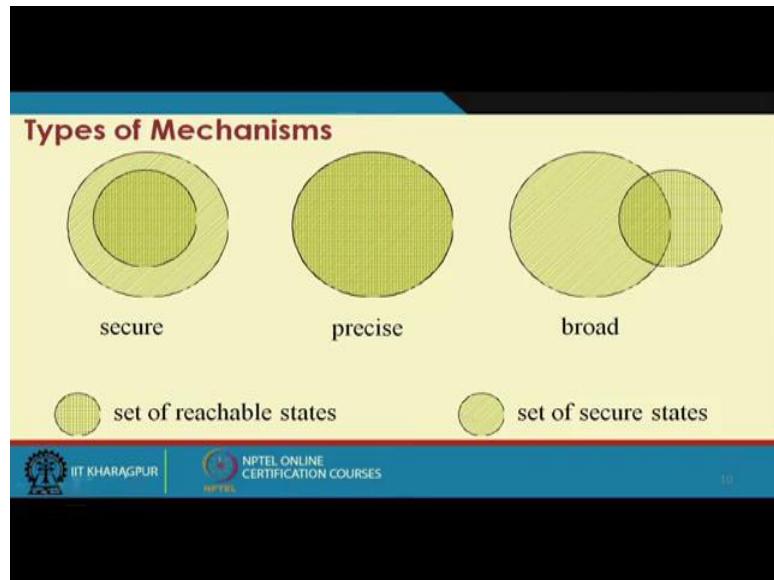
IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES  
NPTEL

Trust and assumptions that is another aspect. So, underlie all aspects of security right. So, I have some trust and assumption, I trust this system I assume that system will work fine or this or this particular application and so on and so forth things are there. So, it all stress and if you if you look at our day to day life also for security mechanisms we have some sort of a test trust, and assumption like I say I understand what we trust that the security person who is guarding that particular installation or particular premise is can be trusted right.

So, I assume that this can be trusted to this extent and so on and so forth and type of things. So, this also is important thing. So, policies unambiguously partition system states right. So, that is if you look at the system state system goes on different state because first of all it is a dynamic thing right. It is not that it is statically one defined things are there. So, it is it should unambiguously partition the system state; that means, I am in this state or this state which state I am there. So, it should be ambiguous partition correctly capture the security requirement of every stage right.

So, it will not only partition it should also security requirements of every stage. Mechanisms assume to enforce policies. So, mechanisms are there to enforce policies. Supports mechanisms support mechanisms work correctly. So, that the mechanism is basically a implementation or realization of the policies and that should be they are in place.

(Refer Slide Time: 15:03)



Now, if we look at a little bit holistically. So, if I have like set of reachable states of a system right.

So, if the set of reachable state is in this type of mess, and if I have a set of secured stages like this type of hash line right. So, one we say that if the set of reachable state is within the overall set of secured state then I say it is fully secure right. So, what I am trying to say that the system goes over different state all are within the security state. So, I have say I have security state security state at s1 to s20 as secure my system basically hover between s5 to s16.

So that means, it is always in the security state. It can be precise that the it totally matches with this the set of security and set of reachability is matches. Other can be broad; that means, all are not in the security zone or the in the secure state, but a, but there are some state which is there right. The one thing we should be we should know that how my security policy mechanisms visa-vies work. So, that I can say that how much secured I am.

(Refer Slide Time: 16:38)

**Assurance**

- Specification
  - Requirements analysis
  - Statement of desired functionality
- Design
  - How system will meet specification
- Implementation
  - Programs/systems that carry out design

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, there is a issue of assurance, like which consists of specifications right. Like requirement analysis statement of desired functionalities designs; how the system will meet the specification and implementation, program systems that carry out design right. So, what it tries to do that in order to have properly design the thing, I can assure that this much security has been can be assured based on my design specification etcetera right.

(Refer Slide Time: 17:25)

**Operational Issues**

- Cost-Benefit Analysis
  - Is it cheaper to prevent or recover?
- Risk Analysis
  - Should we protect something?
  - How much should we protect this thing?
- Laws and Customs
  - Are desired security measures illegal?
  - Will people do them?

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, this is these are the best practices which need to be put into place. So, that my security level goes up. Now there are issues of operational issues, or sometimes there are

economical issues right. Cost benefit analysis is it typical it is cheaper to prevent or recover right. So, so whether which is costly like, it is it is recovering is costly or things like I say that I have a lap which has some Linux installation, or windows installation or combined both of them. And we run the lap on day to day basis, but as such I we do not store anything in the system, right.

So, a in that case that is students are supposed to bring their documents or download their codes etcetera run, and then release the thing right, but end of the day there is no question of storing any data. Or there is no responsibility from the authority, so that the data will be saved etcetera. For that I may not look for much interested for preventing the attack right. Even some attack is there if I can recognize I can always reinstall that I can have a already a image of the whole system and I can reinstall the image, right.

But on the other hand if there is a data intensive or say research data etcetera, then I am more interested in preventing the attack right. Or a system which is online running on things I want to prevent the data hum. So, that is cost analysis benefit other is the risk analysis right; should we protect something how much should we protect this thing, right. How much risk is there are laws and customs right. Let are desired security measures illegal will people do them etcetera. So, we have different operational issues which are there.

(Refer Slide Time: 19:21)

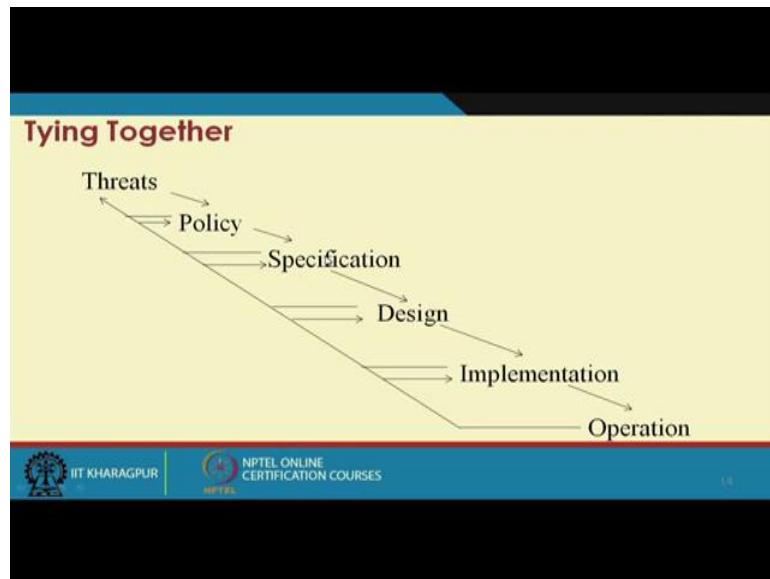
**Human Issues**

- Organizational Problems
  - Power and responsibility
  - Financial benefits
- People problems
  - Outsiders and insiders
  - Social engineering

IT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

There are of course, some of the human issues rights; organizational problems or people problems. So, there are always human in the loop, and there are human issues of responsibility per says authority And so on and so forth.

(Refer Slide Time: 19:39)



So, if we tie them together. So, threats are there policies are a based on the threats policies are made based on the specification based on the things policy specification the design is there based on design that implementation and then operation. These operational issues are feedback either to the implementation or design specification and things or operational issues comes as a threat.

(Refer Slide Time: 20:13)

**Passive and Active Attacks**

- Passive attacks
  - Obtain information that is being transmitted (eavesdropping).
  - Two types:
    - Release of message contents:- It may be desirable to prevent the opponent from learning the contents of the transmission.
    - Traffic analysis:- The opponent can determine the location and identity of communicating hosts, and observe the frequency and length of messages being exchanged.
  - Very difficult to detect.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, this is if we try to make them together bring them together it is like sort of it. Now what we are looking as of now is more from the point of view of like from the providers or the from the system point of view. Like what are the what could be the possible threats what could be important possible policies, how to implement what is the mechanisms and so on and so forth, right. But if we try to look at that what are the different type of attacks like one is passive attack right. Obtain information that is being transmitted eavesdropping. So, it is not the attacks, but these are more eavesdropping.

So, 2 types release of message content it may be desirable to prevent the opponent from learning the contents of the transmission. So, release of the message content may be one of the attack traffic analysis. Like I do not look at the message, but say, but I want to look at the traffic right. If the traffic is highly volatile or heavy or low I try to predict that what could be the effect of the effect of the type of mechanisms going on right. I can say that if there is a very high traffic that may be some sort of a video conferencing or video chat is going on.

And it may be something need to be looked attached to me something if is a low traffic or medium traffic I can say that this is the type of things. And based on not only the traffic persist time it said also plays different important role.

(Refer Slide Time: 21:55)

□ Active attacks

- Involve some modification of the data stream or the creation of a false stream.
- Four categories:
  - Masquerade:- One entity pretends to be a different entity.
  - Replay:- Passive capture of a data unit and its subsequent retransmission to produce an unauthorized effect.
  - Modification:- Some portion of a legitimate message is altered.
  - Denial of service:- Prevents the normal use of communication facilities.

Usually passive attackers are difficult to detect because they do not do direct harm and very difficult to detect, whereas on the other hand we have active attackers. Like involve some modification of the data stream or creation of false streams. So, that is these are all active attackers right, So 4 categories. So, one is masquerade one entity pretends to be different entity. So, that is the one attacker replay. Passive capture of the data unit and it is subsequent retransmission to produce unauthorized effect right. So, this is a replay attack right.

So, contained passive capture of the data units and it is subsequent retransmission huge amount of retransmission of the date modification. Some portion of the legitimate message is altered. So, that is the attack of modification denial of service prevents the normal use of communication facilities. So, this is a dos type of attack or denial of services attack. So, all these attacks actually create problem in our in the operation of the active system.

(Refer Slide Time: 23:04)

**Security Services**

- Confidentiality (privacy)
- Authentication (who created or sent the data)
- Integrity (has not been altered)
- Non-repudiation (the order is final)
- Access control (prevent misuse of resources)
- Availability (permanence, non-erasure)
  - Denial of Service Attacks
  - Virus that deletes files

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Now, in the security services as those things we have seen that the these are the security threats security services try to give provide this sort of services, right. Like confidentiality, authenticity, integrity, non repudiation, access control, accessibility, a availability. So, this first thing we had discussed non repudiation what we say the order is final right. It is like that you say you order something like you say that the bank you instruct your bank that you transfer over online that you transfer x amount from my account to somebody else's account and next day I go to the bank that I never gave this right.

So, there is a there should be a way of handling this. So, there is non repudiation is why is such things is basically says the order is better. Access control is a big field that how this access control will be there are works on role based access control mechanisms and so on and so forth. It is basically say the prevent misuse of resources. So, you should have that particular resource particular access to a particular resource center So That you can use that resources

Availability performance and non erasure type of services. So, that is denial of services service attack and there can be virus that deletes files.

(Refer Slide Time: 24:39)

**Role of Security**

- A security infrastructure provides:
  - **Confidentiality** – protection against loss of privacy
  - **Integrity** – protection against data alteration/ corruption
  - **Availability** – protection against denial of service
  - **Authentication** – identification of legitimate users
  - **Authorization** – determination of whether or not an operation is allowed by a certain user
  - **Non-repudiation** – ability to trace what happened, & prevent denial of actions
  - **Safety** – protection against tampering, damage & theft

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, that is your that is the also case of non availability, so, role of security. So, if when you when you talk about computer security network security information security cloud security and so on and so forth. What are the role of thing? The security infrastructure should provide first of all confidentiality. That means protect against loss of privacy the integrity protection against data altered data alteration or corruption.

So, that is the protection of this the integrity, because as you have seen integrity that during the message transfer the data is being altered. Availability protection against denial of service authentication identification of legitimate users so that how to identify an authenticated legitimate user. Authorization is determination of whether or not operation is allowed by a certain user. Non repudiation as we have discussed the order is final, safety protection against tampering damage etcetera right.

(Refer Slide Time: 25:44)

**Types of Attack**

- Social engineering/phishing
- Physical break-ins, theft, and curb shopping
- Password attacks
- Buffer overflows
- Command injection
- Denial of service
- Exploitation of faulty application logic
- Snooping
- Packet manipulation or fabrication
- Backdoors

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

And we have a series of attacks based on that different type of vulnerabilities and so on and so forth, from social engineering to phishing, password attacks, buffer overflow, command injection and etcetera etcetera. So, these are different type of attacks which are there in the information system, which are true in some sense for the cloud infrastructure also.

(Refer Slide Time: 26:05)

**Network Security...**

- Network security works like this:
  - Determine network security policy
  - Implement network security policy
  - Reconnaissance
  - Vulnerability scanning
  - Penetration testing
  - Post-attack investigation

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Now, if we look at a typical scenario like say network security which is very prominent because the cloud is based on a this term is basically build on the distributed systems

which are leverage or network and so, it is important that the basic network level security is high. So, network security works like this determination of the network security policies that what should be the security policy, implementing those policy then reconnaissance. So, that should this to see that whether the security things are in place or not vulnerability scanning like how vulnerable I am.

So, that look at the vulnerability scanning, there is a concept of penetration testing; that means, or what we say self attacking sort of scenario is a self and safe attacking scenario, that how much I can predict it to the system. So, it is a penetration testing and there is a need of post attack investigation, if there is a attack then post attack investigation.

(Refer Slide Time: 27:05)

**Step 1: Determine Security Policy**

- A security policy is a full security roadmap
  - Usage policy for networks, servers, etc.
  - User training about password sharing, password strength, social engineering, privacy, etc.
  - Privacy policy for all maintained data
  - A schedule for updates, audits, etc.
- The network design should reflect this policy
  - The placement/protection of database/file servers
  - The location of demilitarized zones (DMZs)
  - The placement and rules of firewalls
  - The deployment of intrusion detection systems (IDSs)

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, determination of security policy, that the security policy is a full security roadmap and for any organization. So, smaller things if it is a inter organization. So, that what will be the security policies need to be placed right? It is a full road map have to be there. The network design should reflect these policies. So, if it is a network thing.

(Refer Slide Time: 27:35)

**Step 2: Implement Security Policy**

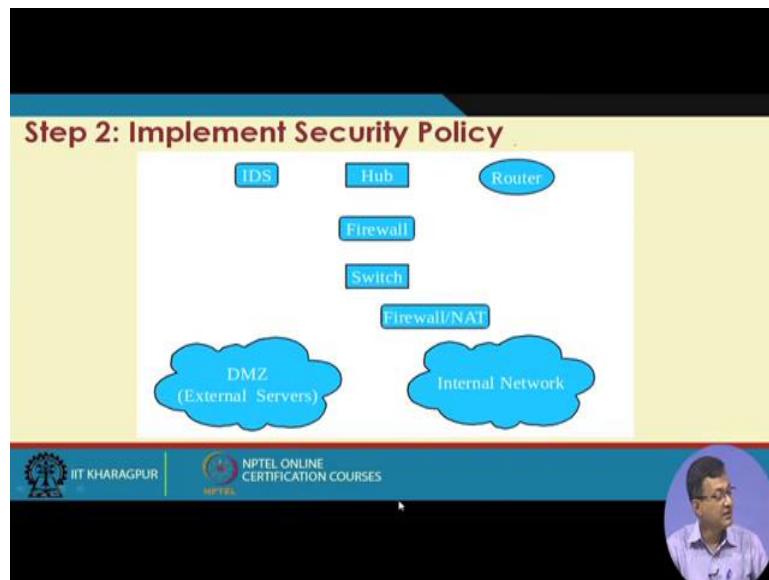
- Implementing a security policy includes:
  - Installing and configuring firewalls
    - *iptables* is a common free firewall configuration for Linux
    - Rules for incoming packets should be created
      - These rules should drop packets by default
    - Rules for outgoing packets *may* be created
      - This depends on your security policy
  - Installing and configuring IDSSes
    - *snort* is a free and upgradeable IDS for several platforms
    - Most IDSSs send alerts to log files regularly
    - Serious events can trigger paging, E-Mail, telephone

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES



So, whenever you are designing. It should confirm these security policies. So, implementing the security policies implementing policies include installation and configuration of security measures like firewalls, installation of configuration of ID's and there are several other type of things which need to be there.

(Refer Slide Time: 27:53)



So, if we look at it is a big picture like this, where you have different there is demilitarized zone internal network. And that firewall or network address translator nat,

switch, firewall and type of things right. So, it is it is dual homing or 2 firewalls are there.

(Refer Slide Time: 28:11)

**Step 2: Implement Security Policy**

- Firewall
  - Applies filtering rules to packets passing through it
  - Comes in three major types:
    - Packet filter – Filters by destination IP, port or protocol
    - Stateful – Records information about ongoing TCP sessions, and ensures out-of-session packets are discarded
    - Application proxy – Acts as a proxy for a specific application, and scans all layers for malicious data
- Intrusion Detection System (IDS)
  - Scans the incoming messages, and creates alerts when suspected scans/attacks are in progress
- Honeypot/honeynet (e.g. honeyd)
  - Simulates a decoy host (or network) with services

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, implement security policies either the policies or the excess rules in the firewall or ID's or there is a concept of honeypot or honeynet where vulnerable things are there so, that lot of attacks will be there and the security means security personals understand that what sort of attacks is there. Based on that signatures they basically do they basically fine tune there are RDA ID's or firewall policies.

(Refer Slide Time: 28:45)

**Step 3: Reconnaissance**

- First, we learn about the network
  - IP addresses of hosts on the network
  - Identify key servers with critical data
  - Services running on those hosts/servers
  - Vulnerabilities on those services
- Two forms: passive and active
  - Passive reconnaissance is undetectable
  - Active reconnaissance is often detectable by IDS

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, the next thing is that need to learn about the network right. So, in order to whether to attack or prevent you need to none of the network. So, IP address of the host identify key servers with critical data and so on and so forth. So, there are 2 forms are there one is passive which is undetectable, one is active it is not often detected by the ID's right.

(Refer Slide Time: 29:12)

**Step 4: Vulnerability Scanning**

- We now have a list of hosts and services
  - We can now target these services for attacks
- Many scanners will detect vulnerabilities (e.g. nessus)
  - These scanners produce a risk report
- Other scanners will allow you to exploit them (e.g. metasploit)
  - These scanners find ways in, and allow you to choose the payload to use (e.g. obtain a root shell, download a package)
  - The payload is the code that runs once inside
- The best scanners are updateable
  - For new vulnerabilities, install/write new plug-ins
  - e.g. Nessus Attack Scripting Language (NASL)

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, there are this is a need. There is a vulnerability scanning that as we are discussing as we are discussing couple of minutes back that I need to basically scan my vulnerabilities right, so that how vulnerable I am in the other system wise.

So, there are different scanner there are in case of a network there are different like there is a open source thicknesses in map and so and so forth. So, that you can basically scan that which are the ports open, what are the possible vulnerabilities and type of things right. So, this is important that you scan and see that; what is the security quote unquote security health of your installation.

So, other scanner will allow to exploit then they are they are called metasploit and type of things which has is security database, there are difference security vulnerability database like one such is that in NVD national vulnerability database where which basically says that what are the different vulnerabilities. So, scanners are need to be updatable. So, that it goes on as in case of antivirus etcetera, which are primarily scanner. So, need to update with the signatures.

(Refer Slide Time: 30:21)

**Step 5: Penetration Testing**

- We have identified vulnerabilities
  - Now, we can exploit them to gain access
  - Using frameworks (e.g. metasploit), this is as simple as selecting a payload to execute
  - Otherwise, we manufacture an exploit
- We may also have to try to find new vulnerabilities
  - This involves writing code or testing functions accepting user input

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES



And then we have the penetration testing, like one we do a vulnerability analysis of a say a network and then looking at the vulnerabilities, we do a penetration testing of the system that how much I can penetrate into the systems and type of things. These are safe attacks and late as late as means a organization or the security personnel can know that what are the different vulnerability points and put appropriate patches.

(Refer Slide Time: 30:53)

**Step 6: Post-Attack Investigation**

- Forensics of Attacks
- This process is heavily guided by laws
  - Also, this is normally done by a third party
- Retain chain of evidence
  - The evidence in this case is the data on the host
  - The log files of the compromised host hold the footsteps and fingerprints of the attacker
  - Every minute with that host must be accounted for
  - For legal reasons, you should examine a low-level copy of the disk and not modify the original

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

And finally, we have a post attack investigation. The forensics of the attacks the process is heavily guided by the laws that how this post attack or post mortem scenarios screaming there then retain chain of evidences that how things happens etcetera.

So, these are post mortem or post attack scenarios. Now if you look at in our in case of a cloud all these things also come in different in same or different forms right. Because these are more generic though we discussed at the end little bit of network related, but these are primarily more generating attacks. And then we have this post attack investigations to look at that what are the different attack pattern etcetera. And we will try to look at in our next lecture or so, that how what are the implications or what are the specialty of this security in case of cloud computing. So, will stop here today.

Thank you.

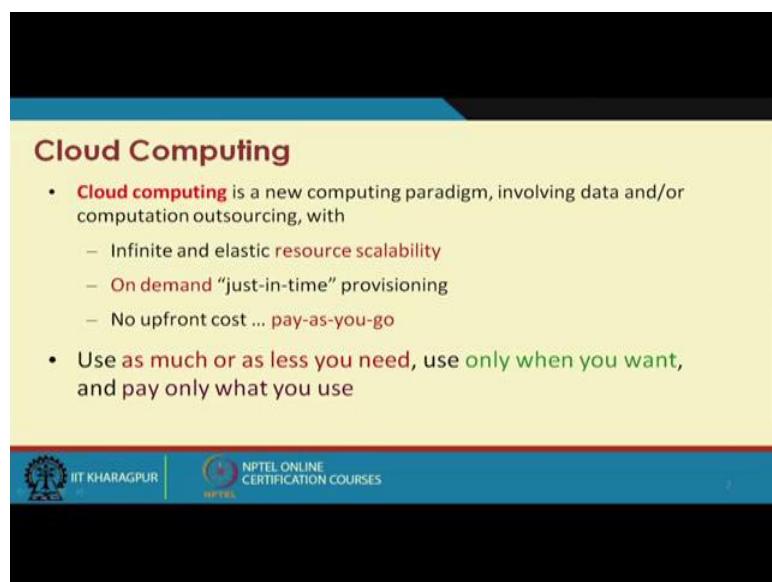
**Cloud Computing**  
**Prof. Soumya Kanti Ghosh**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**

**Lecture - 27**  
**Cloud Security – II**

Hi. Welcome to this Cloud Computing lecture series. Today, we will be discussing we will be continuing our discussion on Cloud Security. So, we will more now try to look at thus security with respect to more with respect to cloud perspective. So, as we have seen in our last lecture that the security has different aspects, right, like one is that security concepts or security components other part is the threats, there are issues of policies mechanisms there are issues of trust assumptions and those or risk.

So, all those things are need to be looked into when we are review on to implement the things. So, as it is; as we have to; as we have seen or discussed that these are manifested in any type of information system in including cloud, but cloud has different other some few more characteristics, right. So, we will try to see that what are the different characteristics and why this security becomes the important component; when we talk about cloud computing.

(Refer Slide Time: 01:38)



**Cloud Computing**

- **Cloud computing** is a new computing paradigm, involving data and/or computation outsourcing, with
  - Infinite and elastic **resource scalability**
  - **On demand** “just-in-time” provisioning
  - No upfront cost ... **pay-as-you-go**
- Use **as much or as less you need**, use only when you want, and pay only what you use

**IIT KHARAGPUR** | **NPTEL ONLINE CERTIFICATION COURSES**

So, as If we try to boil down cloud to a very simplistic are what we do one part is it is a resource elastic resource scalability, right, you can go up and go down in your resources

and resources can be anything right it is studying from computing power to memory to any type of a network resources bandwidth and so on. So, it is infinite and elastic resource scalability theoretically another thing is on demand just in time provisioning if I require, it should be on demand just in time provisioning should be there; there is another important aspect, thirdly it should be no its should be in a model what we say pay a metered service, right, pay as you go model, right as you use or as you go model, right.

So, when a what I pay for the things; that means, the resources are being acquired released escalated skill down as path the ways of the things this whole paradigm of security need to be over around these type of these policies, right; so that becomes a serious challenge and for that that number of organisation are not are little reluctant in going to the fully cloud even that is economically at times beneficial, right.

So, use as much as you use as much or as less as you need use only when want and pay only what you use this is the whole philosophy of going towards that.

(Refer Slide Time: 03:21)

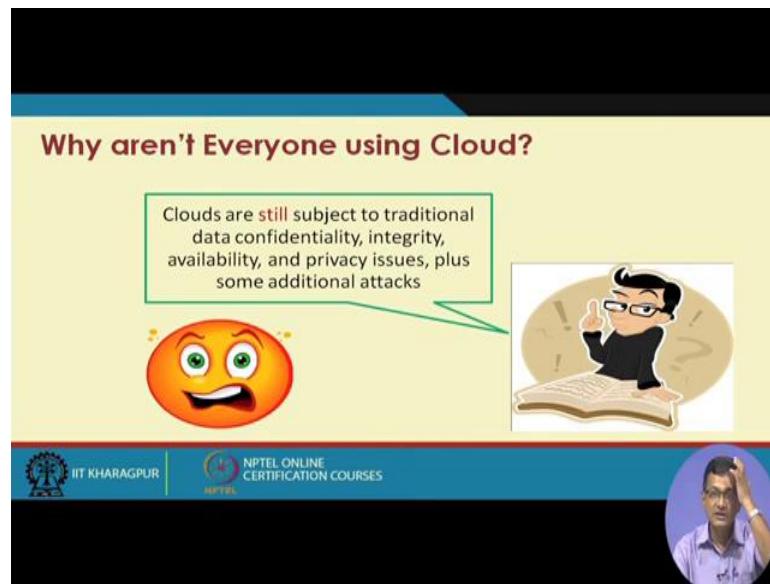
**Economic Advantages of Cloud Computing**

- For consumers:
  - **No upfront** commitment in buying/leasing hardware
  - Can **scale** usage according to demand
  - Minimizing start-up costs
    - Small scale companies and startups can reduce CAPEX (Capital Expenditure)
- For providers:
  - **Increased utilization** of datacenter resources

**IIT Kharagpur** | **NPTEL ONLINE CERTIFICATION COURSES** | 

So, if we already you have seen, but just to have a quick look economic advantage of cloud computing for the consumer no upfront cost can scale uses as and when required minimize that of course, for provider increased utilization of the data centre resources. So, provider has a huge volume of resources and that has a increase utilization of the resources is the one of the major aspects.

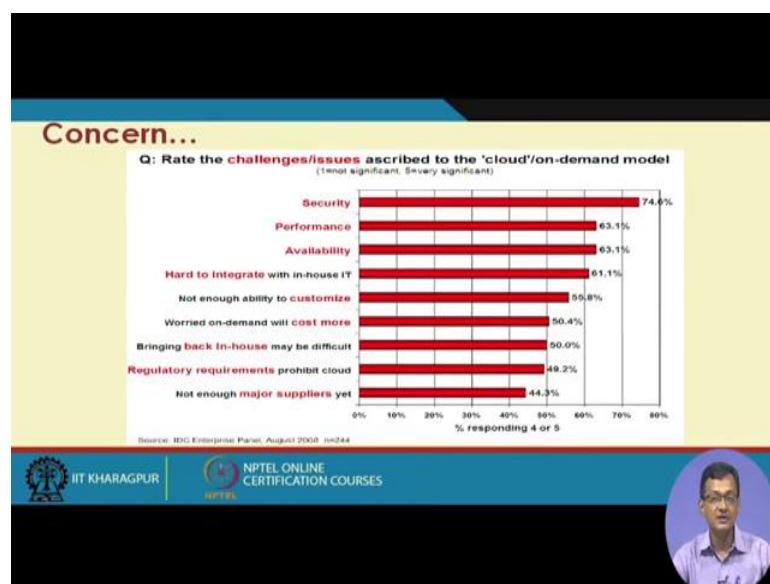
(Refer Slide Time: 03:46)



So, if it is win-win situation; why not everybody is using cloud, right.

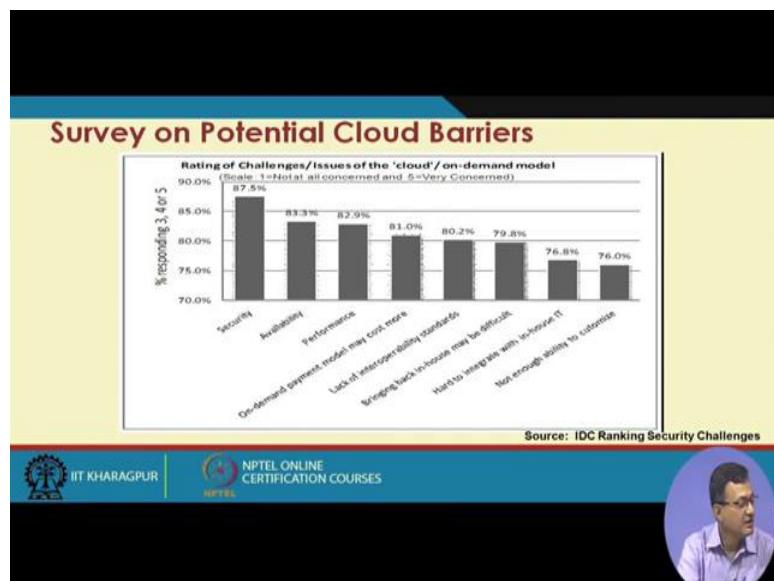
So, one of the major thing is that cloud are still subject to traditional data confidentiality, integrity, availability, privacy issues plus some additional attacks, right. So, this is a serious concern that any state. So, if it is if the provider is happy if the consumer is happy why not the things everything a going to cloud immediately because of this type of scenario because of SaaS scenario of that there are issues of data confidentiality integrity availability privacy, etcetera plus some additional cloud related things.

(Refer Slide Time: 04:28)



If we look some references that IDC inter price panel in 2008, they say that the major challenges or issue is the security right then the performance then availability and so and so forth, right. So, this is a from their survey it has been seen that the security still at the top level when you look at the challenges and issues in the things, it is not like that is insecure it is like at I am not able to define it is not only insecurity that thing, but also I failed to defined the things which we are define in more precisely there or I am not having that much trust or confidence on systems or the providers.

(Refer Slide Time: 05:09)



So, similarly survey on potential cloud barriers that also say that what we is blocking that going to the things is also that you look at it is a security plays a major role here.

(Refer Slide Time: 05:24)

**Why Cloud Computing brings New Threats?**

- Traditional system security mostly means keeping attackers out
- The attacker needs to either compromise the authentication/access control system, or impersonate existing users
- But cloud allows **co-tenancy**: Multiple independent users share the same physical infrastructure
  - An attacker can legitimately be in the same physical machine as the target
- Customer's **lack of control** over his own data and application.
- Reputation fate-sharing

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, what this may new threats come into play here one is the traditional systems security mostly keeps means keeping attackers out right. So, if I say that IIT Kharagpur need to be secured as a enterprise with this network and I have a very very very strong. So, that the attackers are out I have different mechanism to keep my internal attackers also out right maybe. But my concern is that how to keep these attackers out is the thing the attacker needs to either compromise the authentication or access control system or impersonate existing user in order to do that whereas, in case of cloud I voluntarily they provide consumer or the user voluntarily gives keep their data services etcetera at the providers place; that means, it is by nature it is co-tenancy is there.

So, it is cotenant. So, multiple impenent users share the same physical infrastructure. So, I know that the infrastructure I am sharing my some attacker or some other parties also sharing the infrastructure. So, attacker can legitimately use the same physical machine as the target it is not like that digit into or for into the things it can legitimately use customer's lack of control over his own data and application right it is on the premises of the service provider.

So, it is less control or lack of control over the services applications and there can be reputation fate sharing right it is a; what we say that I; if as I go together as we go together. So, I share the fate of each other, right. So, that is also there is also a challenge. So, these are the things if you see co tenancy lack of control reputation fate sharing these

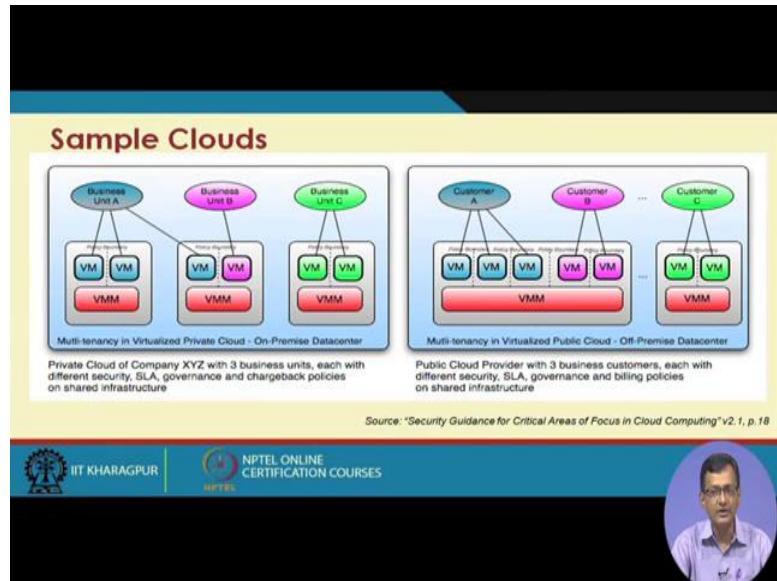
are the things which are not there in that a big way in case of a traditional things traditional security measures. And this becomes a new way of looking at the security in some cases.

(Refer Slide Time: 07:33)



Now, if we look at the different 3 prominent service model IaaS, PaaS and SaaS. So, in case of the IaaS, the infrastructure why is the provider is there rest of the thing is the responsibility of the consumer so; that means, the increase providers responsibility is whenever I go to IaaS, PaaS to SaaS, right, the provider has more responsibility or in other sense if I have the increase consumer responsibility. So, the when it goes the IaaS, it is the maximum, right. So, whenever somebody is taking IaaS, PaaS, etcetera, this one is definitely need the organizational of the individual need along with that need to look at that the type of security aspects now we need to need to deploy on my system.

(Refer Slide Time: 08:28)



Now, to typical scenario whatever we are discussing is one is that in case of private cloud say organisational cloud say IIT Kharagpur cloud. So, it has 3 business; you need business A, business B and business C and there is a chance that I the business A is sharing one of the VMs of the; in the data centre or the infrastructure where the business B is there where the C is isolated.

So, this type of scenarios are there so; that means, the services or data are residing on one physical or one or in the physical systems, right whereas, in case of a public; similarly for a public cloud I can have different customer who are sharing the same infrastructure and there is a possibility of a channel of communication between the things, it can be thing the VMM if might having compromised or there are some attack on though those things which are there. So, these are the there can be different type of things which is beyond a control of the consumer do not have or the cloud service consumer or the users do not have any control over this or not much control over this other than basically relying on the SLAs and the how the reporting of the providers are there.

(Refer Slide Time: 10:00)

**Gartner's Seven Cloud Computing Security Risks**

- Gartner:
  - <http://www.gartner.com/technology/about.jsp>
  - Cloud computing has “unique attributes that require risk assessment in areas such as data integrity, recovery and privacy, and an evaluation of legal issues in areas such as e-discovery, regulatory compliance and auditing,” Gartner says
- Security Risks
  - Privileged User Access
  - Regulatory Compliance & Audit
  - Data Location
  - Data Segregation
  - Recovery
  - Investigative Support
  - Long-term Viability

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, Gartner's have seven cloud computing security risk parameters, right. So, there is a Gartner's seven point things. So, rather Gartner's cloud computing according Gartner's has a unique attributes that require risk assessment in areas such as data integrity recovery and privacy and evaluation of legal issues in the areas the of e-discovery regulatory compliance and type of things, right.

So, these are five securities which Gartner point out in a report that is the one is privileged user access that is one securities regularity compliance and audit is another thing data location where the data is located with why whether; I how much control; I am having data segregation is another problem recovery mechanisms investigative support like if I want to do some post mortem type of things then how much investigated support what I am having and long term viability, right. So, there is a there is a chance of vendor locking then that will see that how that long and viability will be there.

(Refer Slide Time: 11:04)

**Privileged User Access**

- Sensitive data processed outside the enterprise brings with it an inherent level of risk
- Outsourced services bypass the “physical, logical and personnel controls” of traditional in-house deployments.
- Get as much information as you can about the people who manage your data
- “Ask providers to supply specific information on the hiring and oversight of privileged administrators, and the controls over their access,” Gartner says.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, if we look at user privileged access sensitive data process outside enterprise brings with it an inherent level risk; right, any sensitive data which is beyond going beyond your premise as a risk into the things outsourced services bypass the physical logical and personnel controls, right which is there if you are doing those of traditionally in house deployment. So, all these traditional in house deployments we have we bypass this. So, like a Gartner says that ask providers to supplies specific information on hiring and oversight of privileged administrator and controls over their access. So, that is there, but as such I a organization may feel insecure that while will it loses its number of controls to the provider.

(Refer Slide Time: 11:54)

**Regulatory Compliance & Audit**

- Traditional service providers are subjected to external audits and security certifications.
- Cloud computing providers who refuse to undergo this scrutiny are "signaling that customers can only use them for the most trivial functions," according to Gartner.
- Shared infrastructure – isolation of user-specific log
- No customer-side auditing facility
- Difficult to audit data held outside organization in a cloud
  - Forensics also made difficult since now clients don't maintain data locally
- Trusted third-party auditor?

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES



Next is the regulatory compliance and audit like traditional services are subject to external audits and security certification, right. So, so our traditionally in house services are there computing cloud computing provider who refuse to undergo a scrutiny signalling that the customer can only use them for the most trivial functions etcetera. So, in case of a cloud computing providers this type of things making audit etcetera become a tricky things right. So, because your data is there, but you do not have the control over the infrastructure. So, making the audit successful or compliance successful whether it will be compliance of that what the provider sends or what the provider supposed to do it or your compliance and things are like that. So, though the SLA tries to address this is but still there is a there are risk or what we say security loopholes there.

So, there are usually no customer side audit facilities difficult to audit data held outside organisation in a cloud trusted third party auditor maybe a thing then again how this auditor will be there and said that there is another question.

(Refer Slide Time: 13:09)

**Data Location**

- Hosting of data, jurisdiction?
- Data centers: located at geographically dispersed locations
- Different jurisdiction & regulations
  - Laws for cross border data flows
- Legal implications
  - Who is responsible for complying with regulations (e.g., SOX, HIPAA, etc.)?
  - If cloud provider subcontracts to third party clouds, will the data still be secure?

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES



Data location is a major issue, right where I share the data in the things where my data are hosted I do not have any clue whether in this country or outside country whether these jurisdiction of our own country or not or etcetera we do not wont state it and either types of things up we do not have think.

So, that becomes a major issue data centres located at graphically dispersed location different jurisdictions and regulations and legal implications these has different legal implications like say held data they keep of protected in UAS or other some other countries, but we do not have; we have a different type of things here and that it creates a problem that if are the data is store there that whose law will prevail on the thing.

(Refer Slide Time: 13:58)

## Data Segregation

- Data in the cloud is typically in a shared environment alongside data from other customers.
- Encryption is effective but isn't a cure-all. "Find out what is done to segregate data at rest," Gartner advises.
- Encrypt data in transit, needs to be decrypted at the time of processing
  - Possibility of interception
- Secure key store
  - Protect encryption keys
  - Limit access to key stores
  - Key backup & recoverability
- The cloud provider should provide evidence that encryption schemes were designed and tested by experienced specialists.
- "Encryption accidents can make data totally unusable, and even normal encryption can complicate availability," Gartner says.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES



Data segregation is another issue another which a pointed out by Gartner the data in the cloud is typically in a shared environment alongside data from other end customers, right encryption effective, but is in that cure all type of solution, right find out what is done to segregate data at rest.

So, encryption data encrypt data in transit needs to be decrypted at the time of processing another major issue, right. So, where the key will lie at types of things. So, there should be a secure key store resource the cloud provider should provide evidence that the encryption schemes were designed and tested by experienced specialist or what is the test mechanisms and what should the encryption scheme and type of things. So, these are several challenges which are data segregation related things which are not there in a big way when we have used additional systems.

(Refer Slide Time: 14:58)

**Recovery**

- Even if you don't know where your data is, a cloud provider should tell you what will happen to your data and service in case of a disaster.
- "Any offering that does not replicate the data and application infrastructure across multiple sites is vulnerable to a total failure," Gartner says. Ask your provider if it has "the ability to do a complete restoration, and how long it will take."
- Recovery Point Objective (RPO):** The maximum amount of data that will be lost following an interruption or disaster.
- Recovery Time Objective (RTO):** The period of time allowed for recovery i.e., the time that is allowed to elapse between the disaster and the activation of the secondary site.
- Backup frequency
- Fault tolerance
  - Replication: mirroring/sharing data over disks which are located in separate physical locations to maintain consistency
  - Redundancy: duplication of critical components of a system with the intention of increasing reliability of the system, usually in the case of a backup or fail-safe.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Another point and what we are concerned is the recovery right if something goes wrong what sees the recovery mechanism even if you do not know where the data your data is data providers is to tell you what happens to your data and services in case of a disaster if there is a disaster then or outrage; then what happened to my data, right a store I in a say share data storage I store my data and if goes for some problem, then what happened whether how much time it will take recovery at all whole recovery is possible or not these are the things which will be questioned, right.

So, there are 2 concepts if you will try to use one is the recovery point objective the maximum amount of data that will be lost has to follow a interruption or disaster. So, that is the RPO; recovery point objective; there is RTO; there is a period time period allowed for recovery that the time that is allow to elapse between the disaster and activation of the secondary side, right. So, that they how much time, even it is recovered how long it will take. So, that my business process not does not get much affected. So, fault tolerance 2 type of things, it is followed one is that replication that of the same thing or redundancy or duplication of critical components of the systems and type of things.

(Refer Slide Time: 16:25)

## Investigative Support

- Investigating inappropriate or illegal activity may be impossible in cloud computing
- Monitoring
  - To eliminate the conflict of interest between the provider and the consumer, a neural third-party organization is the best solution to monitor performance.
- Gartner warns. "Cloud services are especially difficult to investigate, because logging and data for multiple customers may be co-located and may also be spread across an ever-changing set of hosts and data centers."

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Then investigative support another risk component as mentioned by things like investigation investigating inappropriate or illegal activity may be impossible in cloud computing like how to investigate on the things especially there is not much control on the customer side. So, neither there is much control on monitoring the things.

(Refer Slide Time: 16:51)

## Long-term Viability

- "Ask potential providers how you would get your data back and if it would be in a format that you could import into a replacement application," Gartner says.
- When to switch cloud providers ?
  - Contract price increase
  - Provider bankruptcy
  - Provider service shutdown
  - Decrease in service quality
  - Business dispute
- Problem: vendor lock-in

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Long term viability; so, I leverage the things my work processes work flows or my deferent organisational processes into the cloud and I end up in a long term viability things or long term arrangement with the things, right.

Ask potential provider; how would you get your data back if it would be in a format that would import from a replacement application etcetera. So, if there is a; from one provider to another provider then how the data will be there and how data can I can recover my data if there is the if there is a problem with the provider. So, when to switch cloud provider contract price increase provider bankruptcy provider service shutdown decrease in service quality business dispute and all those things mainly to for thus consumer to switch cloud providers the major is vendor logging vendor lock in. So, that with the particular provider the consumer gets locked in and it is very difficult to recover from that lock in phase.

(Refer Slide Time: 18:01)

**Other Cloud Security Issues...**

- Virtualization
- Access Control & Identity Management
- Application Security
- Data Life Cycle Management

IT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, these are the major Gartner issues. So, there are few more issues which are which are critical which are critical. So, there is one is virtualization, access control and identity management, application security, data life data lifecycle management, right. So, one is the issue of the virtualization what you have seen the virtualization is primarily done by that VMM or the hypervisor, right.

So, the virtualization becomes the key of this cloud computing. So, if I have a VM so; that means, it is evolved from the basically handled by the VMM. Now if the VMM is compromised or then my I am in trouble even though even though different processes of the VM, etcetera to some level compromise then the whole system is in trouble.

(Refer Slide Time: 18:59)

**Virtualization**

- Components:
  - Virtual machine (VM)
  - Virtual machine manager (VMM) or hypervisor
- Two types:
  - **Full virtualization:** VMs run on hypervisor that interacts with the hardware
  - **Para virtualization:** VMs interact with the host OS.
- Major functionality: resource isolation
- Hypervisor vulnerabilities:
  - Shared clipboard technology—transferring malicious programs from VMs to host

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, if you look at the virtualization there are 2 component one is virtual machine one at VMM or the hypervisor or virtual machine monitor as we have seen. So, 2 type of primarily 2 type of virtualization one is full virtualization VMs run on the hypervisor that interacts with the hardware.

So, that the VM is there in between hypervisor and the rest of the hardware it interacts another is a para virtualization when a VM interacts with the host OS directly; that means, it penetrates to a level higher; so that 2 type of things major functionality resource isolation, right. So, what it tries to do it tries to isolate this consumer or the user with the rest of the infrastructure at the back bone and so that it basically tries to provide difference scalable services over the things, right. So, hypervisor vulnerabilities; now if there is a hypervisor vulnerability that will cropping and basically put the whole system in trouble; so, shared clipboard technology transferred malicious programs from VMs from VMs to the host and type of things. So, hypervisor vulnerability key stroke logging; so, 1 bun one such things that some VM technologies enable logging of key stores and the screen updates to be passed across virtual terminals in the single virtual machine.

So, these are some of the properties of the things and that becomes a threat right there are hypervisor risk like there can be a rogue hypervisor root kits initiate a rogue hypervisor and it its creates a havoc into the system hide itself from the normal malware detection

system can create a covert channel to dump unauthorised codes right it can even create a covert channel to dump with the unauthorised codes.

(Refer Slide Time: 20:59)

**Virtualization (contd...)**

- Hypervisor Risks
  - External modification to the hypervisor
    - Poorly protected or designed hypervisor: source of attack
    - May be subjected to direct modification by the external intruder
  - VM escape
    - Improper configuration of VM
    - Allows malicious code to completely bypass the virtual environment, and obtain full root or kernel access to the physical host
    - Some vulnerable virtual machine applications: Vmchat, VMftp, Vmcat etc.
  - Denial-of-service risk
- Threats:
  - Unauthorized access to virtual resources – loss of confidentiality, integrity, availability

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES  
NPTEL

There are other hypervisor risks like that external modifications of the hypervisor or VM escape in proper configuration of the VM. So, there can be other issues there are denial of services attacks.

So, there are issues of threats unauthorized access to virtual resources loss of confidentiality integrity availability and these are the different issues of these are the different threats which are there is a high loss of confidentiality integrity availability. That means, what we refer to these types of CIA related issues will come in to play.

(Refer Slide Time: 21:42)

**Access Control & Identity Management**

- Access control: similar to traditional in-house IT network
- Proper access control: to address CIA tenets of information security
- Prevention of identity theft – major challenge
  - **Privacy issues** raised via massive data mining
    - Cloud now stores data from a lot of clients, and can run data mining algorithms to get large amounts of information on clients
- Identity Management (IDM) – authenticate users and services based on credentials and characteristics

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Access control is a big gain as those who have gone through access control things like one is that troll base access control and different type of MAC; DAC type of things. So, those issues are there. So, access control similar to traditional in house it network in here also proper access control to address CIA tenets of information security, right. So, prevention of identity theft major challenge primarily privacy issues via massive data mining.

So, that I whether I can have some learning techniques and data mining techniques to find out the identity of the user or the cloud service consumer identity management is another challenge it is a challenge not only here it is a challenge across if any distributed or in type of system. So, identity management authenticate users and services based on credential and characteristics right. So, it based on different features said it tries to look at that um that I have to authenticate the users and services.

(Refer Slide Time: 22:56)

**Application Security**

- Cloud applications – Web service based
- Similar attacks:
  - **Injection attacks:** introduce malicious code to change the course of execution
  - **XML Signature Element Wrapping:** By this attack, the original body of an XML message is moved to a newly inserted wrapping element inside the SOAP header, and a new body is created.
  - **Cross-Site Scripting (XSS):** XSS enables attackers to inject client-side script into Web pages viewed by other users to bypass access controls.
  - **Flooding:** Attacker sending huge amount of request to a certain service and causing denial of service.
  - **DNS poisoning and phishing:** browser-based security issues
  - **Metadata (WSDL) spoofing attacks:** Such attack involves malicious reengineering of Web Services' metadata description
- Insecure communication channel

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, at the application level it is mostly there is cloud applications are web based; right.

Most of the applications are web based. So, similar type of attacks like injection attacks x xml signature element wrapping attack cross site scripting attack flooding DNAs poisoning and phishing metadata like WSDLs spoofing attacks; so, these are the different attacks which are still prevailed in case of a in case of application level cloud security, right. So, there can be insecure communication channel because at the application level your data is more vulnerable right. And that insecure communication channel can lead to interrupts and of the services eavesdropping and so and so forth.

(Refer Slide Time: 23:56)

**Data Life Cycle Management**

- Data security
  - Confidentiality:
    - Will the sensitive data stored on a cloud remain confidential?
    - Will cloud compromise leak confidential client data (i.e., fear of loss of control over data)
    - Will the cloud provider itself be honest and won't peek into the data?
  - Integrity:
    - How do I know that the cloud provider is doing the computations correctly?
    - How do I ensure that the cloud provider really stored my data without tampering with it?

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES



Data lifecycle management; so, need to look at that over all data lifecycle; so, one is that your confidentiality right will the sensitive data stored on cloud remains confidential that is one major question or major challenge we will cloud compromise leak confidential client data right fear of loss of control over the data. So, that is another problem will the cloud provider itself be honest and wont peek into the data that is a how much trust into the things. So, a trusting a provider is a is another challenge that is for of in our day to day life also if we need to trust or we need to build trust on deferent service provider.

So, there are lot of work going on; we will try to if time permits, we will try to see some of the aspects of this; how this task risk competence were together and we have a mechanism of a more security or how can I select a more trusted provider for a particular work; if there are more than one provider for that. So, that is one the confidentiality another aspect is the integrity; how do I know that the cloud provider is doing computations correctly right. So, I do some processing then how do I know that it is things because I push my data and process and I expect is result out of it.

How do I ensure that a cloud provider really stored my data without tempering it? So, how do I ensure that right availability?

(Refer Slide Time: 25:48)

**Data Life Cycle Management (contd.)**

- Availability
  - Will critical systems go down at the client, if the provider is attacked in a Denial of Service attack?
  - What happens if cloud provider goes out of business?
- Data Location
  - All copies, backups stored only at location allowed by contract, SLA and/or regulation
- Archive
- Access latency

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

With critical system go down at the client if the provider is attacked in a denial of service attack, right; so, this is another availability with the critical system go down at the client if the provider is at attacked in a denial if there is a dos type of a attack on the provider end; what will happen to my things right if that a even if the cloud provider goes out of business what will happen to my data and processes. So, these are very tricky issues and extremely difficult to address this type of challenges data locations as we have seen; if we look at the data lifecycle data location all copies beck up stored only at location allowed by the contract SLA or regulation, etcetera, right.

So, where the data are located which extension etcetera we do not have much control over the things then archive access latency these are the different other issues which are which are there in this type of scenarios. So, if we if we look at holistically that the overall cloud aspects. So, one major problem is co-tenancy; that means, you are your data processes are residing on the same system.

Another issue what we have seen that which is which is making it different from the traditional thing another issue is your data is located in somewhere where I do not have any control over the things data even my application function processes are located in the premises where I do not have much control other than looking at the SLAs and type of things. So, this is another major challenge of handling those type of a scenarios there are the other tricky issues which come up because if there are inter cloud communication,

then the issues are become more tricky, like a process at cloud 1 communicating to the cloud 2 communicating to the cloud 3 and so forth in doing. So, whether it is able to again that it is coming back to the originating cloud in doing. So, is it possible that I can there is a possibility or there is a chance that I violate the basic principle of access control like I am I am able to access a data which are otherwise I am not able to access it, right.

So, this is major challenge when there is a inter cloud communication things right. So, there is way that can be very much true because you are you have different provider consumers and a provider can be consumer for some other services and so on and so forth. So, that is another issues and there are other underlining threats like what will happen if the VM VMM is compromised if the hypervisor is compromised, then likely that all the VMs can be compromised, right or all the VMs are in a spin like which VM is up or down etcetera whether it is functioning properly or not we do not have any control.

So, there are underlining challenges at the IaaS level itself. So, these are the which need to be; access and finally, when selecting cloud providers or things; how can I trust each other, right how whether the SLA in the things or if I have more than one providers for a things whether there is a possibility or whether there is a mechanism that I can know that this is these are the different trust, etcetera.

So, the trust competence risk also plays a serious role into the things in looking at all these aspects we see that the cloud is this cloud security or the security issues in cloud plays a extremely vital role in making this cloud computing popular other than coming this coming that resource availability and other type of cross benefits of traditional versus cloud etcetera. This security issues become a major bottleneck going from say traditional to the cloud computing things.

So, with this we will wrap up our today's lecture and in the subsequent lecture we will see that other aspects of cloud.

Thank you.

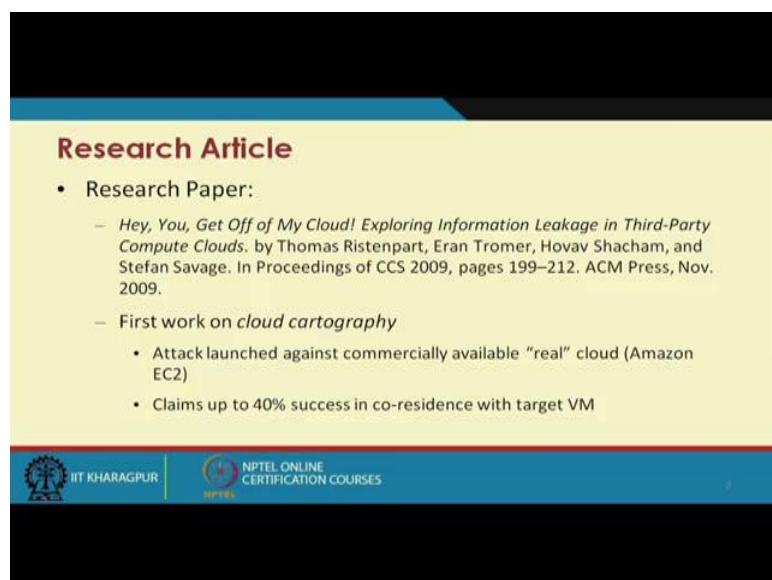
**Cloud Computing**  
**Prof. Soumya Kanti Ghosh**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**

**Lecture – 28**  
**Cloud Security – III**

Hi. We will be discussing on Cloud Computing, primarily looking into the Cloud Security. So, this may be this series of lectures third talk on cloud security. So, in this discussion, we will be basically trying to look at a case study taken from a well-known research article. It might so happen that many of you have gone through the article or if not it will be good to go through these articles that is pretty one of the interesting article which shows that where these security in cloud computing varies, or how it differ from our generic network or computing security or what are the extra things we need to look at when we look at the cloud security per say.

So, it may not be possible to go for all deep into the technical details of this article this such article, but I will try to give you that a overview of the problem which will help us in understanding that how security matters even when you are using is a standard secured, trusted, well used public cloud computing platform.

(Refer Slide Time: 01:51)



**Research Article**

- Research Paper:
  - Hey, You, Get Off of My Cloud! Exploring Information Leakage in Third-Party Compute Clouds. by Thomas Ristenpart, Eran Tromer, Hovav Shacham, and Stefan Savage. In Proceedings of CCS 2009, pages 199–212. ACM Press, Nov. 2009.
  - First work on *cloud cartography*
    - Attack launched against commercially available “real” cloud (Amazon EC2)
    - Claims up to 40% success in co-residence with target VM

 IIT KHARAGPUR |  NPTEL ONLINE CERTIFICATION COURSES

So, one that article we are talking about is came in ACM CCS 2009 and title says that hey you get off my cloud exploring information leakage in third party compute clouds

right. So, what they demand is the first work in cloud cartography, but apart from that it is interesting to see that what are the different way things will be there. So, our major objective of this particular discussion is to more look into the security aspects of the cloud; it is not on looking into any loophole of a particular cloud provider or security of any particular provider. We try to look at this paper which is there are some practical experiments which may help us in understanding the security aspect of the cloud in a better way that is our objective. It is not to analyze the work, but say but to take that work as an example case and see that how security it plays a important role in this cloud computing aspect.

So, the experiment that done in this particular work is this some sort of a so called quote unquote attack launched against a commercially available real cloud like typically Amazon EC2. And what they claimed that 40 percent are success in co-residence with the target VM right. So, if we remember our earlier lectures, what we are telling that one of the major issue is that if you whether we can co-residence a particular what we say attacking VM or a malicious users VM two way target VM, so that is a very challenging task. Because I really do not know how a cloud provider allocates the VM to other things.

(Refer Slide Time: 04:00)

**New Risks in Cloud**

- Trust and dependence
  - Establishing new trust relationship between customer and cloud provider
  - Customers must trust their cloud providers to respect the privacy of their data and integrity of their computations
- Security (multi-tenancy)
  - Threats from other customers due to the subtleties of how physical resources can be transparently shared between virtual machines (VMs)

 IIT KHARAGPUR |  NPTEL ONLINE CERTIFICATION COURSES



So, what we have seen in the new risk in cloud, one is that trust and dependency right. Establishing new trust relationship between the class customer and cloud provider that is

important because I am basically as a customer leveraging all my means most of my data most of my processes on the cloud and I am somewhat going dependent on that cloud infrastructure. So, customer must trust their cloud provider to respect the privacy of the data and integrity of the computations. So, when we look at the security point of view, the customer must trust the cloud provider for the preservation of the privacy of the data and the integrity of the computation. If there is a process, the process is supposed to be performing the way it is supposed to perform that is one of the objectives of any customer right, so that is expected.

Now, the other problem, so that is how much you trust and dependent on the thing; other thing is the multi tenancy right. Threats from other customer due to the say they are basically deciding on the same VMs and physical resource can be transparently shared. So, what is happening that the virtual machine what I have been allocated the in the same physical machine some other customers are also allocated. So, what is the chance that there is a path establishment between these two VM; and if there is a malicious VM or this processes running in a malicious VM, what is the chance that my VM or my process is likely to be compromised. So, that is the usual problem or multi tenancy when you have multi tenant data, these are the things which becomes a big issue.

(Refer Slide Time: 05:59)

**Multi-tenancy**

- Multiplexing VMs of disjoint customers upon the same physical hardware
  - Your machine is placed on the same server with other customers
  - Problem: you don't have the control to prevent your instance from being co-resident with an adversary
- New risks
  - Side-channels exploitation
    - Cross-VM information leakage due to sharing of physical resource (e.g., CPU's data caches)
    - Has the potential to extract RSA & AES secret keys
  - Vulnerable VM isolation mechanisms
    - Via a vulnerability that allows an "escape" to the hypervisor
  - Lack of control who you're sharing server space

 IIT KHARAGPUR |  NPTEL ONLINE CERTIFICATION COURSES

So, multi-tenancy as we have discussed earlier multiplexing VMs of disjoint customers upon the same on the same physical machine. So, your machine is placed in the same

server with other customer problem you do not have the control to prevent your instant from being co-resident with the some adversary instant. So, if it is a multi-tenant, so this multi-tenancy, how it will be residing that is at that is the logic the cloud provider doing. So, for the cloud provider point of view it is one of the major thing is resource management right it has a limited resource or it has a particular resource and it has to manage the thing, so that it optimize the performance of the customer level. So, based on that it basically try to try to deploy the VMs based on this analysis of that how resource can be properly managed and maximum performance level can be provided to the respective customer vis-a-vis their service level agreements right.

So, they here the with this with the multi tenancy some of the new risk factor came into picture, one is that slight channel exploitation that means, cross VM information leakage due to sharing of the physical resource, so that is another a big challenge. Across VM information leakage sharing of the physical resources is there. Has the potential to extract RSA and AES secret keys. We do not know, whether there is a potential to extract RSA and AES secret keys of this cross channel all this side channel exploitation. There are vulnerable VM isolation mechanisms like via a vulnerability that allows an escape to the hypervisor.

So, if there is a hyper, if there is a vulnerability, so if that can be exploited and this hypervisor some extent can be compromised. Lack of control you are sharing the server space. So, lack of control who you are sharing with. So, you do not have any control that who you are co-residing with right or this you are saying the simple way. So, these are the new risk which come into play.

(Refer Slide Time: 08:11)

**Attack Model**

- Motivation
  - To study practicality of mounting cross-VM attacks in existing third-party compute clouds
- Experiments have been carried out on real IaaS cloud service provider (Amazon EC2)
- Two steps of attack:
  - *Placement*: adversary arranging to place its malicious VM on the same physical machine as that of the target customer
  - *Extraction*: extract confidential information via side channel attack

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

And the attack model specifically the attack model which in being also followed in this particular work we are what we are discussing is that the one of the motivation of this attack model is to study the practicality that whether it is practical or mounting cross VM attacks in existing third party compute clouds. So, if I am having existing third party compute cloud, is it possible to do to launch some sort of a cross VM attacks right if on the things right. So, what they did the experiments have been carried out on realize cloud provider like I am as an EC2 and this can be carried out to any type of IS provider.

So, there are two steps and this is these two steps are irrespective whether is a cloud or network or anything that is one is placement, adversary arranging to place the malicious VM on the same virtual machine as that of the target customer this is important. If I want to do across some sort of attack or something, first thing I have to do is that whether I can physically place my VM into the same space or the same physical machine or the same server space where the adversaries machine is there. So, that is one important thing And secondly, it is a extraction thing. So, once I am placed, so extract confidential information by side channel attack. So, these are the two type of.

(Refer Slide Time: 09:45)

**Threat Model**

- Assumptions of the threat model:
  - Provider and Infrastructure to be trusted
  - Do not consider attacks that rely on subverting administrator functions
  - Do not exploit vulnerabilities of the virtual machine monitor and/or other software
  - Adversaries: non-providers-affiliated malicious parties
  - Victims: users running confidentiality-requiring services in the cloud
- Focus on new cloud-related capabilities of the attacker and implicitly expanding the attack surface

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, next is the threat model like assumption of threat model is that the provider and the infrastructure to be trusted right. So, what we do when we this is one of the basic assumption is the provider and the infrastructure need to be trusted, do not consider attack that that rely on subverting the administrator functions all right do not exploit vulnerabilities of the virtual machine monitor or others software. So, it will not exploit that hypervisor and other things.

So, adversaries non-providers affiliated malicious parties. So, advisories are not provider affiliated they came as a user or customer. Victims, user running confidentiality requiring services of the cloud. So, victims are running some of the operations which needs some basic privacy and confidentiality is not public operations of public data and services through the cloud. So, focus on new cloud related capabilities of the attacker and implicitly expanding the attack surface right. So, we try to see that what are the things and try to see that what type of other attacks.

(Refer Slide Time: 10:54)

**Threat Model (contd...)**

- Like any customer, the malicious party can run and control many instances in the cloud
  - Maximum of 20 instances can be run parallel using an Amazon EC2 account
- Attacker's instance might be placed on the same physical hardware as potential victims
- Attack might manipulate shared physical resources to learn otherwise confidential information
- Two kinds of attack may take place:
  - Attack on some known hosted service
  - Attacking a particular victim's service

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, there are other threat models consideration like any customer a the malicious party can run and control many instances of the cloud, so that is another thing. Attackers instance might be placed on the same physical hardware as the potential victims are. Attacks might manipulate the shared physical resource to learn otherwise confidential information, so that attacker can basically do some surface cause VM attacks. So, two type of attacks can take place, attack on some known hosted services or attack on a particular victim services. So, one attack is that I know that these are the hosted services I want to attack on the things or I want to have particular victim service to be attacked. So, it is more what we say targeted attack.

(Refer Slide Time: 11:45)

**Addresses the Following...**

- Q1: Can one determine where in the cloud infrastructure an instance is located?
- Q2: Can one easily determine if two instances are co-resident on the same physical machine?
- Q3: Can an adversary launch instances that will be co-resident with other user's instances?
- Q4: Can an adversary exploit cross-VM information leakage once co-resident?

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, in order to do that, so what they proposed or what they did is basically need to answer a few questions, one is that can one determine where the cloud infrastructure and instance is located right. So, is it possible to determine that where a particular instances located, very, very difficult not only difficult, something apparently impossible. You take a login from Amazon or Azure or Google platform or any other sales force or anything and that is they are way of handling the things like at the back of his management or the backbone management things. Question two, can anyone can one easily determine if two instances are co resident on the same physical machine right. One is that finding that where the one instance is there, another is that whether it is possible that I can determine that whether these two instances are co-resident right.

Number three, is that can a adversary launch instances that will be co resident with the user instances right. So, other question is that whether the adversaries launch instances, so that you want to co-residents with some targeted instances. And number four can an adversary exploit cross VM information leakage one co resident. So, if it is a co-resident somewhere rather whether there is a possibility that then I can have a cross VM information leakage. So, these are what they did the experiment or what we are trying to it get overview of the whole thing is primarily working on one of the very popular cloud provider, and it follows all possible best practices. Even with that, whether it is possibilities they are or not that is that thing what we are trying to look at.

(Refer Slide Time: 13:42)

**Amazon EC2 Service**

- Scalable, pay-as-you-go compute capacity in the cloud
- Customers can run different operating systems within a virtual machine
- Three degrees of freedom: *instance-type, region, availability zone*
- Different computing options (instances) available
  - m1.small, c1.medium: 32-bit architecture
  - m1.large, m1.xlarge, c1.xlarge: 64-bit architecture
- Different regions available
  - US, EU, Asia
- Regions split into availability zones
  - In US: East (Virginia), West (Oregon), West (Northern California)
  - Infrastructures with separate power and network connectivity
- Customers randomly assigned to physical machines based on their instance, region, and availability zone choices

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, if you look at the Amazon EC2 service per say. So, it is a scalable pay as you go compute capacity of the cloud customer can run different operating system within the virtual machine, three degree of freedom instance type, region and availability zone. So, when you do select. So, there are three degree of freedoms. So, you can have instance type what type of instant or which region you want to launch, and the availability zone in the things. So, three computing options instances are available one is m1 small, m1 medium 32 bit architecture, m1 large, m1 extra large and so on and so forth. So, these are the different instances are available.

So, there are different region available right US, EU and Asia, this is the time one what the paper tells that is what when they are the came up in 2009. So, region split into availability zone. So, if you look at that UCI in us it has a east in the East Virginia, West Oregon and west another thing is Northern Carolina right. So, similarly infrastructure will separate power and network capacity connectivity. So, these are the different physically different located different places. So, they have not only different power line, that means, they are not on the same power backbone that in other sense that is they are not subject to failure if there is a power failure of one instances, and it is likely they are network also is different. So, that means, the IP block used in one will be different another end type of things. Customers randomly assigned to a physical machine based on their instance region and availability zone choices. So, customer has this option of choice

in taking a make a choice of this, and based on that they are given to the different machine.

(Refer Slide Time: 15:43)

**Amazon EC2 Service (contd...)**

- Xen hypervisor
  - Domain0 (Dom0): privileged virtual machine
    - Manages guest images
    - Provisions physical resources
    - Access control rights
    - Configured to route packets for its guest images and reports itself as a hop in traceroutes.
  - When an instance is launched, it is assigned to a single physical machine for its lifetime
- Each instance is assigned internal and external IP addresses and domain names
  - External IP: public IPv4 address [IP: 75.101.210.100/domain name: ec2-75-101-210-100.compute-1.amazonaws.com]
  - Internal IP: RFC 1918 private address [IP: 10.252.146.52/domain name: domU-12-31-38-00-8D-C6.compute-1.internal]
- Within the cloud, both domain names resolve to the internal IP address
- Outside the cloud, external name is mapped to the external IP address

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, typically Amazon EC2 service using hp hypervisor and if you a XEN hypervisor sorry; so if you look at the XEN hypervisor, so there is a Dom0 what we say portion what is the privileged virtual machine which have manages guest images, it provisions physical resources access control rights configure to route packets in its guest images and reports itself as a hop to the trace route right. So, it routes the things and that can get it is a hop on the thing. So, when an instant is launched, it is assigned to a single physical machine for its lifetime. So, the instant particular anything.

So, secondly, each instance is assigned internal and external IP address and a domain names, so that is the philosophy of Amazon. So, external IP address something internal IP address based on some standard and respective domain name. Within the cloud both the domain names resolve the internal IP address. So, within the cloud both whatever you have seen internal external add thing outside the cloud external name is made to the external ip address. So, when we go to the outside the cloud then the external name is mapped to the external IP address.

(Refer Slide Time: 17:01)

**Q1: Cloud Cartography**

- Instance placing is not disclosed by Amazon but is needed to launch co-residency attack
- Map the EC2 service to understand where potential targets are located in the cloud
- Determine instance creation parameters needed to attempt establishing co-residence of an adversarial instance
- Hypothesis: *different availability zones and instance types correspond to different IP address ranges*

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES



Now, if we look at the is different aspects or the different queries which we which we raised or where which the article raised that try to address those and those queries. The query one is the cloud cartography, instant placing is not disclosed by Amazon, but is needed to launch whole residency attack. So, if I want to do some sort of a co-residency attack then I require the instant to be placed into the victims things, but the Amazon will definitely not disclose this. And map the EC2 service to understand where the potential targets are located in the cloud. So, we need to map or what they have shown they have tried to map the EC2 service to understand that where the targets are located in the cloud.

So, determine the instance creation parameters needed to attempt establishing co-residence at an adversarial instance. So, it is needs a create a parameter to attempt establishing a co-residence on the thing. And the basic hypothesis different availability zone and instance types correspond to different IP address right. So, if I have different availability zone and different instance types, it is likely that they are in the different IP ranges. So, whether we are able to whether we can there is a possibility of exploiting this.

(Refer Slide Time: 18:29)

## Network Probing

- Identify public servers hosted in EC2 and verify co-residence
- Open-source tools have been used to probe ports (80 and 443)
  - nmap – perform TCP connect probes (attempt to complete a 3-way hand-shake between a source and target)
  - hping – perform TCP SYN traceroutes, which iteratively sends TCP SYN packets with increasing TTLs, until no ACK is received
  - wget – used to retrieve web pages
- External probe: probe originating from a system outside EC2 and has an EC2 instance as destination
- Internal probe: originates from an EC2 instance, and has destination another EC2 instance
- Given an external IP address, DNS resolution queries are used to determine:
  - External name
  - Internal IP address

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, in order to do that those who have worked on network security or networking per se you know that there are different type of network probing tools are available. So, network probes are available and which are many of them are open source and fairly able to map the thing. So, similarly here also we require a network probing. Identify public servers hosted in EC2 and verify the co-residency. So, open source tools have been used to probe the ports port 80 and 443 that is the http and https secure port right RSN port. So, because these are these are mostly used for external access and likely that they will be opened and allowed the things.

So, one such tool is nmap, other is hping and wget right. So, these are the three popular there are several others tools and it is sometimes someone can write their own tool and type of things, but they are using the tool of probe. So, external probe, probe originating from a system outside EC2 and has an EC2 instance as the destination, so that can be the external. And internal originates from EC2 instant and it has destination another e c two instance right. So, this is internal and external thing. So, given external IP, DNS resolution queries are used to determine external name and internal IP address right. So, this is by the DNS query

(Refer Slide Time: 20:03)

**Survey Public Servers on EC2**

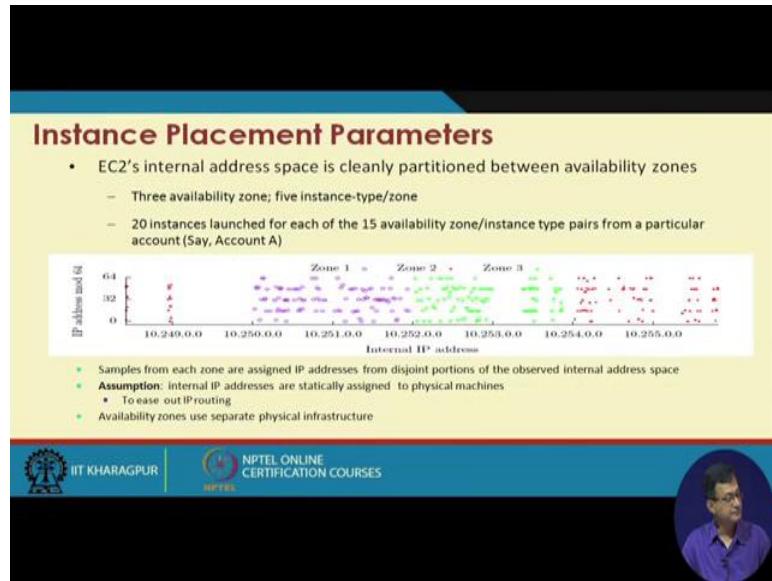
- Goal: to enable identification of the instance type and availability zone of one or more potential targets
- WHOIS: used to identify distinct IP address prefixes associated with EC2
- EC2 public IPs: /17, /18, /19 prefixes
  - 57344 IP addresses
- Use external probes to find responsive IPs:
  - Performed *TCP connect probe* on port 80
    - 11315 responsive IPs
  - Followed up with *wget* on port 80
    - 9558 responsive IPs
  - Performed a *TCP scan* on port 443
    - 8375 responsive IPs
- Used DNS lookup service
  - Translate each public IP address that responded to either the port 80 or 443 scan into an internal EC2 address
  - 14054 unique internal IPs obtained

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, survey a public server on EC2 because to have a if we survey a goal to enable identification of instant type and availability zone of one or more potential targets. So, our primary or the primary goal of this particular work is to whether I can basically identify the instant type and availability zone of one or more potential targets that is one of the major thing. EC2 public IPs are in this prefixes like and there are public IPs of those tuned as reported by the particular article use external probes to find the responsive IPs right which are responsive for from TCP connect probe on port 80 and followed by wget port at port 80 and performed TCP scan at port 443 and then they see that what are the IPs which responses.

So, use DNS lookup translate each public IP that correspond to either port 80 or port 443 to an internal EC 2 address and then do again the probing on the things. So, some 14,000 odd unique internal IPs are obtained

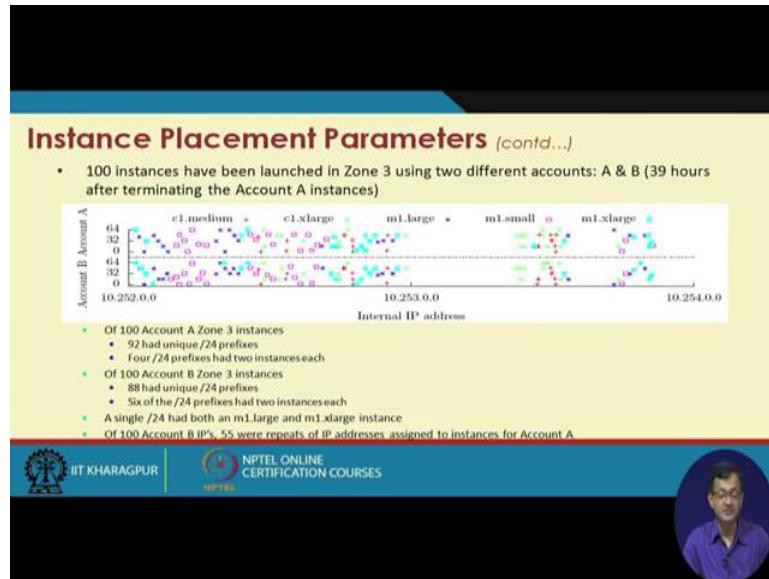
(Refer Slide Time: 21:20)



Now, next is the instance placement parameter, what parameter you need to say that the things. Now, EC2's internal address space is cleanly partitioned between the availability zone. They are partitioned into three availability zone five instance type of instance type and zone as we have seen. 20 instance launched for each of the 15 what the experimental things they have done, and they have shown that samples from each of the zone are assigned IP address from disjoint portion of the observed internal IP address spaces. Assumption, internal IP address are statically assigned to physical machines to ease out IP routing otherwise it will again routing parameters will be there. Availability zone are used physical infrastructure. So, these are the things which are there.

So, in other sense what we try holistically see there are these are the different zones and different type of instances they are on different IP block. In other sense if I somehow select the same zone and etcetera whom I am targeting, it is likely I may be in the same IP block. If I know that in the same IP block then whether it is possible to launch again some of the probes and some of the attacks with the same IP blocks.

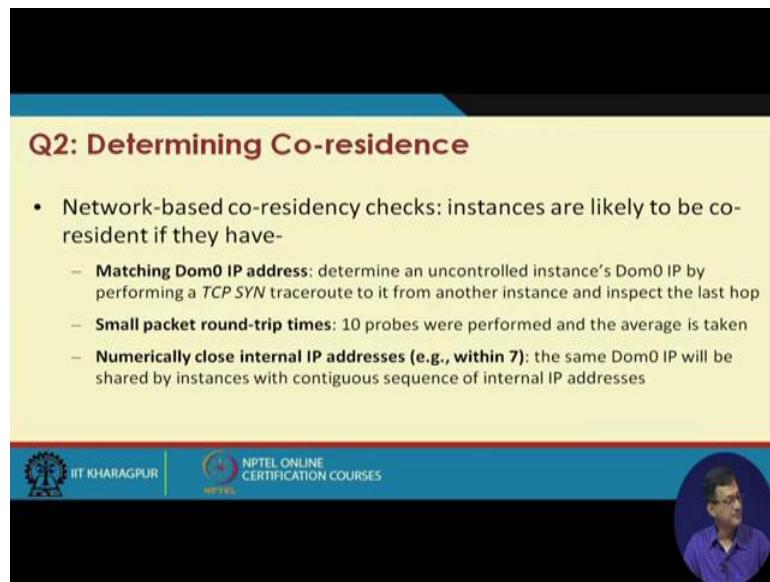
(Refer Slide Time: 22:49)



So, they what they experimentally shown that hundred instances has been launched in zone three using two different account A and B, 39 hours after terminating the account instance A and of hundred zone 3, 92 has a unique slash 24 prefixes, 4 prefixes has two instances each. So, these are their resultant. Out of 100 b zone three instances 88 unique and 6 this. A single slash 24 had both m1 large and m1 extra large instance. Of 100 accounts of these IPs 55 were repeats of IP address assigned to the instance of that account A that is interesting. So, out of that assign thing which are A and B.

So, what they are launched in different time scale like A and after terminating 39 hours of B, so I can I got that address etcetera. So, if you look at it gives some sort of it tries to give in some may be very grossly some cartography of or in other sense that the IP address blocks and etcetera how they are spared and so on and so forth.

(Refer Slide Time: 23:58)



**Q2: Determining Co-residence**

- Network-based co-residency checks: instances are likely to be co-resident if they have-

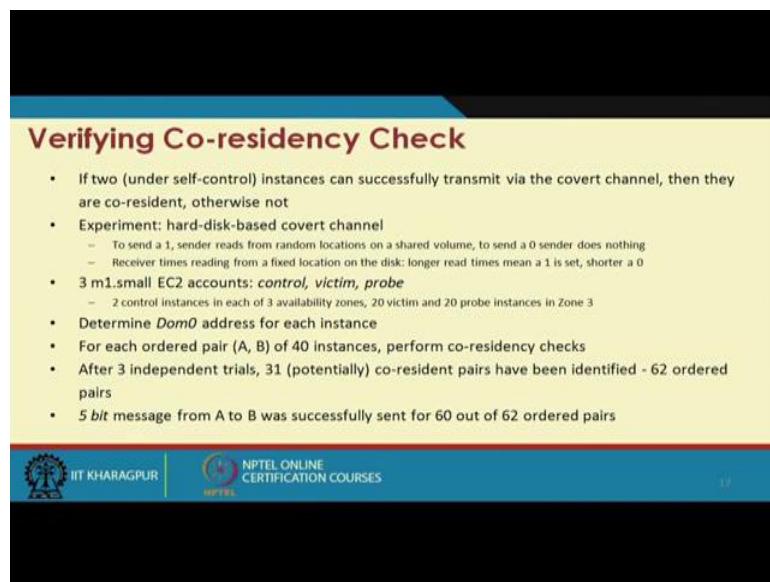
  - Matching Dom0 IP address:** determine an uncontrolled instance's Dom0 IP by performing a TCP SYN traceroute to it from another instance and inspect the last hop
  - Small packet round-trip times:** 10 probes were performed and the average is taken
  - Numerically close internal IP addresses (e.g., within 7):** the same Dom0 IP will be shared by instances with contiguous sequence of internal IP addresses

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES



Now, if I have this type of things like roughly know that these are the IP address blocks and type of things are there then whether determining co-residency can help or what can be done. So, network based co-residency at a checks instances likely to be co-residence if they are matching Dom0 IP address that as we have seen that domain zero is that primarily do we their management part. Small packet rounder trips if I have a round-trip times, so that will be the small packet in is in the same block, numerically close IP range typically within seven, so that is another things what we are having.

(Refer Slide Time: 24:43)



**Verifying Co-residency Check**

- If two (under self-control) instances can successfully transmit via the covert channel, then they are co-resident, otherwise not
- Experiment: hard-disk-based covert channel
  - To send a 1, sender reads from random locations on a shared volume, to send a 0 sender does nothing
  - Receiver times reading from a fixed location on the disk: longer read times mean a 1 is set, shorter a 0
- 3 m1.small EC2 accounts: *control, victim, probe*
  - 2 control instances in each of 3 availability zones, 20 victim and 20 probe instances in Zone 3
- Determine Dom0 address for each instance
- For each ordered pair (A, B) of 40 instances, perform co-residency checks
- After 3 independent trials, 31 (potentially) co-resident pairs have been identified - 62 ordered pairs
- 5 bit message from A to B was successfully sent for 60 out of 62 ordered pairs

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES



So, verifying co-residency is check is another challenge that if two under self control instances are successfully transmit via the covered channel they are co-residents and so on and so forth. You can, if you can connect them experiment the hard disk base covert channels they have shown. So, three m one small account control victim probe what they need determine dom0 address for each pair of A B. So, what they try to do that what is a checking the co-residency of that 2 to VMs and type of things

(Refer Slide Time: 25:21)

### Effective Co-residency Check

- For checking co-residence with target instances:
  - Compare internal IP addresses to see if they are close
  - If yes, perform a TCP SYN traceroute to an open port on the target and see if there is only a single hop (Dom0 IP)
    - Check requires sending (at most) two TCP SYN packets
    - No full TCP connection is established
    - Very "quiet" check (little communication with the victim)

So, effective co-residency check for checking constancy with the target instances compare internal IP address to see if they are closer. So, if it is within the within the typical seven thing then it is a closer if he has performed TCP sync trace route to open port to the target and see and see if there is only one single hop dom0 IP. Check requires sending at most two TCP SYN packets, no full TCP connection is established very quiet check little communication with the victim. So, these are the checks which are done known some sort of a quote unquote non invasive manner that means, the victims is not aware of the things. So, you are basically cross checking that whether it is some co-residency thing.

(Refer Slide Time: 26:13)

**Q3: Causing Co-residence**

- Two strategies to achieve “good” coverage (co-residence with a good fraction of target set)
  - Brute-force placement:
    - run numerous *probe* instances over a long period of time and see how many targets one can achieve co-residence with.
    - For co-residency check, the probe performed a wget on port 80 to ensure the target was still serving web pages
    - Of the 1686 target victims, the brute-force probes achieved co-residency with 141 victim servers (8.4% coverage)
    - Even a naive strategy can successfully achieve co-residence against a not-so-small fraction of targets
  - Target recently launched instances:
    - take advantage of the tendency of EC2 to assign fresh instances to small set of machines

And the third thing is causing co-residency two strategies to achieve good coverage co-residency with a good functions target one is brute force placement you want to do some brute force placement of the things by run numerous probe instances and find out that where things are there, and you do a brute force. Other is target recently launched instances take advantage of the tendency of EC2 to assign fresh instance a small set of machines.

So, if it is after studying etcetera that is it can be think that the service provider are doing that is that close which are very instances launched within a particular small time span are placed into the thing into the same type of hardware or server. Then there is a chance of doing a target recently launched, instances and try to co-residency.

(Refer Slide Time: 27:05)

## Leveraging Placement Locality

- Placement locality
  - Instances launched simultaneously from same account do not run on the same physical machine
  - *Sequential placement locality*: exists when two instances run sequentially (the first terminated before launching the second) are often assigned to the same machine
  - *Parallel placement locality*: exists when two instances run (from distinct accounts) at roughly the same time are often assigned to the same machine.
- *Instance flooding*: launch lots of instances in parallel in the appropriate availability zone and of the appropriate type

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES



So, leveraging placement of locality, one is the placement of locality instances launched simultaneously from the same account, do not run on the same physical machine. Sequential placement locality existing when two instances run sequentially as we have seen that previously A and B. Parallel placement, locality exist when two instances run at roughly the same time are often assigned to the same machine. So, these are the things which can be exploited. And there is a other ways that instance flooding launch lots. So, parallel instances in the appropriate availability zone and appropriate type and try to see that what they happen.

(Refer Slide Time: 27:53)

## Q4: Exploiting Co-residence

- Cross-VM attacks can allow for information leakage
- How can we exploit the shared infrastructure?
  - Gain information about the resource usage of other instances
  - Create and use covert channels to intentionally leak information from one instance to another
  - Some applications of this covert channel are:
    - Co-residence detection
    - Surreptitious detection of the rate of web traffic a co-resident site receives
    - Timing keystrokes by an honest user of a co-resident instance

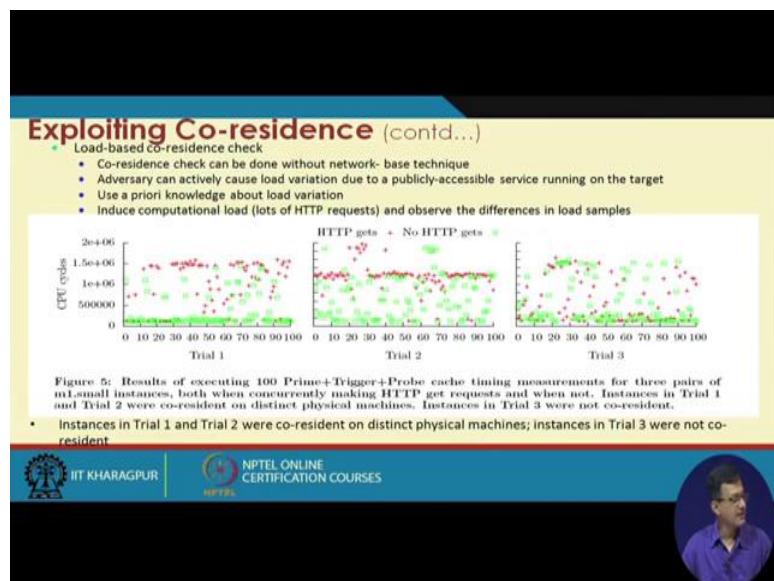
IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES



Similarly, they did a lot of experimentation to see that how this locality can be exploited. And finally, exploiting co-residents that is cross VM attacks can allow information leakage how can we exploit shared infrastructure is one that it means if I am co-resident then how to exploit this like gain information about the source uses of the other instances. Create and use covert channels to intentionally leak information from one instance to other. So, these are the other things which can be exploited by the attacker. Some application of these covert channels are co residence detection whether it is the thing that is whether I can have some secret detection scheme to look at it whether timing of the keystroke allow me to look at the password and so on and so forth.

And other type of techniques which are there in other cases other this type of covert channel attacks is one is that measuring the cache uses that and try to see that the what is the normal pattern and whether there is a any attacks etcetera there. And try to map that what sort of processes are they are based on the cache uses pattern.

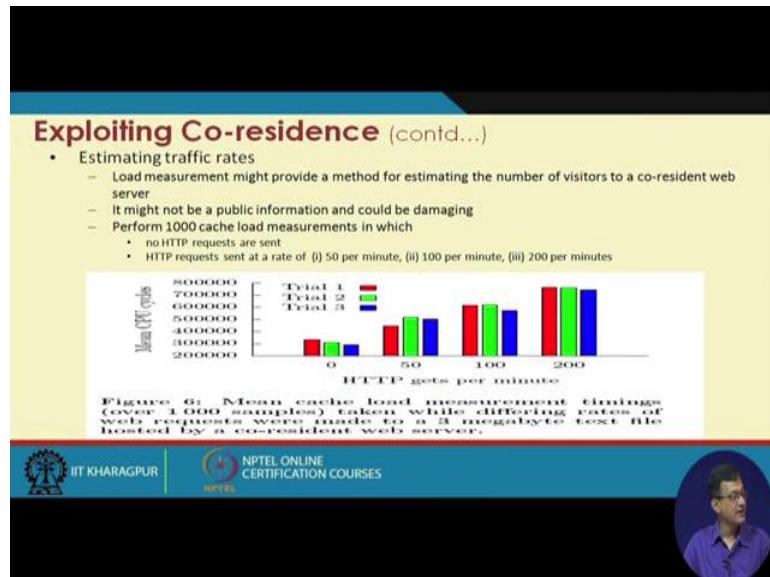
(Refer Slide Time: 29:08)



So, one is that exploit a load-based co-residence checking that exploit in co-residency. So, co-residency check can be done without network base network or adversary because I am co-residence. So, I do not require again now the resident a bigger other network infrastructure I am on the same machine. So, here they have shown this with the experiment that the trial one and two were co-resident on distinct physical machine

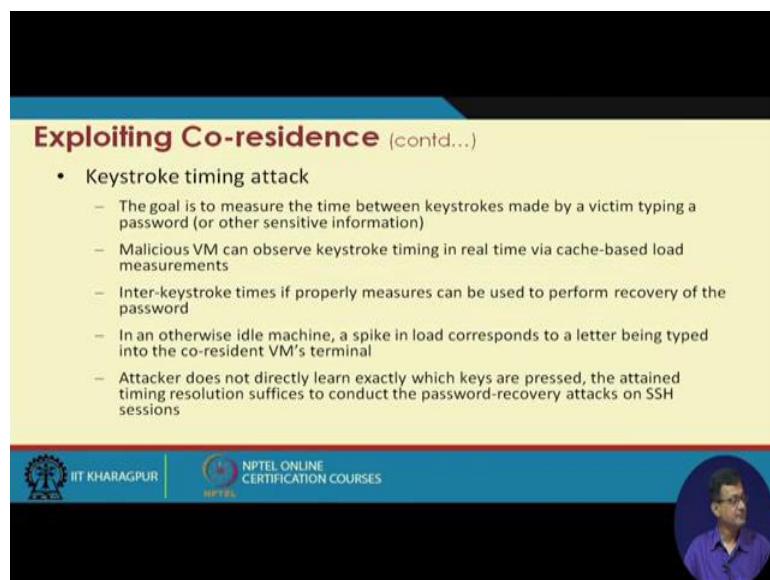
instant, three were not co resident. So, that what if you look go through the paper there this has been shown.

(Refer Slide Time: 29:42)



And it has been shown that estimation traffic rates that what sort of traffics are there with no http, with http connection and so on and so forth. These are the figures and data taken from that particular research article.

(Refer Slide Time: 29:57)



And other things are keystroke timing attacks right. So, based on that keystroke and this is a very popular or very well known type of things based on that whether you can basically look at that passwords and against the password and try type of things.

(Refer Slide Time: 30:18)

**Preventive Measures**

- Mapping
  - Use a randomized scheme to allocate IP addresses
  - Block some tools (nmap, traceroute)
- Co-residence checks
  - Prevent identification of Dom0
- Co-location
  - Not allow co-residence at all
    - Beneficial for cloud user
    - Not efficient for cloud provider
- Information leakage via side-channel
  - No solution

So, finally that whether the what type of preventive measures we can think of one is the mapping use randomized scheme to allocate IP address block some tools like in map and trace route. So, blocking tools may be, but randomized allocation may basically may be going against and some cases that the optimization or resource management of the things right. So, what they are from the prospective of the service provider, so that is that may be a challenge. So, co-residency check prevent identification of the dom0 that may be one of the way. So, what of that, so that that they cannot check that what is the domain zero.

Co-location not allow co residence at all right so, that means, beneficial for cloud user, but not efficient definitely not efficient for the cloud provider. And information leakage via side channel still is a big challenge like different type of sites challenge things are there and it is not only that you create a covert path like that you basically judge different other parameters look at like maybe the as well looking at that the looking at the cache behavior or the how the cache uses pattern, I am trying to look at that what sort of activities are going on.

So, with this we end our talk today that that new risk from the summary of that security thing that new risks from cloud computing they are which is little different from our conventional computer or information or network security. Shared physical infrastructure may and most likely will cause problems exploiting software vulnerabilities are not addressed properly here. Practical we have not there may be some software vulnerability if that the SaaS level cloud etcetera which are not being addressed. Practical attacks are in that particular paper they have shown that particular attacks are performed and some counter measures also proposed in this work

So, I encourage you to go through this paper, so that it is a good again I am repeating it is not to particularly look into particular service provider or look at the they are loopholes etcetera, more things we want to see a look at an overview that these are the things possible or this at the things which open up new risk etcetera, which are not there in our traditional network or computer or information security.

So, with this we will end our talk today.

Thank you.

**Cloud Computing**  
**Prof. Soumya Kanti Ghosh**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**

**Lecture – 28**  
**Cloud Security – IV**

Hello. We will be continuing our discussion on Cloud Computing. Today we will talk about some aspects of a cloud security rather we will look at a sort of a scenario where how this security plays a role and what are the different aspects. This is primarily with the SaaS a type of cloud more of collaborating SaaS clouds. So, this what we are looking that you had presently or in near future that lot this sort of clouds will be communicating between each other; that means, in other sense this a this consumer or different stakeholders having their application in the in the cloud will be communicating with other applications in other cloud.

So, in other sense it is a collaborative SaaS cloud or collaborative collaboration at the application level of the cloud. So, today's discussion will be looking at one of this what are different security aspects when we collaborate between each other we will see that there are very tricky issues of each comes into play. So, this one of this approach this is work of one of my PhD scholar Nirnoy Ghosh a goes he his work will be describing, but we will be looking at more broader aspects that how things should be there.

So, this will be good to for many of you who are looking at some sort of a research or some sort of a more studying into these aspects of the things.

(Refer Slide Time: 02:11)

The slide has a dark blue header and footer. The main content area is light yellow. The title 'Security Issues in Cloud Computing' is in bold red font. Below it is a bulleted list of security concerns. At the bottom, there are logos for IIT Kharagpur and NPTEL, along with the text 'NPTEL ONLINE CERTIFICATION COURSES'.

**Security Issues in Cloud Computing**

- Unique security features:
  - Co-tenancy
  - Lack of control on outsourced data and application
- General concerns among cloud customers [Liu'11]:
  - Inadequate policies and practices
  - Insufficient security controls
- Customers use cloud services to serve their clients
- Need to establish trust relationships
- Beneficial to both stakeholders

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, it is security issues in collaborative SaaS cloud. So, as a just to recap if you look the look at the security issues in cloud computing. So, which are typically or unique to this cloud is one is co tenancy right. Numbers of applications are residing or different user are residing in the same physical infrastructure. So, co tenancy is a major issue. And lack of control on outsource data and applicants that is another typically or uniqueness of this type of cloud platform right.

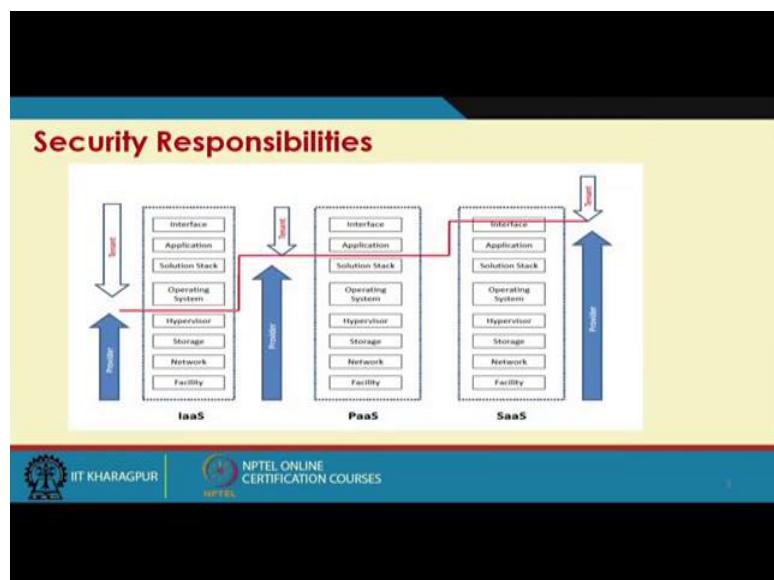
So, we have once I off load my data and application on or outsourced in a cloud, then we do not have much control over that things. Or our control is decided by the provider right. The whatever we I can control or whatever the handlers I am having for the control is primarily decided the by the service provider. So, these are this in other sense if we look at the security point of view this, this is a constraint of how my data and you need to be secured, what is the what is the how much it is exposed to the external other applications and other type of users and all those things.

There are other general concern like inadequate policies and practices, that is another concerned and insufficient security control as we are looking talking about. So, customers use cloud services to serve their clients. So, customers and can use a cloud services to serve their clients. So, running the applications on those cloud services needs to establish trust relationships right. So, it is there is requirement that how much I trust this service provider. So, there is a major requirement for that and there are this can be

beneficial for both the stakeholders the customer and provider. So, it is not only the provider how the customer trust the provider there is a question of how much whom I am if I am a service cloud service provider if the customers for me or the consumer of the service services, should be also trust out this.

So, there should not be malicious customer who will create a problem out of the things it is not always that it is not that thing because the cloud service provider their business is selling the services. So, they may not be malicious or they may not have any malicious instead any malicious intent. However, you can in the process you can have some malicious customers which are usually the scenario, who can use the services use the platform to attack or peep into others data and type of things.

(Refer Slide Time: 05:10)



So, if I this is a thing which is available in it is a various literature we have also seen. That if you look at the security responsibilities in case of IaaS the responsibility up to the hypervisor end after that it is having the operating system or the guest operating system so and so forth, it goes to the tenant right. So, the providers responsibility up to hypervisor in case of PaaS cloud provider responsibility is up to that platform or that were are the solutions stack is there.

In case of a SaaS it is responsibility goes up to the interface application interface right. So, it is the things like if I am using a say a API for what processing. So, it is up to that level the responsibility of providers coming to play, for the for the consumer it is it is

up to that up to that application level the services are there this security are handled by the provider.

So, we can see that at various type of clouds we have different type of level of security. So, in case of a SaaS there is lot of things which depend on the on the providers end. Like I am if I am using a say what processing service or any type of text editing service. So, as I am using that API somebody else also maybe using that API.

(Refer Slide Time: 06:46)

**SaaS Cloud-based Collaboration**

- APIs for sharing resources/information
  - Service consumer(customers): human users, applications, organizations/domains, etc.
  - Service provider: SaaS cloud vendor
- SaaS cloud-centric collaboration: valuable and essential
  - Data sharing
  - Problems handled: inter-disciplinary approach
- Common concerns:
  - Integrity of data, shared across multiple users, may be compromised
  - Choosing an “ideal” vendor

Hemay Ghosh, Resource-Limited-coopetition in Cloud through Risk Estimation and Access Control Mechanism Thesis, IIT Kharagpur, 2016

So, it is the same application level I can have different instances which are working for different type of things. So, SaaS cloud base collaboration. So, what broadly we try to mean that API for sharing resources, and information service consumer or customers human users applications organization domains. And anybody service provider are the cloud vendor SaaS cloud, SaaS clouds centric collaboration. So, they are there are some of the essential things like data sharing issues problems handled like interdisciplinary approaches human to be taken to handle different type of issues.

So, common concerned is integrity of the data shared across multiple user may be compromised things right. As there is a the data is being shared across or the basic platform is share across multiple users. So, there may be a compromises and there may be a chance of being compromised. And how do I choose a ideal vendor or a service provider is one of the major challenge if there a number of provider then how do I choose the provider.

So, as am I send this is work of my one of my PhD student is Dr. Nirnay Ghosh who worked on this area. And we will be taking some part of his work to describe and the more we will be taking the challenges we taken up in this particular problem. So, it will be good to look at those a type of things.

(Refer Slide Time: 08:18)

**SaaS Cloud-based Collaboration**

- Types of collaboration in multi-domain/cloud systems:
  - Tightly-coupled or federated
  - Loosely-coupled
- Challenges: securing loosely-coupled collaborations in cloud environment
  - Security mechanisms: mainly proposed for tightly-coupled systems
  - Restrictions in the existing authentication/authorization mechanisms in clouds

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, type of collaboration in multi domain or cloud systems is tightly coupled or federated can be one way of looking at it. Where I have strong connectivity between this type of federating clouds or they are loosely coupled systems. So, that they there are federally cloud, but they are loosely coupled system. So, I have instances in different cloud may be in the same cloud, but they are loosely coupled. So, they are not very strongly coupled.

So, there are various changes securing loosely coupled collaboration in cloud environment is a major problem, and security mechanisms mainly proposed for tightly coupled systems. So, what loosely coupled there are not much security mechanism. So, whenever you look for the security mechanism as we discuss earlier, it send to in phenomena. So, there is the requirement is goes hand in hand. So, in turn it comes to be more tightly coupled thing restriction in the existing authentication authorization mechanisms in cloud there is another problem that the type of each mechanisms you are having at in the present day cloud may be a restrictive to having those secretive phenomena in place.

(Refer Slide Time: 09:30)

## Motivations and Challenges

- SaaS cloud delivery model: maximum lack of control
- No active data streams/audit trails/outage report
  - **Security:** Major concern in the usage of cloud services
- Broad scope: *address security issues in SaaS clouds*
- Cloud marketplace: rapid growth due to recent advancements
- Availability of multiple service providers
  - Choosing SPs from SLA guarantees: not reliable
    - Inconsistency in service level guarantees
    - Non-standard clauses and technical specifications
- Focus: *selecting an "ideal" SaaS cloud provider and address the security issues*

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES



So, there are a lot of challenges and which motivates or if you look at in the other way these are the motivation for having research or study in this area, like SaaS cloud delivery model, So maximum lack of control right. So, whole control on the service provider end; so these has the minimal that the control on the consumer end. No active data stream audit trails outage reports are directly available to the things, whatever is provided by the consumer need to be looked into.

So, major concern in uses of the cloud services; so broad scope address security issues in the cloud we need to address. So, there is a concept of cloud marketplace coming up like that rapidly growing due to recent advancements. So, we have a typically a cloud market place where numbers of providers number of consumers there is a economic model is goes on I am not talking about a cloud economics talking about that where you go for better services not only pricing quality of services better a SLAs is better security and things are there.

So, availability of multiple service provider is a major challenge of choosing that which service provider we need to look at like. So, there is inconsistent in service and a guarantees no standard clauses. So, there is a selecting an ideal SaaS cloud provider and is a issue, and how to if I after selection what are the different other security challenges can come up.

(Refer Slide Time: 11:10)

**Motivations and Challenges**

- Online collaboration: popular
- Security issue: unauthorized disclosure of sensitive information
  - Focus: *selecting an ideal SaaS cloud provider and secure the collaboration service offered by it*
- Relevance in today's context: *loosely-coupled collaboration*
  - Dynamic data/information sharing
- Final goal (problem statement): *selecting an ideal SaaS cloud provider and securing the loosely-coupled collaboration in its environment*

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

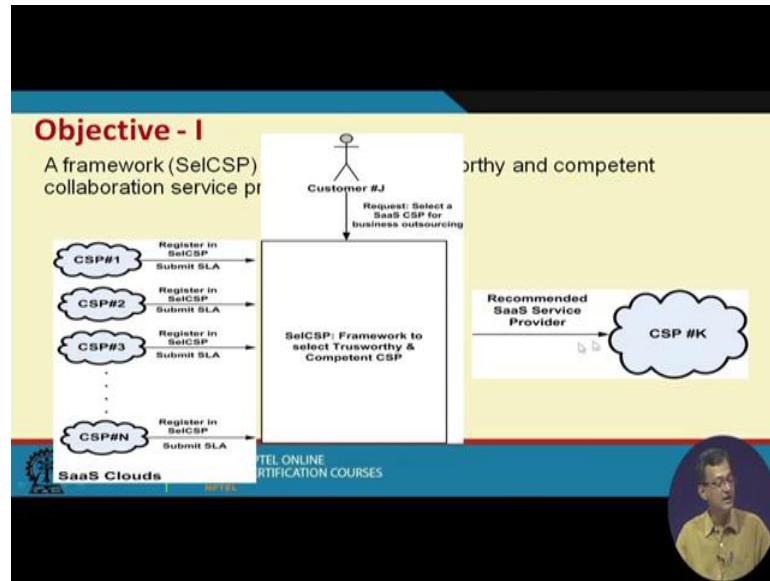


So, there are other things like online collaborations are becoming pretty popular right. There are several security issues I finding a ideal provider. Relevance of today's context there is a loosely coupled collaboration dynamic data information sharing like if you look at any e marketplace or look at any type of service provider like what we see that any type of things where you purchase and over online purchase selection etcetera, even your travel booking centers.

So, there are different parties which are being connected and mostly they are loosely coupled there are parties who are provider of the products, there are parties who are provider of the financials area like credit card debit card to other types of services, there are parties who are courier services and type of things then they are being connected over in a loosely couple way.

So, our goal is to select an ideal SaaS cloud provider and securing loosely coupled collaboration in it is environments. So, what are the different aspects. So, what are what they looking for a typical approach for that it is not like that that there are there are cannot be other approaches, but what way we can go into this particular problem.

(Refer Slide Time: 12:34)



So, if you look at our one of the objective is to whether we can developed a framework like as I mentioned that in this particular work we developed a framework or sel a SelCSP selecting a trustworthy and competent collaboration service provider right.

So, there can be different CSP's in set of CSP's and registered in that particular some sort of a central authority. And the customer requesting to select a SaaS provider for business outsourcing, and it recommends that CSP particular k, or CSPi is the base suited for it is requirement looking primarily at the security aspects right. So, that is the goal of the thing.

(Refer Slide Time: 13:23)

## Objective - II

Select requests (for accessing local resources) from anonymous users, such that both access risk and security uncertainty due to information sharing are kept low.

The diagram illustrates a cloud computing environment where four domains, labeled D1 through D4, are represented as boxes within a cloud icon. Below the cloud is the label 'CSP #K'. Three arrows, each labeled 'Collaboration Request', connect the domains: one from D1 to D2, another from D1 to D3, and a third from D2 to D3. This visualizes the process of selecting requests for local resources from anonymous users within a single cloud provider's infrastructure.

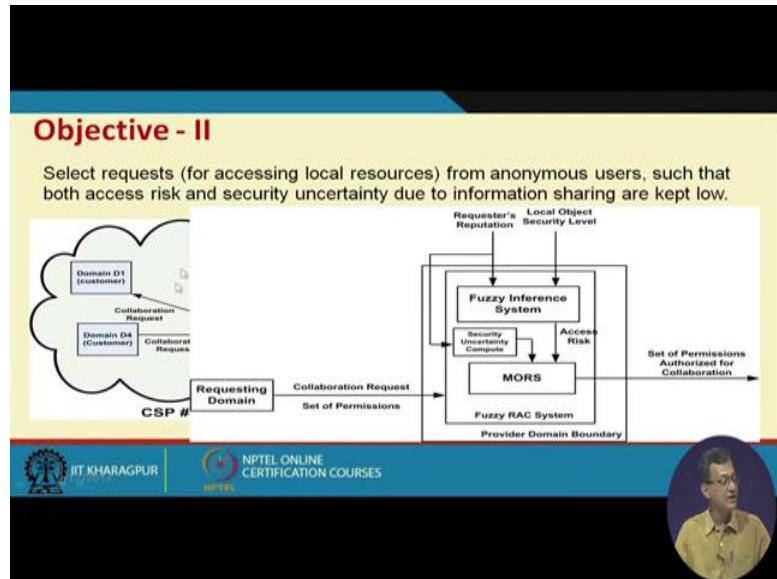
IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

A circular portrait of a man with glasses and a yellow shirt, identified as the speaker or professor.

There can be there after the selection there can be select the request for accessing the local resource. So, once I once I select the particular CSPs, then we want to look at the select request for accessing local resources within the cloud, for anonymous user because we do not know who are the users, such that both access risk and the security uncertainty due to the information sharing are kept low.

So, our objective is to that access risk and security uncertain c for information sharing should be kept low or minimum. So, if I have that different customer in different domain, like a domain one domain 2 domain 3.

(Refer Slide Time: 14:18)



And they are collaborating in some somewhere other, say we need to have some sort of a mechanisms here we worked on a fuzzy inference system, which keeps a that. So, requesting domain collaborate request and set up permissions right. And say request reputation request are reputation local object security level like, and set up permission authorized for the collaboration. So, it is it is may So happen that I request for set of operations, and then I based on the basic policy engine and based on requester any reputation and local objects security level, I grant a set up permission authorized for the collaboration right.

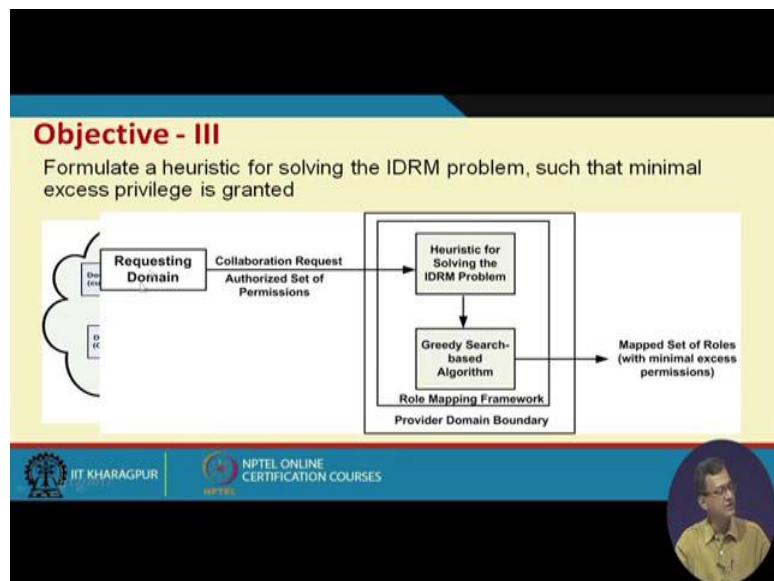
So, it is if some sort of a analogy we try to do, like I want to access some a particular office or type of things. And then based on the based on my reputation or credential I maybe given access to different type of things like I can be said that you can enter the campus you can enter the visitor lounge, but you cannot enter the actual office with based on my reputation another type of credential I may be allowed to inter the actual office, but; however, I cannot enter the say computing system lab or the where actual labs are there.

So, it based on your level of authority and your requirement and requesting domain you go on things like I go to a bank. And if I am just going to deposits some check or some documents then I go somewhere, if I want to look at meet the manager I go to some other level of accessibility and type of things and it depends that my requirement type of

things. Or in other sense if this miss access one misses or access role are decided by the by the my requesting role, and what is the permissible things for different type of objects like if my accessing a particular section of the thing is a if they if will take a object then based on needs access policy I be I need to be filter.

So, in case of a collaborative cloud, then when the customer comes with a type of request based on it is reputation another type of access policies on the objects. It has been granted a set of things, it not likely that whatever the particular customer has requested everything has been permitted, but a subset of that as can be permitted based on it is reputation.

(Refer Slide Time: 16:51)



So, other objective can be formulate a heuristic to look at that IDRM problem inter domain role mapping problem, such that minimal access privilege is granted.

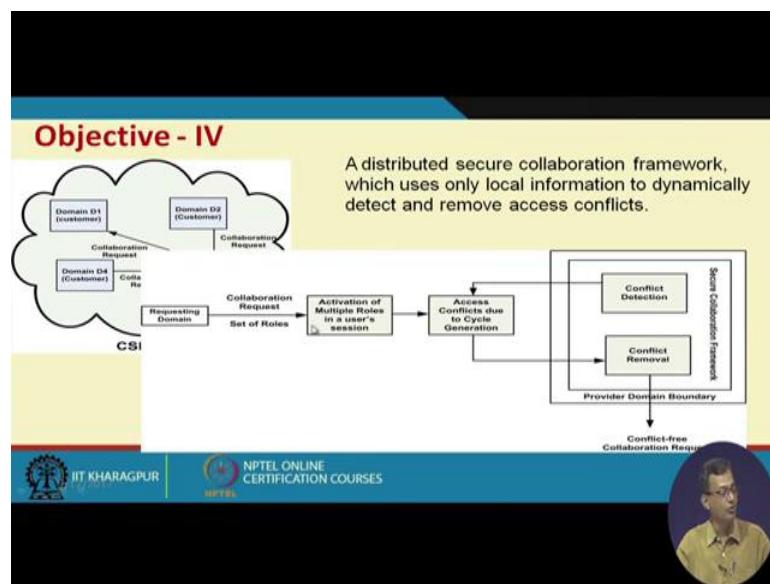
So, that is that is that is a problem for different type of things like, if I say from organizing a I want to access something at organization b, then the role of organization the role I am having in organization a need to be map to a equivalent role in the other organizations right. Like I am accessing as a financial organization from organization say 1, 2 another some as some data in the organization 2, and here I am accessing as a say a manager of level 1. And that data in order to which is equivalent to manager of level 2 they are. So, then my role need to be map to that particular level otherwise I cannot

access the thing. So, this is a this is a roll mapping problem is a is a problem which is there already and when need to look at it is there in this type of collaborating cloud also.

So, here also we try to say that requesting domain collaborating request authorized a set of permissions, and based on some heuristic for solving the IDRM problem we will see that this is a hard problem. And so, we need to have a some greedy research based algorithm and try to have a mapped a set up roles with a minimal excess permissions. So, minimal excess permission which it tries to say that I need to given that level of permission, which that minimal set of permission which I need to which is the which is required to execute the things.

Suppose I want to read a document. So, I can be given only read permission I can be given read and right permission. So, the minimal set maybe the reading the thing right. So, so that it is no excess permission are no excess permissions are given to the thing. And another objective may be a distributed secure collaborative framework which uses only local information to dynamically detect and remove excess conflicts there is another major challenge in any type of loosely coupled these systems.

(Refer Slide Time: 18:52)



So, how I can have a dynamic framework with a only local informations? Now I want to excess organization one from or a organization one running into cloud instance in a cloud say in cloud providers cs CSP 1, want to access another data in CSP 2 of another

organization, then I may not have all the information of this either the CSP or the excess write of the things.

So, I need to look at my local resources or local information and try to have the maximum security. In other sense when you have a loosely coupled things you may not carry all the credentials whatever it is having into collaborative. So, you should be there should be a mechanisms or there should be a way or there should be an approached that how I map it into the into my way of looking at it.

So, here also we tried a requesting domain collaborating request with a set of roles, activation of multiple roles in the users sessions right. And access conflicts due to the cyclic cycle generation there can if there is a cycle generation there can be a access conflict; that means, some document one way I am not able to access due to my particular role in a particular organization, but if it is goes to a cycle like an reach to this type of things.

(Refer Slide Time: 20:39)



So, there can be access conflicts. So, I need to have a conflict detection and conflict remover and then I should have a conflict free collaboration request right. So, this sort of mechanisms we need to look into.

(Refer Slide Time: 20:46)

## Trust Models in Cloud

- Challenges
  - Most of the reported works have not presented mathematical formulation or validation of their trust and risk models
  - Web service selection [Liu'04][Garg'13] based on QoS and trust are available
    - Select resources (e.g. services, products, etc.) by modeling their performance

• Objective: Model trust/reputation/competence of service provider

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES



So, one is the selection of trustworthy and competent SaaS cloud provider for collaboration. So, there are challenges of most of the reported works have not presented. So, there are several challenges objective is the model trust model trust reputation competence of the service provider. So, there are we are looking at 3 component, trust, reputation, competence. So, they are very much interlinked, but again there are they have some distinct property.

(Refer Slide Time: 21:19)

## Service Level Agreement (SLA) for Clouds

- Challenges:
  - Majority of the cloud providers guarantee “availability” of services
  - Consumers not only demand availability guarantee but also other performance related assurances which are equally business critical
  - Present day cloud SLAs contain non-standard clauses regarding assurances and compensations following a violation[[Habib'11](#)]

• Objective: Establish a standard set of parameters for cloud SLAs, since it reduces the perception of risk in outsourced services

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

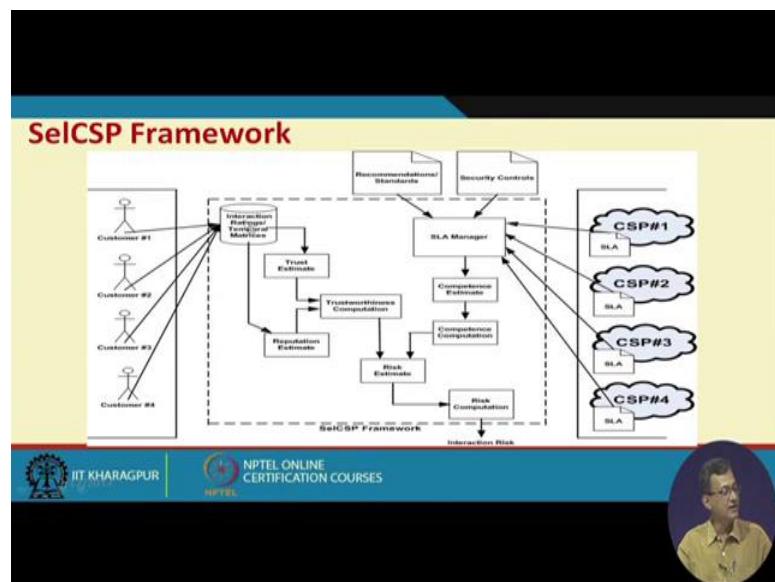


So, how much it is trusted? Whether it is competent to do that? And how what is its reputation right of doing a particular things or security type of things?

So, there are again challenges if you look at the SLA's because someone may argue that the SLA's tries to cover this, like majority of the cloud providers guaranteed availability of services right. Consumer not only demand availability of guarantee, but also other performance related assurance which are equally business critical. So, I am not only looking at the availability as a consumer, but also that assurances that this will be done this type of in timeframe or in a up to the compensation.

Present day clouds SLA's contain nonstandard clauses regarding assurance and compensation following a violation. So, there are compensation of the penalty scheme they follow some nonstandard present, nonstandard in the since there is no standardize mechanism across the cloud thing. So, one again establish a standard set of parameters for cloud SLA's since it reduces the perception risk of the outsourced services. So, there should be way to reduce the perception resource.

(Refer Slide Time: 22:21)

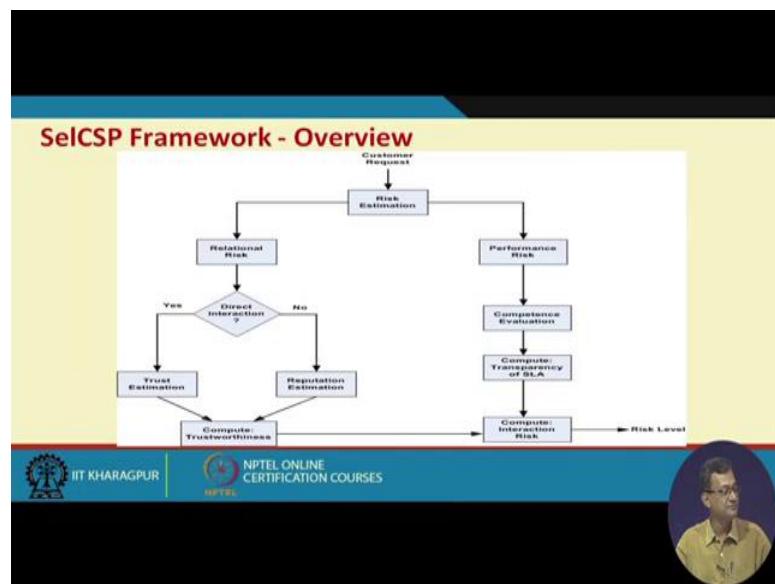


So, this is again a we try to look at a framework whether there are different customer. So, interaction rating and temporal matrices are complete computed. So, trust estimation reputation estimation trust worthy the worthiness of the computation then risk estimation and risk computation. On the other hand what we have recommendation and standard

security controls which drives the SLA's managers and this all this with the service provider there are a SLA's, right.

They provides some level of SLA's then competence estimation competence computation one side that we calculate trust worthiness another side competence, risk estimation, risk computation of the particular things and interests in risk. Out of that we try to find out interaction these between for different service provider.

(Refer Slide Time: 23:10)



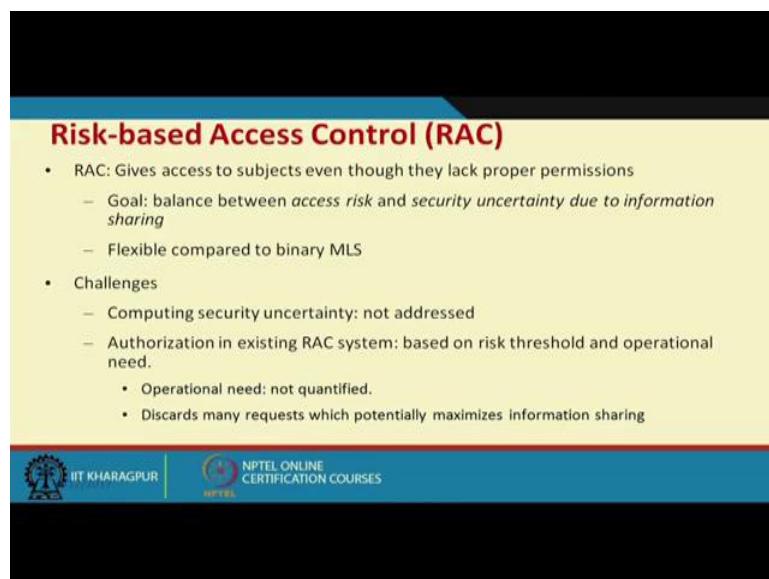
So, there are different flow of these is SelCSP framework, that is a one is risk estimation and can be relational risk direct interaction trust estimation same thing We want to put in it in the flow chat.

(Refer Slide Time: 23:30)



The second is recommending access request from anonymous users for authorization.

(Refer Slide Time: 23:36)



So, one is that risk based access control right. So, though we have heard about other type of access control, I role based access control this we term as a risk base access control.

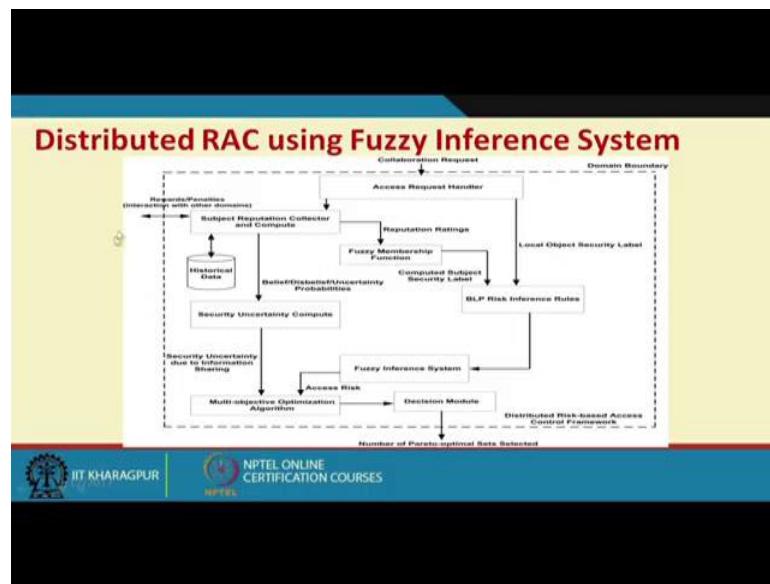
So, gives access to a subjects, even though they lack proper permissions right. So, I have I do not have to have the whole set of permissions which are fully coupled. So, can whether I can give the access with some amount of risk involving it right. It is not that is

binary stop on or off, but those sort of things. Goal balance between the accesses risk an security uncertainty due to information sharing right.

Flexible compared to the binary MLS. So, that is little bit of as we are talking about that instead of binary I take care of little bit of risk right. So, challenges computing security uncertainty is not a fully addressed stuff right. So, how to I look at computing security uncertainty right? Authorization in existing risk base access control system based on risk threshold sold and operational needs, right. Operational need not is quantified. It is difficult to quantified, operational risk did discards many request which potentially maximize information sharing right.

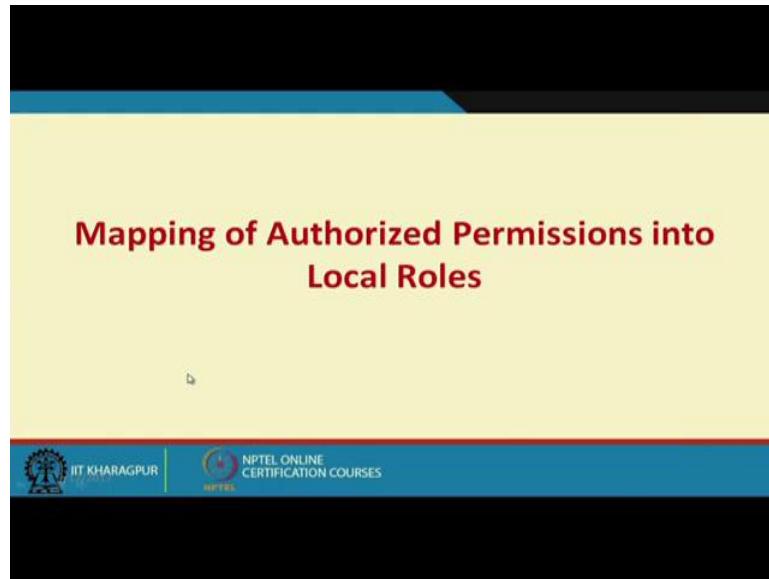
So, in order to reduce it we discards many request which prove purse potentially maximize the information sharing so that my overall risk come down. So, in other sense in order to reduce the risk we try to reduce the collaboration itself right.

(Refer Slide Time: 25:29)



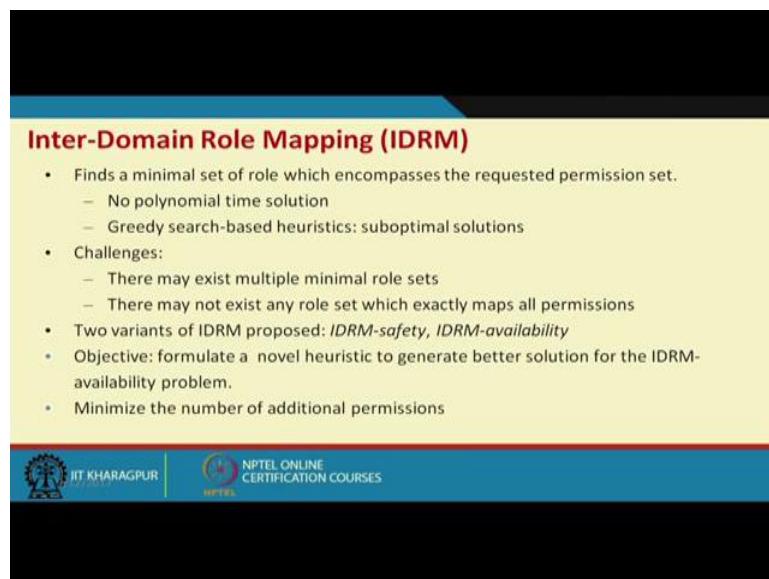
So, that it is one of the looking at it. So, there is a distributed frame work as we used a file fuzzy inference system to look at it. And it tries to find out a distributed racking to the thing.

(Refer Slide Time: 25:40)



The next one is mapping authorize permission into local roles right. So, inter domain roll mapping thing IDRM.

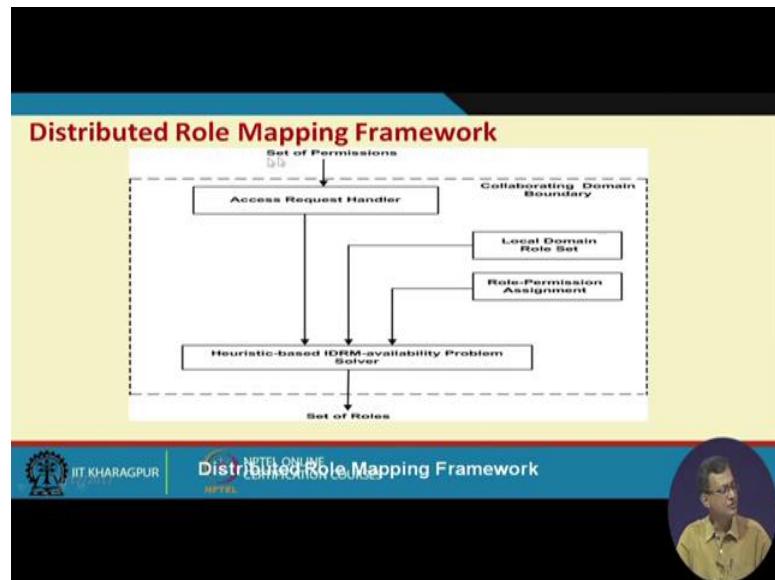
(Refer Slide Time: 25:45)



So, we what you have the finds a minimal set of roles which in compasses the requested permission set. No polynomial time solutions are available greedy search based heuristics sub optimal solutions. Challenges they are may exist multiple minimal role set right. So, that there can be existing minimum multiple minimal role set they are may not exist here any roles set which exactly map to the all permissions.

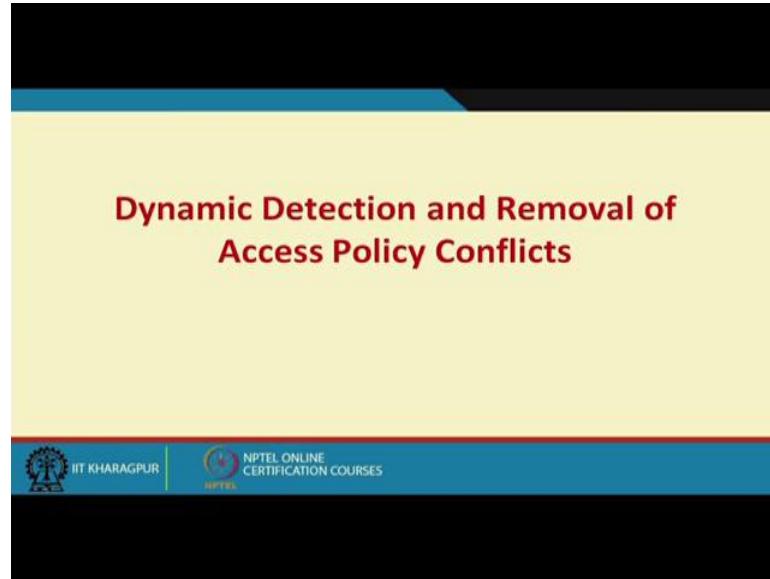
So, there are different types of problems or challenges. So, 2 variant of a IDRMs are there one is IDRMs safety and IDRMs availability; so IDRMs availability and safety. Objective to formulate a novel heuristic generate better solution to IDRMs availability is problem minimize the number of ability permission.

(Refer Slide Time: 26:37)



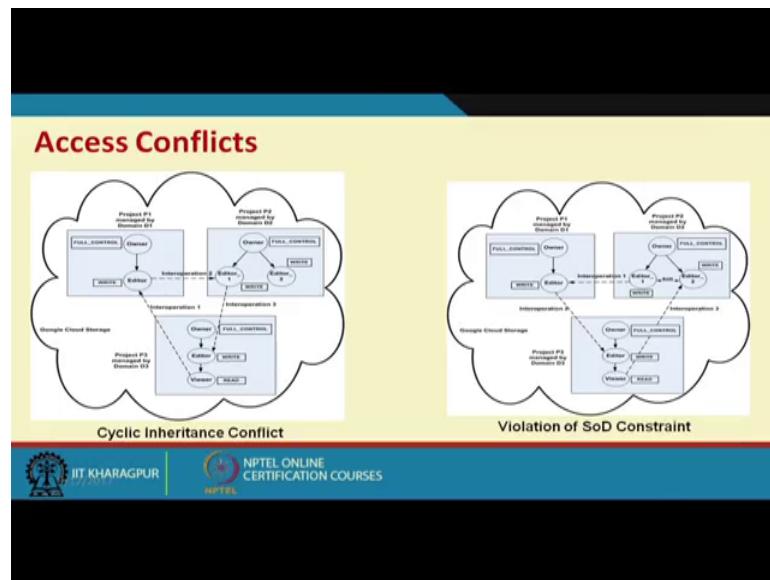
So, here also if you look at the distributed role mapping framework; so I have a set of permission access request handler. And we have local domains role set roll permission align assignment that role to permission set. So, which are set of permissions and heuristic, based idea availability problem solver what we try to formulate or propose and which keeps a set of role which is a minimal set up role.

(Refer Slide Time: 27:04)



And finally, there is the other aspect of dynamic detection and removal of access conflict. So, this is another major problem whenever we have multiple collaborations. So, there may be a chance that you may there may be a cyclic access cycle. And it may lead to accessing some objects which are other way a particular subject is not suppose to access right.

(Refer Slide Time: 27:36)



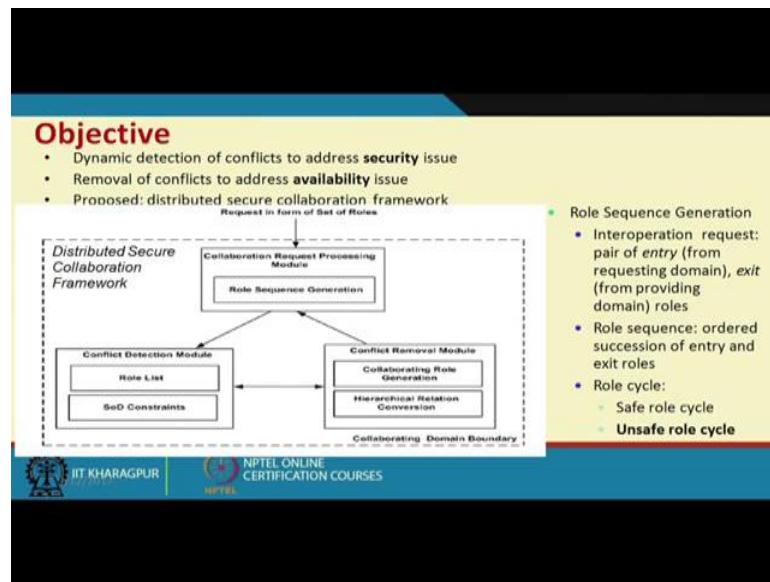
Like If you look at this cyclic inheritance conflict like this viewer this particular things is not allow to write or as editor permission, but I can have access to another domain which

has a right permission that is allowed from they are to another editing which has a reading permission.

So, in other sense I cannot right. To this particular subject cannot write to this particular object, but I can have a cyclic way of this. So, inheritance cyclic way and write to this. So, in other sense I have done a conflicting situations which other way I am not supposed to do. There can be violation of sod constant. So, sod constant is the separation of duty really, like I can say a typical analogy like say in a bank the person who is issuing a demand draft cannot verify the demand draft like I am I am in the issue counter. So, if I am issuing the thing, issuing the draft then the verification I cannot do the same thing right. So, there should be has to be a separate things. So, it is a separation of duties has to be there sod what we to popularly known as a sod constant, which is they are in any security information security mechanisms.

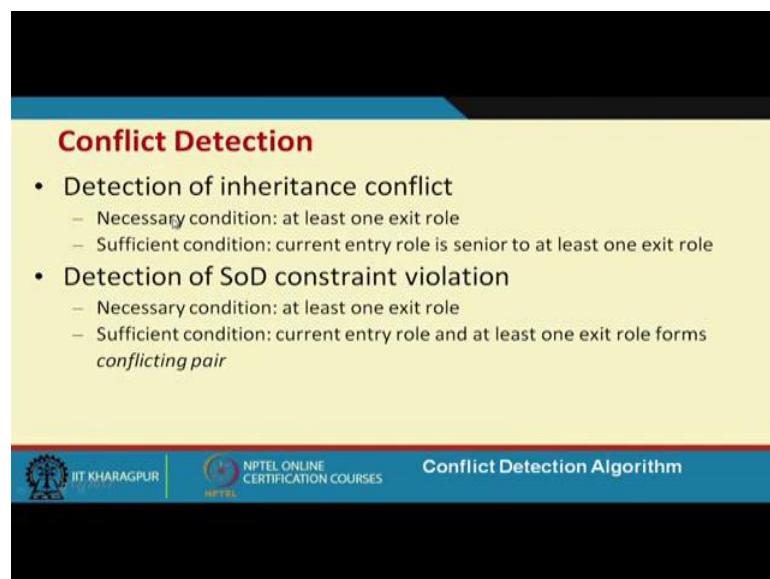
So, here also we can see there may be a conflict in the sod constant itself. Like here I can have a right. On the things and I can have another channel with having this editor writing on these editor there is a communication here. So, I can basically though there is a sod constant, I communicate between through a different channel right. So, there is there can be a violation of things, these things happens whenever there is a multiple communicating partner and specially when they are loosely coupled; that means, you do not know the whole security scenario of the other things or security settings of the other party.

(Refer Slide Time: 29:53)



So, here also we try to for a particular distributed security collaborative frame work, which takes a set of roles and based on collaborating request processing modules, and conflict detection and conflict remover module come up with a set of a scenario which will be complete. So, role sequence interoperation request pair of a entry from the domain exist from a providing roles right role sequence order success of entry and exist role. So, I can have a safe role cycle unsafe role cycle.

(Refer Slide Time: 30:38)



So, it as we understand it has too think one is that role a detection, that is there if there is a conflict they are need to detection it has to part detection of the inheritance conflict detection of the sod constant violation. So, this to need to be detected what and other is the now we need to remove the things.

(Refer Slide Time: 30:50)

**Conflict Removal**

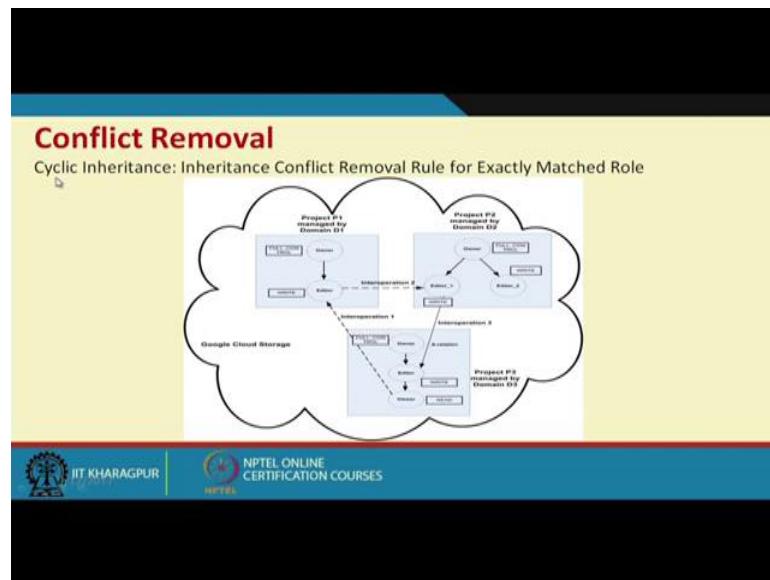
Cyclic Inheritance

- Two cases arise:
  - Exactly matched role set exists
    - RBAC hybrid hierarchy
      - *I-hierarchy, A-hierarchy, IA-hierarchy*
    - Replacing *IA-relation* with *A-relation* between exit role in previous domain and entry role in current domain
  - No-exactly matched role set exists
    - Introduce a virtual role

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES  
NPTEL

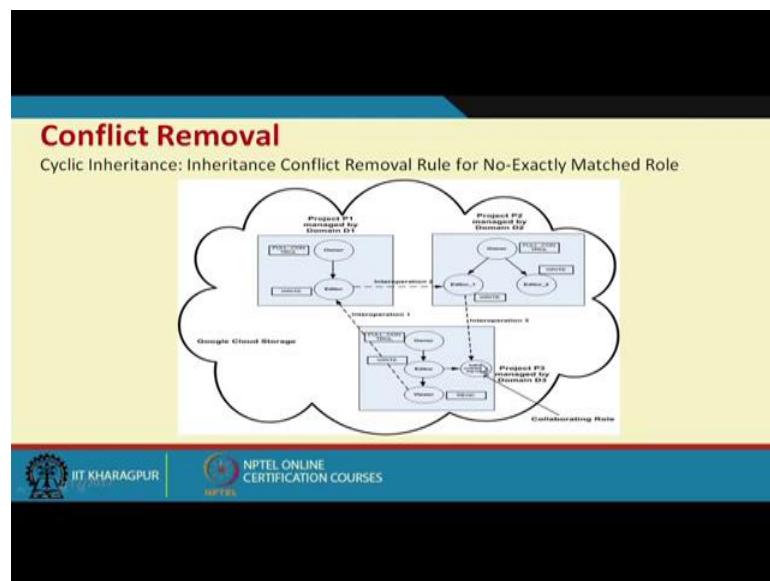
Then once detected that those conflicts need to be removed; so one is that 2 cases may arise exactly match role set exist. So, r back hybrid hierarchy or there can be no exactly role set exists right I can. So, I can have a virtual role into the things. So, I can create a dummy role and look at it.

(Refer Slide Time: 31:20)



Like if you look at cyclic redundancy inheritance the cyclic inheritance conflict removal role for exactly match roles what we do here.

(Refer Slide Time: 31:40)



So, instead of this after the conflict removal we create a collaborating role. See here it was basically it was a cycle to end up in this editor whereas, here inheritance conflict removal for no exactly match role. So, as there is no exactly match role of looking at it. So, we create a another sub roll or a new roll which is fall back.

Now, this viewer there is no way of going to this particular editor. So, that I am not going a there is no conflict.

(Refer Slide Time: 32:12)

## Conflict Removal

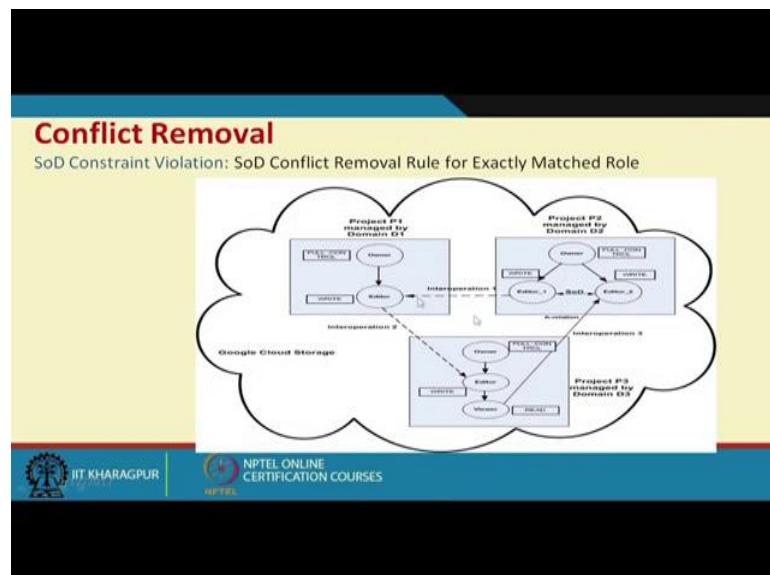
SoD Constraint Violation

- Two cases: similar to removal of inheritance conflict
  - Additional constraint: identifying *conflicting permission* between collaborating role and entry role in current domain
  - Conflicting permission
    - Objects are similar
    - Hierarchical relation exists between access modes
- Remove conflicting permission from permission set of collaborating role

JIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

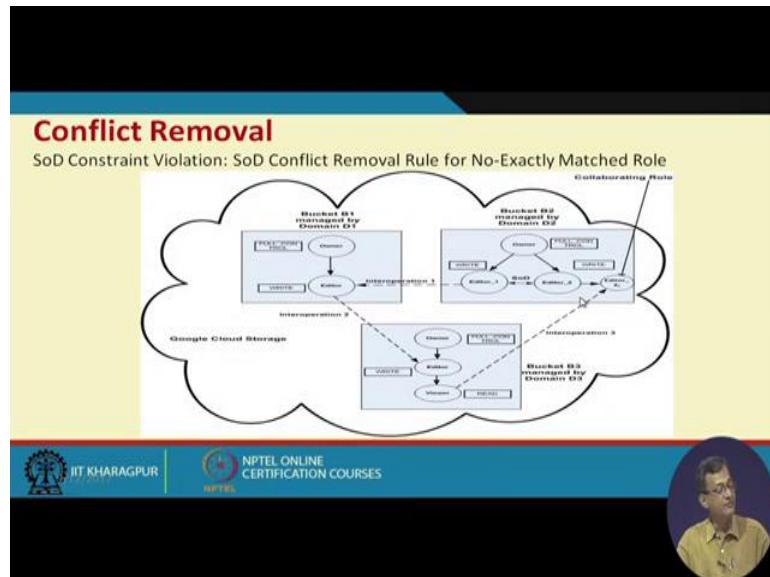
Similarly, for conflict removal also, we can have a to sod constraint removal thing here also we have. So, this was our earlier scenario where this editor rights here and this viewer this editor can right has access to these and to and finally, it goes to this editor one and took though there as sod. So, there is a chance of a cycle which violates this sod.

(Refer Slide Time: 32:21)



So, here also if there is no exact match we created a collaborating role right. So, it collaborating role of the editor 2.

(Refer Slide Time: 32:42)



So, editor 2 c and it is ends up there. So, there is no sod violence in between the editor one and editor 2, right. So that means what we are trying to look at that in doing so.

(Refer Slide Time: 33:02)



We may in the or may basically formulate a scenario where this where there is healthy collaborations between the things without security violences. So, this is a typical approach we try to show that how secure collaborated in the sass cloud can be possible,

and definitely this is a very brief overview, but that is good enough or it will help you those who are interested in this sort of research. See there is in need of infrastructure is minimal, but; however, you can basically work on a this sort of problem right.

So, one is that selection of the trustworthy and competent cloud provider and after the selection we have the recommending access request from the anonymous users for authorization. Then mapping of authorized permission to the local roles, and detect dynamic detection removal the access control conflicts. Like there can be cyclic inheritance problem conflict or sod type of conflict can be there which can be it. So, what we try today what we discussed today is that looking at one of the one of the very tricky issue of security where collaborating SaaS cloud in a in a loosely coupled way.

So, what are the major what are the typical security issues can come up, and how to approach those problem to address is there can be different other approaches. You can find in the literature and even you can think of other approaches, but this is a typical way of looking at it. And this is a problem which is very much part intent, and which has very much true for today's cloud scenario collaborating cloud scenario.

Thank you.

**Cloud Computing**  
**Prof. Soumya Kanti Ghosh**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**

**Lecture – 30**  
**Broker for Cloud Marketplace**

Hello. Today we will discuss a aspect of Cloud Computing, where we will see that if there are number of providers. So, how do I select whether there is a way to select the things and how what should be the approach of the things. So, what we see in a present day scenario that there are increasingly availability of number of provider, at various level IaaS level, PaaS level SaaS level. So, at various level and there are several requests on the from the customer side right.

So, customer has a expectation or requirement per se, and the providers provides a particular things with some SLAs right. So, first of all the customer wants to know that where which provider, how to select a provider out of a branch of providers available, where the customers expectation will be satisfied and the SLA will be honored and the providers also wants to maximize his profit right.

Some of the aspects already we have been seen. So, given all these security and other aspects in place we do not to look at that whether in this cloud quote unquote marketplace, how do I whether the what whether there is a possibility or approach to look at the things. So, that is exactly we like to look at. So, it is broker for cloud marketplace. So, what we look at there is a number of providers available there are several customers. So, it is a marketplace to how to select the thing.

(Refer Slide Time: 02:01)

## INTRODUCTION

- Rapid growth of available cloud services
- Huge number of providers with varying QoS
- Different types of customer use cases – each with different requirements

**Machines in the sky**  
Amazon's virtual computers, "000 created per day"

Source: Douglass \*3-month moving average

IIT KHARAGPUR     
 NPTEL ONLINE CERTIFICATION COURSES  
NPTEL

So, if you look if we see that there is a rapid growth on available cloud services right. So, it is there are several providers, several service providers right; and there is a rapid growth of the things like and huge number of providers with varying quality of services are things are there right.

So, different providers has different quality of services and type of things different type of customer use cases. So, customer based use cases each with different requirements or with several requirements customer use cases are there. So, thus the review the requirement of the customer is based on the need of the customer, it varies from customer to customer. So, that is another aspects and from a from some reference what we see that this number of amazons VMs created per day from 2007 to somewhere 2011 it has it has remarkably changed right. So, there is a huge demand for the things.

(Refer Slide Time: 03:08)

**INTRODUCTION**

- Rapid growth of available cloud services
- Huge number of providers with varying QoS
- Different types of customer use cases – each with different requirements

- *Need for a "middle man" (Intelligent Broker!) to*
  - Suggest the best cloud provider to the customer
  - Safeguard the interests of the customer

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, keeping all those things in the offering or in keeping in this mind there is a need of a middleman right. So, we there is a need of a middleman we can called as a agent or broker or sometimes it may be a intelligent broker to suggest the best cloud provider to the customer or safeguard the interest of the customer right.

So, I require a quote unquote middleman or intelligent broker, serve broking service which safeguard the definitely safeguard the interest of the customer, and also suggest the best cloud provider or cloud service provider of the customer based on its requirements right. So, this is need this is a need of our or people are working on it that how can I select how can I broke, how can I select a things with a intelligent agent or broker.

(Refer Slide Time: 04:13)

The slide has a dark blue header bar at the top. Below it is a light blue bar containing the word "MOTIVATION" in white, bold, uppercase letters. The main content area is yellow and contains a bulleted list of four items: "Flexible selection of cloud provider", "Trustworthiness of provider", "Monitoring of services", and "Avoiding vendor lock-in". At the bottom of the slide is a dark blue footer bar. On the left side of this bar are two logos: the IIT Kharagpur logo and the NPTEL logo, which includes the text "NPTEL ONLINE CERTIFICATION COURSES". On the right side of the footer bar is a circular portrait of a man with glasses and a yellow shirt.

- Flexible selection of cloud provider
- Trustworthiness of provider
- Monitoring of services
- Avoiding vendor lock-in

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, there are definitely motivation that these flexible selection of cloud provider, there is one of the things trustworthiness of the provider plays a important role how much I trust or I can be able to trust a provider is there, how do I calculate the trust is a big challenge yes some of the aspects we have looked into or discussed during the security, when we discussed about cloud security. So, starts worthiness is important right.

So, that is one of the factor monitoring the service that whatever the things are there, how services are monitored and of course, there is important aspects of vendor lock in right. So, whenever there is a vendor. So, there is a problem of vendor lock in so that means, you work with some provider and you get locked with the provider because of the services right now if the providers is not giving the service up to the mark, or the provider or the CSP does not provide the service then going to another vendor is difficult; and it is not only that now you need to go for a in the terms and condition of the or the customer has to go for the terms and condition of the provider.

So, vendor lock in is also important aspects which we need to look into the thing. So, there are several motivations there may be several other motivation, but end of the day what we look at that I may need I need a the mechanisms to or a broker to find in find the best possible cloud things.

(Refer Slide Time: 05:54)

## OBJECTIVES

- Selection of the most suitable provider satisfying customer's QoS requirements
- Calculation of the degree of SLA satisfaction and trustworthiness of a provider
- Decision making system for dynamic service migration based on experienced QoS

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES



So, based on this motivation if you look at the objectives, so one is the selection of most suitable provider satisfying customers QoS requirements that is the major objective; calculation of the dis degree of SLA satisfaction trustworthiness of the provider may be one of the sub objective right like how I can guarantee that the degree of SLA satisfaction and trustworthiness of the provider, that is the other objective which one is selecting and type of things decision making system for dynamic service migration based on the experienced QoS is another aspects right.

So, it is if the provider if I do not get that appropriate service or that is from the customer point of view or from the aspects of the that a middle ware, that whether I can migrate the service from one provider to another provider right like we look at that VM migration if the my if there is a any outage on the VM whether the application will be migrated on the things.

This type of things has been supported by various organizations as we look at on the hardware and type of things like at the VM level, but whether this type of migration on the whole application from the one provider to another provider that is may be there. So, once we do a migration or any of this or many of these aspects, we need to take a call all right.

So, there should be a decision making process into the things right that is one aspect. So, I need to have a some sort of a decision making support, which allows me to do that.

Secondly, another aspect is there that most of the cases things are not very crisply defined right the requirement wise the your performance wise, the thing is the thing is the all these parameters not very crisply reach us there is lot of fuzziness; specially in fine in giving the details of the customer. So, what we need to do we need to account for the overall those aspects also.

(Refer Slide Time: 08:13)

The slide has a blue header bar with the title 'Different Approaches'. Below the title is a yellow rectangular area containing a bulleted list. At the bottom of the slide is a footer bar with the IIT Kharagpur logo and the text 'NPTEL ONLINE CERTIFICATION COURSES'.

## Different Approaches

- CloudCmp: a tool that compares cloud providers in order to measure the QoS they offer and helps users to select a cloud.
- Fuzzy provider selection mechanism.
- Framework with a measure of satisfaction with a provider for keeping in mind the fuzzy nature of the user requirements.
- Provider selection framework which takes into account the trustworthiness and competence of a provider.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, there are different approaches what its tried few of them are here, but there are different approaches people have tried like one is that cloud Cmp, a tool that compares cloud providers in order to measure the QoS the offer and helps user to select a cloud. So, there is a cloud there is a tool which compares the cloud providers in order to measure the quality of service they provide. So, based on that the a user or customer can select a cloud, there can be fuzzy selection mechanism right.

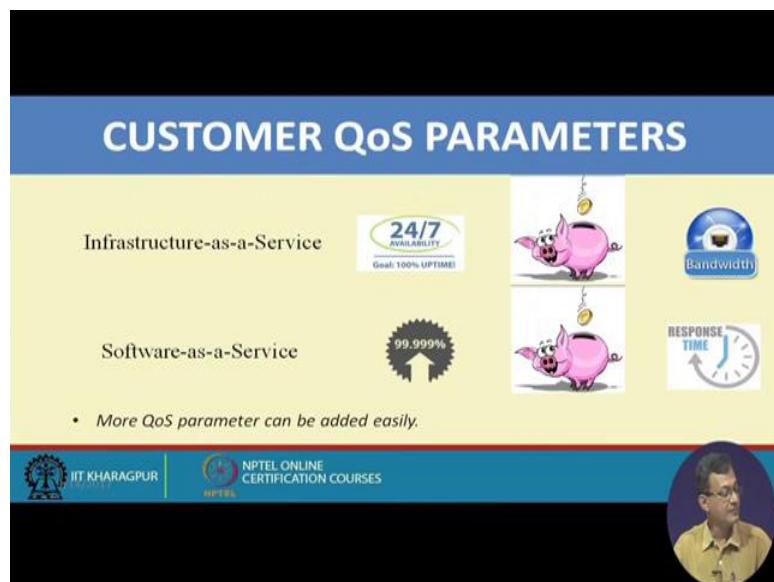
So, there is there can be fuzzy provider selection mechanism. So, that is fuzzy provide a selection mechanism. So, that if which takes care of the fuzziness that is also well suited because there are a lot of fuzziness into the description. There is a framework with a measure of satisfaction with a provider satisfaction with the provider for keeping in mind the fuzzy nature of the user requirements right.

So, keeping considering the fuzzy nature of the user requirement, measure of satisfaction with a provider how to calculate that based on this measure the consumer can take a call; provider selection framework which takes into account the trustworthiness and

competence of the provider. So, there is other way of another way of looking at it is that provide there is a folk provider selection framework which takes into account, the trustworthiness and competence of the provider all right this trustworthiness competence of the provider.

So, if you look at these different approaches. So, this one is there that it attempts to find a suitable provider for the things. Secondly, there are several overlaps between those type of things right may overlap in the sense that how they consider the user or the customer requirements or inputs the overall mechanism of selection and so and so forth, right.

(Refer Slide Time: 10:22)



So, if you look at customer QoS parameters some of the things, like if we consider infrastructure as a service. So, what we look at that 24 cross 7 availability with hundred percent requirement like it is the I get a returned what I pay for it, and there are requirements in terms of the bandwidth requirements other type of things. In case of a software wise service the uptime requirement may be male vary again the response times is critical. Rather if you look at this these are the several type of a components which are there. So, more QoS parameters can be there are like in terms of things like even I can talk about we can talk about storages, we can talk about other quality of services like even security, trustworthiness, competence risk and so and so, forth like availability and so and so forth.

(Refer Slide Time: 11:27)

**PROVIDER**

- Promised QoS values :  $Prom_i^1, Prom_i^2, \dots, Prom_i^L$
- Trust values :  $TRUST_i^1, TRUST_i^2, \dots, TRUST_i^L$

*Note: They have been kept independent as they pertain to different parameters*

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

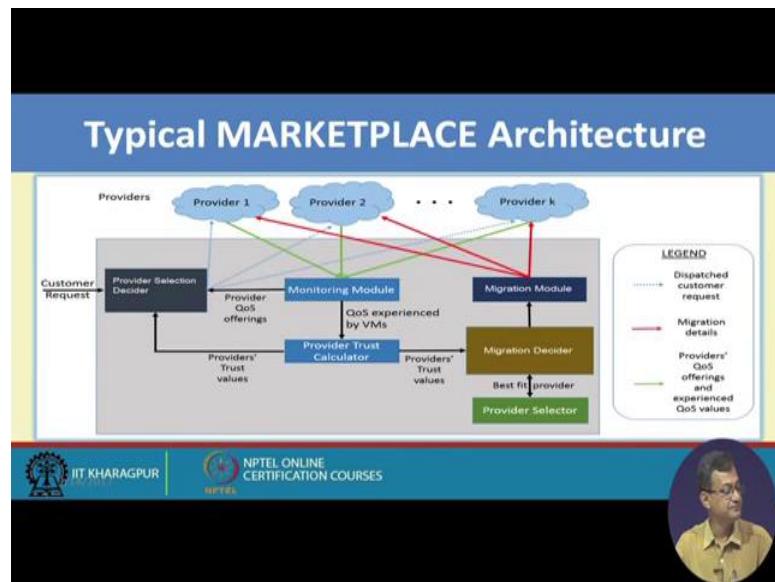


So, that is in case of any cloud provider, we have a set of parameters which the provider provide or the promised QoS values right they based on thing that they provider promise different type of provider, different type of QoS values or QoS parameters like I the provider says there is a 99.99 percent uptime or so much security level or maybe this much bandwidth available and so on and so forth and there are based on that we can have different trust.

On those things the trust can be overall over all the provider or I can even trust on a on different parameters like I like availability things. I say that that this based on the experience this trust on these things trust on a particular provider is something right 0-29 scale I trust it in 0.029 scale I trust it in point 8.8 value and trust value of 0.8 and so forth right.

So, there are different trust values and they have been kept independent as they pertain to different parameters right, there the trust value can be kept independent as there can be different type of parameters for looking at the as those trust values like promise a particular promise 1 can have a trust value 1, promise 2 and trust value 2 and that's it. There are different mechanism for trust calculation several things for trust computation type of things it goes for a for history of those things, and looking at the history and then calculate and or third party auditor or third party monitoring units, which monitors those trusts and there are mechanisms to calculate the trust.

(Refer Slide Time: 13:29)



So, if we look at a typical marketplace architecture. So, there can be several components right; one is there are set of providers right like a provider 1, 2, 3 p 1, p 2, p 3 pk there are components like the major black box may be the provider selection decider right. So, the customer comes with a set of requirement and requests the this particular selection box and it takes the other inputs and decide the things.

The other blocks are like maybe there can be monitoring module which are constantly monitoring, whether there is a request or not it is monitoring the providers right there can be a migration module. So, this was monitoring module it says that the QoS experienced by the VMs for different providers and you put it put to the provider trust calculator right and there is other component is a migration module, which is the migration deciders. So, based on this provider trust calculation I can this is it has goes for a input of the migration decider, and this migration module no keep a track that which are the providers there what are their loading capacity and type of things, and based on those it goes on a which need to be which providers need to be selected. So, it goes for a provider selection right.

So, one side that requests requirement, so the selection procedure provider selection procedure is dictated primarily this monitoring of the different QoS parameters, and type of things. So, this may be one such approaches, but there are a lot of nitty-gritty into the things, like customer requirement how it will be taken, whether it is a very crisply

defined or there are fuzziness of the things. If there is a fuzziness how to handle this fuzziness, or how to account for these fuzziness, there can be fuzzy inference engine which are can be deployed out here. This particular things which we are discussing is to of my students work for their projects one for Btech and one for Mtech projects, on working on an intelligent broker for cloud market phase using a fuzzy approach using a fuzzy inference engine.

(Refer Slide Time: 16:16)

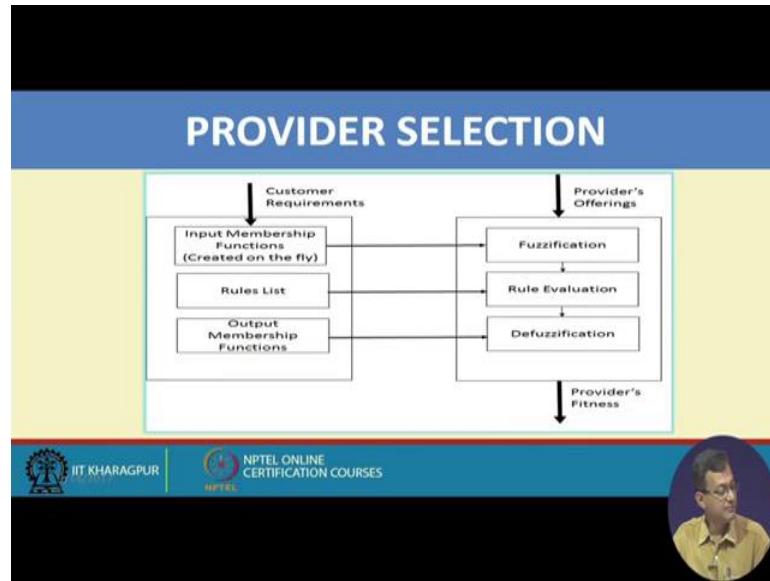
The slide has a blue header bar with the title 'PROVIDER SELECTION'. Below it is a yellow section containing a bulleted list of five items. At the bottom is a dark blue footer bar featuring the IIT Kharagpur logo and the text 'IIT KHARAGPUR', the NPTEL logo and the text 'NPTEL ONLINE CERTIFICATION COURSES', and a portrait of a man.

- Selection of provider is done using a fuzzy inference engine
- Input : QoS offered by a provider and its trustworthiness
- Output : Suitability of the provider for the customer
- Customer request is dispatched to provider with maximum suitability
- Membership functions are built using the user requirements

So, there are a major block diagram of this architecture. So, provider selection if you look at selection of the provider is done using a fuzzy inference engine right. So, there can be other approaches, but in this case what we did in fuzzy inference engine. So, input is the QoS offered by a provider and its trustworthiness right. So, it takes the input as a QoS from offered by the provider and its trustworthiness, output is suitability of the provider for the customer right.

So, I have that one side that provider's competence. So, to say that quality of services offered by things and its trustworthiness other size what we have these customers or the user's requirement right.

(Refer Slide Time: 17:58)

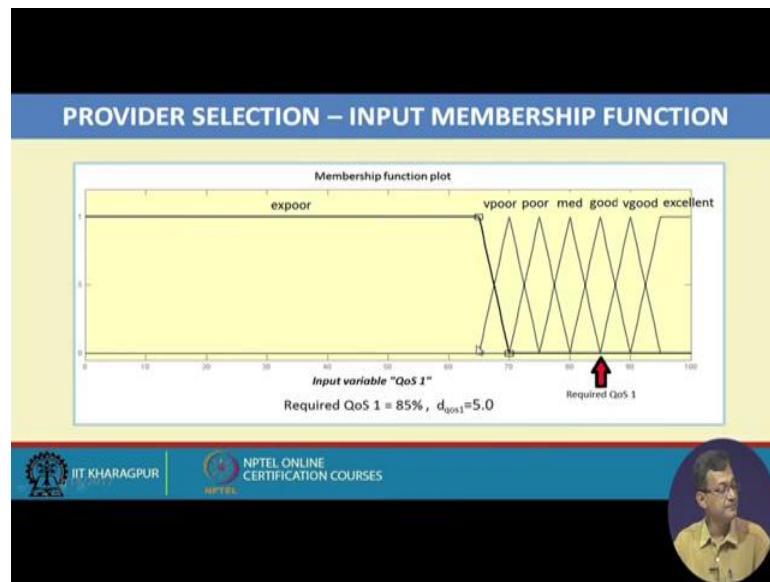


So, and these this fuzzy inference engine in this case takes care or consider these two, and finally, give a output that which should be the base providers for this particular customer.

So, that is the goal of the finding a suitable provider in a cloud marketplace right; customer request is dispatched to provide with maximum suitability, membership function are build using user requirement. So, the membership for function on the providers on the customer side can based on user requirement that can be build on the fly and then put into the system.

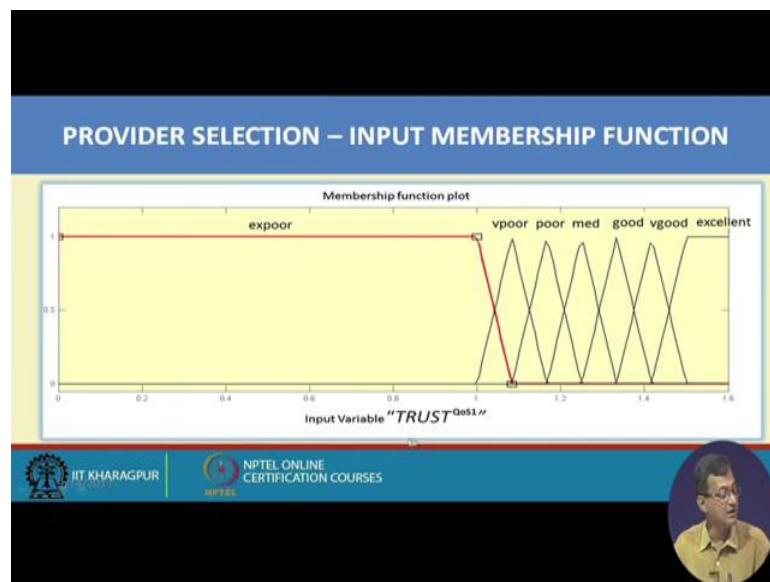
So, they are not only that there will be a set of rule set on the customer side that need to be accounted for also. So, in from the customer requirement input membership function for the fuzzy system, and there is a rule list and there is output membership function based on those things right whereas, the there is providers offerings like what the provider can offer, fuzzification of those things there is a rule evaluation based on the customer rule list and de-fuzzification and providers fitness to the customer requirement, there is other aspects of the other aspect of the thing right.

(Refer Slide Time: 18:44)



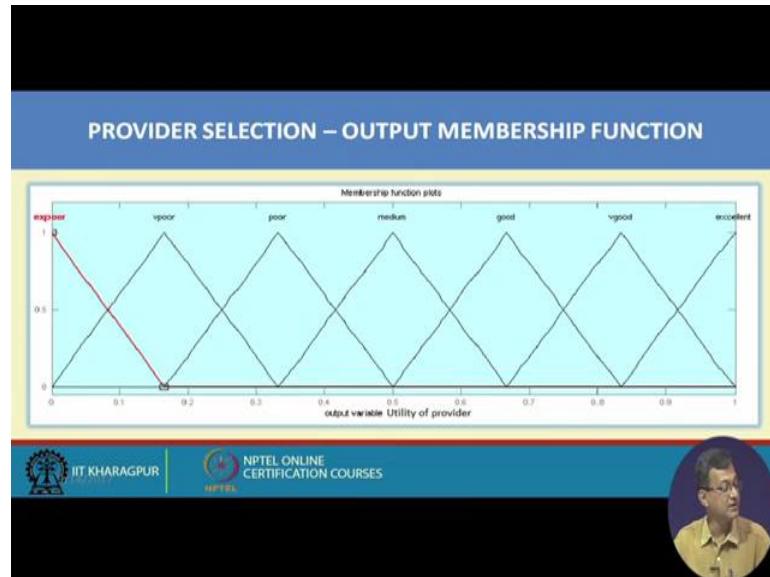
So, this whole thing leads to provider selection. So, this basically fits into the system and it keeps provider selection. So, like a typical scenario where for this experimentation that provider selection input membership function, with different thing like extremely poor very poor medium, good very good excellent and this type of different states and using this input fuzzy membership function.

(Refer Slide Time: 19:15)



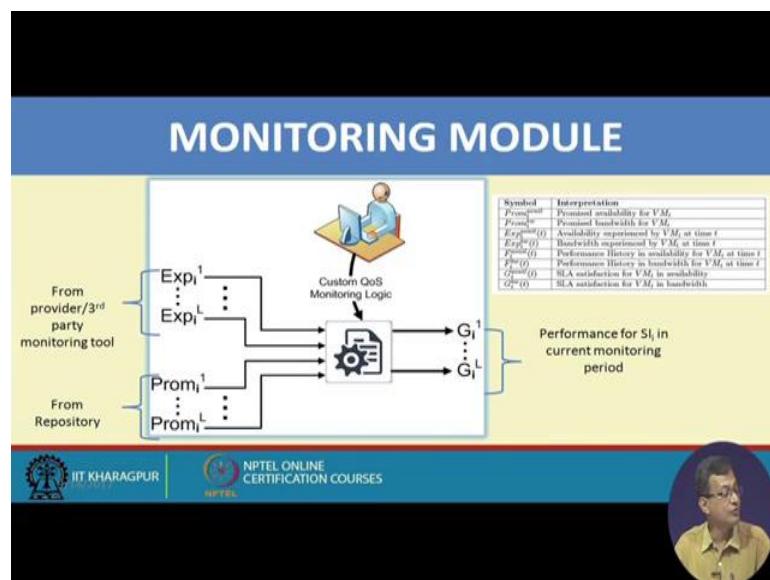
So, what we do that is based on the QoS and based on these, that what is the input variable is the I means with input has trust that having that membership function again at different scale.

(Refer Slide Time: 19:27)



Then we construct that output membership function, generally the output membership function with those levels of extremely poor very poor medium good very good and excellent. Based on this we calculate that what should be the suitable provider for the things.

(Refer Slide Time: 19:41)



There is a monitoring module which monitors the overall processes. So, it takes different available experience values that is availability experienced by the particular VM bandwidth experienced by a particular VM or a particular time then the what are the different promise for different VMs by a particular provider and so and so forth, and generate that SLA satisfaction for a particular VM I on availability SLA satisfaction of a particular VMI on bandwidth and so and so things.

So, performance of this sort of service provider or service instances for a particular service instance I, in the current monitoring period is generated. So that means, what we try to do this, they are we are trying to monitor be service they are performances of different providers and based on different service instances right or a size right that how they are performing this is important for selecting a particular provider.

So, what we are doing if you look at the big story. So, we have one side the customer it has some requirement, I need to fit this requirement to some provider or a set of provider and such as the customer that. See this is the provider is there for the provider in there is a continuous monitoring of the thing right, that with even with performance histories right. So, what are the different performance, what is the promise value and what is the experienced value right one is the promise what the cloud provider gives, and what the experience and the based on this promise like availability bandwidth how much is experienced by the user or the customer with at different point of time.

So, while calculating a at time t, I consider for different provider this monitoring unit gives that for different service instances what how is the performance of this of a particular provider or at a particular provider I based on that. And taking the suitability of the customer we the system sets a then this is the possible provider, or it can even rank that is the first one second one and so and so forth.

There will be a different other factors like here may because we are considering availability and bandwidth there can be other different other factors, even cost is a factor and requirement plays a big role that whether mine is a time critical or say data critical or more accuracy is more important or the time is more important and type of things come into play.

(Refer Slide Time: 22:54)

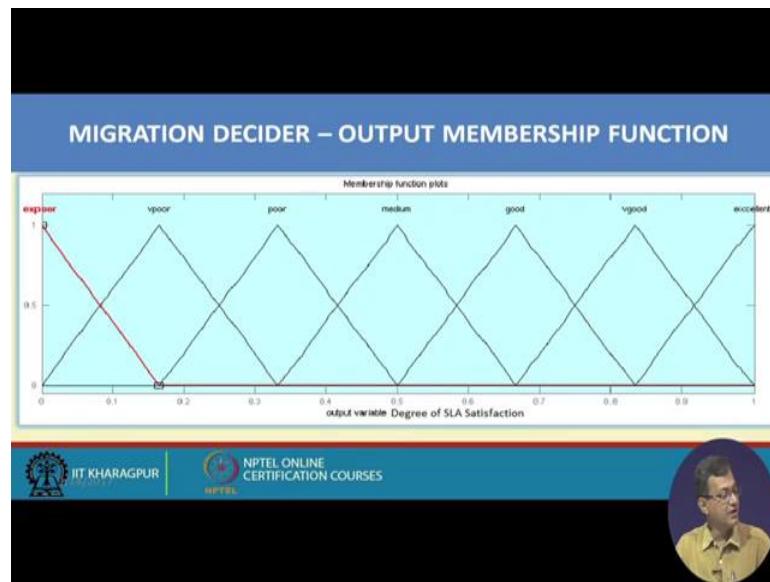
The slide has a dark blue header bar. Below it is a light blue bar containing the title 'MIGRATION DECIDER' in white, bold, uppercase letters. The main content area is yellow and contains a bulleted list. At the bottom is a dark blue footer bar with the IIT Kharagpur logo and the text 'NPTEL ONLINE CERTIFICATION COURSES'. There is also some small text on the right side of the footer bar.

- Makes use of a fuzzy inference engine
- Input:  $F_i^1, F_i^2, \dots, F_i^L$
- Output : *Degree of SLA Satisfaction* for  $SI_i$
- If *Degree of SLA Satisfaction < threshold*, migrate

So, if there is something running on a particular system provider then if there is a need of migration from one provider to another. So, there is a need to have a thing called fuzzy there is a another decider block is needed for migrating for the migration decision right. So, it used again it uses the fuzzy inference engine, there can be different input f one f two f three f three for different providers and output will be the degree of a SLA satisfaction for a service instance I right. If the degree of a SLA satisfaction is less than threshold is then migrate.

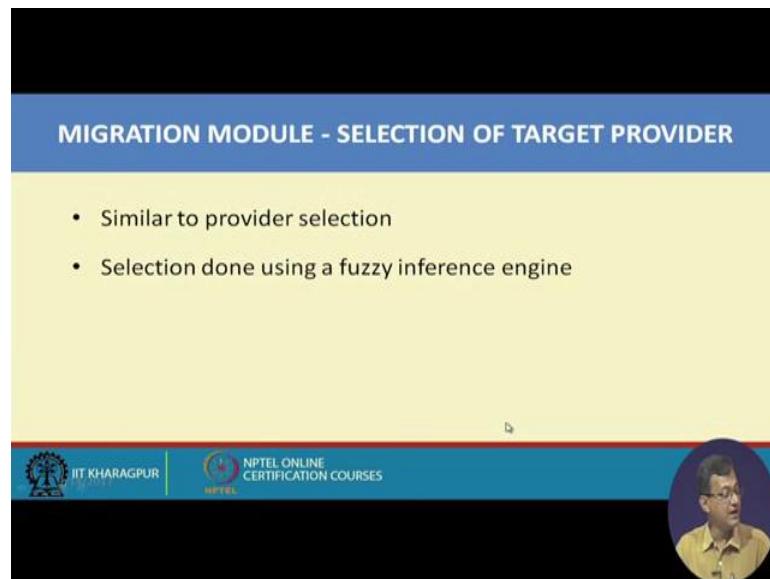
So, what I what we are doing? I am think I am my particular from the customers particular service instance is running. So, and degree of satisfaction is being calculated if the degree of satisfaction is less than the threshold there is a need to migrate to a new provider or a new service instance right.

(Refer Slide Time: 23:59)



Here also this it is done through fuzzy inference engine, and this is a typical instance of a output membership function with different levels.

(Refer Slide Time: 24:19)



We shows that the degree of SLA satisfaction based on that, it is decided that whether the migration is needed or migration is need to be executed for the for a particular service instance.

So, similar the selection of the target provider is somewhat similar there similar to this provider selection procedure, which is being monitored and so on and so forth selection done using a fuzzy inference engine in this case also.

(Refer Slide Time: 24:47)

The slide has a blue header bar with the title "Case study on IaaS Marketplace". Below it is a yellow section containing a bulleted list of requirements:

- 10 providers with varying offered QoS
- 500 requests for VMs
- Year long simulation
- Few providers exhibit performance degradation. Degraded QoS parameters follow a Gaussian distribution
- Comparison made with conventional (minimum cost) crisp broker

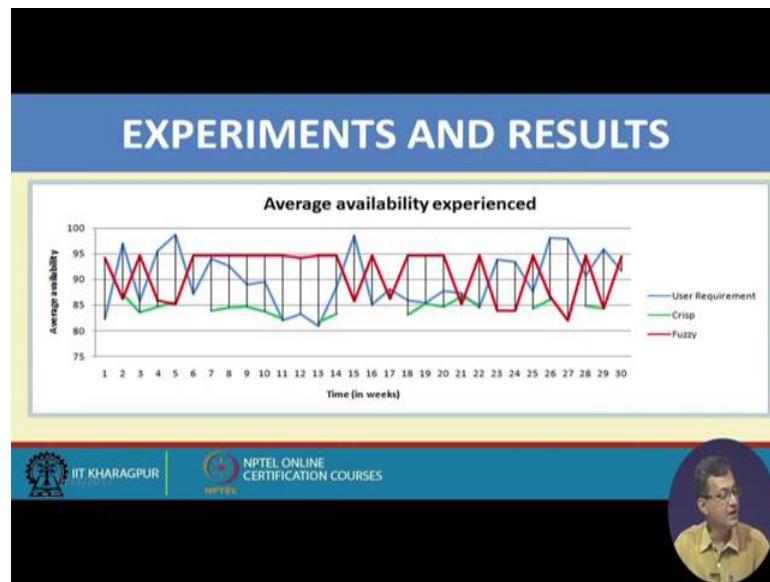
At the bottom of the slide, there are two logos: IIT KHARAGPUR on the left and NPTEL ONLINE CERTIFICATION COURSES on the right.

So, what we have here is that, it is true for any type of service provider though primarily it is more pertinent for is type of market place, where this infrastructure as a service it is being provided, but it is true for any type of cloud right. So, here we did some case study is to. So, that whether how much effective it was that whether it makes sense to fuzzify the or using this fuzzy approach.

So, we did a simulation with 10 providers with varying offer here varying offered keyways 500 requests for VM are considered, there is a done through a yearlong simulation; that means, over a year the time period has been taken as a year taking on a daily basis. Few provider exhibit performance degradation. So, degraded QoS parameter follows a Gaussian distribution. So, the if there is a some sort of a degraded QoS has been done using a Gaussian distribution and comparison made with conventional minimum cost crisp broker.

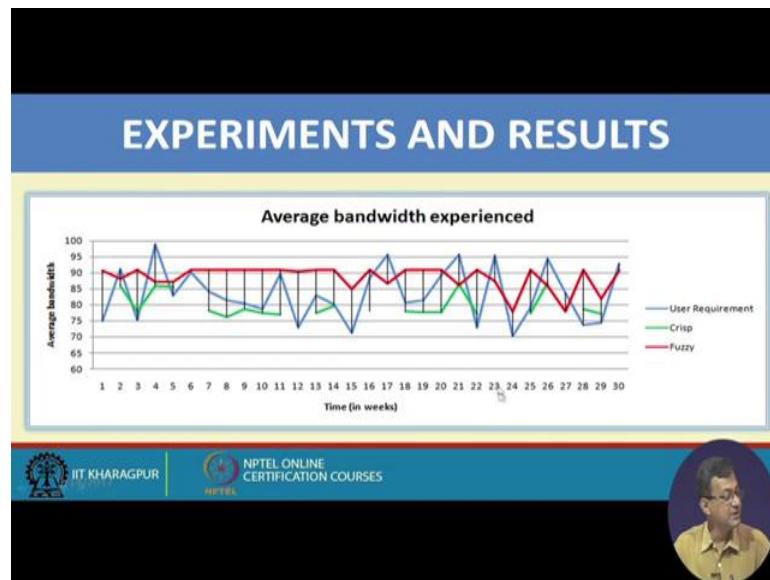
So; that means, there is no this sort of fuzzy system if there is a crisp broker; that means, it takes a call yes or no type of a things then with a with minimal cost on those decision making.

(Refer Slide Time: 26:43)



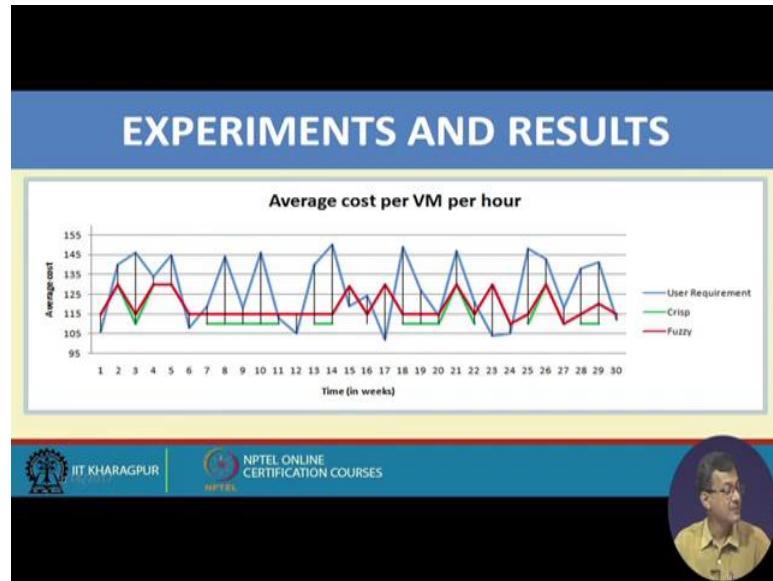
So and then how it performs compared to that. Like one the one is showed that the average availability this blue color is the user requirement green is the for the crisp broker and red is the fuzzy. So, what we see that it is it may not be always following, that user requirement, but in a sense it is better than the crisp approach.

(Refer Slide Time: 27:08)



Similarly, this is for average bandwidth again, here blue is the user requirement, red is the fuzzy and green is the crisp broker.

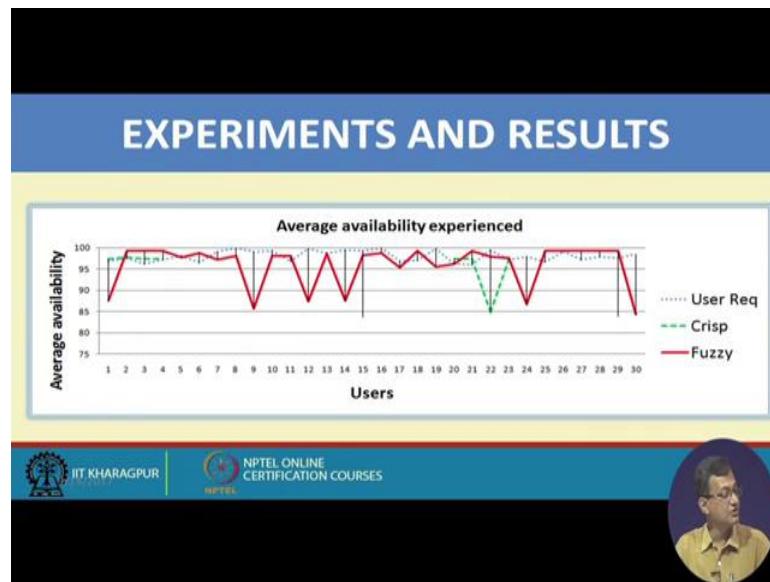
(Refer Slide Time: 27:28)



So, here also we can see that is better than this crisp broking type of thing. Similarly average cost per VM per hour the if we look at the same way, that over a scale the fuzzy gives a better result.

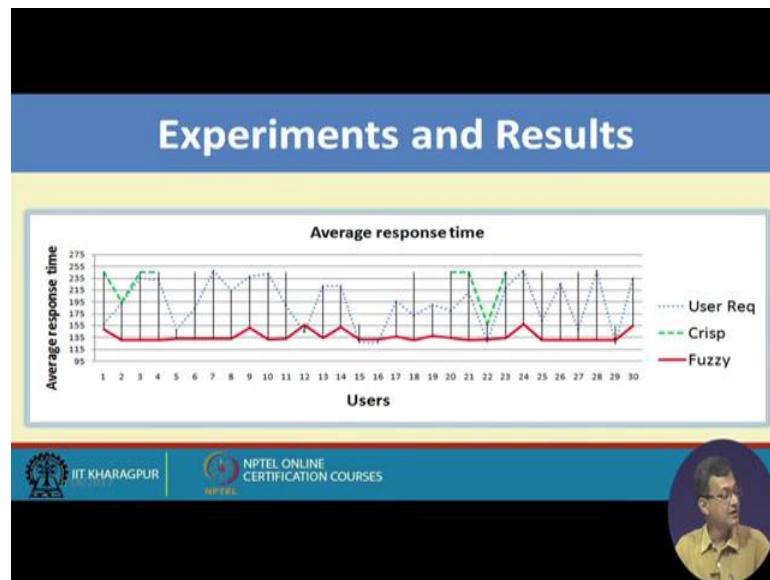
So, similarly though it is predominantly for SaaS is type of provider, but we did some experimentation on the SaaS or some simulation on the SaaS marketplace. So, here also some sort of a 10 providers are considered with 500 service requests again simulated over 365 days, few provider exhibit performance degradation QoS parameter follow a Gaussian means QoS degraded degradation degraded QoS follow a Gaussian thing and compression with the conventional crisp thing.

(Refer Slide Time: 28:27)



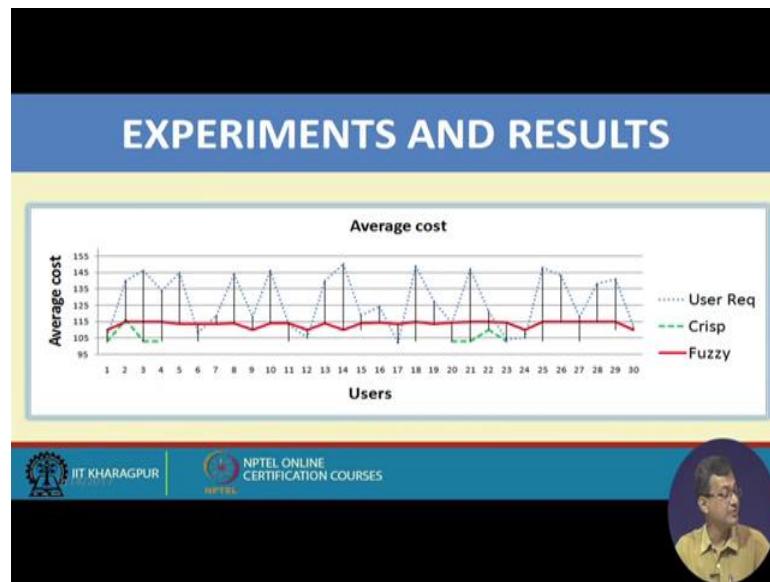
Here also what we see that you the dotted line is the user requirement, red is the fuzzy and crisp is the green. So, over again arrange the fuzzy seems to be following better, that may not be always the case, but it is likely to be better it is likely to be mapping the user requirement better way that fuzzy inference engine.

(Refer Slide Time: 28:54)



Similarly, this is on the average response time over again the particular period of here in this case 30 days and that how the overall performance is there with fuzzy crisp.

(Refer Slide Time: 29:16)



And with respect to user requirement similarly this is the average cost of a particular the instance of the things.

So, what we see that we basically implement this sort of a fuzzy inference engine, in order to map these ah user requirement two way to the rather mess the user requirement with the providers offerings and try to feed the best provider with the user with the particular user requirement or set a for a particular user of the for a particular service instances.

(Refer Slide Time: 30:07)



And what we like to look at there are several future scope, especially in the research formed for a birth for this particular one there is a specification of flexibility in the user QoS requirement is there right that is whether we can make it flexible of QoS the requirement, comparison again existing approaches on production workload right.

So, that on a life workload whether we can compare several classes of customers, now there can be you see there customers can be categorized into different classes of customers right it is instead of taking individual customer or user I say that this is user is this category of customer this user is this class area customer and so and so forth. And it all and also depends on the type of applications type of services is looking for, in those type of things and extending the things for any type of services or XaaS type of services may be a important aspects of the things.

Never delays this type of finding appropriate match in a cloud marketplace is gaining a lot of interest both from the real life point of view or commercial point of view, and from the research point of view how to find a suitable things how to measure or how to say it formally that what has been selected suitable is the best possible things right. If not the best, but it is something near based with some sort of a guaranteed type of services

So, this is again those who are interested in future study research. So, this is a important aspects of a because it involves lot of aspects right. is that how to monitor them how to calculate trust competence and SLA guarantees or SLA values says or how SLAs has been satisfied over time and based on that at a current at a at a present time t how to project that what are the things are there. So, it is it plays the important role in in modern day cloud infrastructure or making cloud computing more popular across the things

So, with this discussion we stop today. And we will continuing with some of some of the new some of these aspects of cloud computing, from different perspective in our future lectures.

Thank you.

**Cloud Computing**  
**Prof. Soumya Kanti Ghosh**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**

**Lecture - 31**  
**Mobile Cloud Computing – I**

Hello. We will continue our discussion on Cloud Computing. Today we will look at one aspect of or on one application of this cloud computing or would you say that amalgamation with other technology, what we look at in a mobile cloud computing. Though mobile cloud computing itself has now became a topic or subject as a whole, but we will try to look at the different aspects of mobile cloud computing basically overview of the things, that where what are the different features characteristics where is the need and so on and so forth.

So, as we see in today's world, these mobile devices or what we say Smartphone devices or smart mobile devices are have a huge purification in our society right from different category of people to different level of application and so on and so forth. So, that has created a revolution in the communication path. So, there are a couple of things one is that this huge purification has is possible or because of easy availability or low cost high bandwidth availability at the back end. So, there are service provider or back end service providers which provides networking to these devices.

Secondly, we also have a scenario that with very resourceful. So, to say quote unquote resourceful in the devices are coming up, which are much resourceful than yesterdays devices right. So, there is a upliftment of the resources till that is not matching with the type of things which we have on the desktop or other type of servers, but nevertheless it is a resourceful device, it is a smart device it has in capability of not of sensing and running lot of apps and you have a backbone networking which allows you to communicate in a bigger way. So, all those things has given up a trend that whether I can use these device for computing purpose right.

So, or if I have some applications which is running, which requires some good amount of computing where there is a provision of uploading this sort of computing activity to some other more resourceful equipment or something. So, here they came as a natural need of that interfacing between the cloud and mobile devices right. So, that mobile

devices it can offload it is some of these computing phases or computing modules to the other to the cloud and get it back and go on running, that could have been marvellous thing right. Like I can say that I am sensing some information environmental things then I want to do a predictive model that what is going to be there very. So, very short term prediction and then that predictive model requires.

So, much more resources than sensing and doing in some initial analysis. So, I what I do that this thing is offloaded to a much higher resourceful infrastructure say cloud, and I get the results and analyze the result at my end or with other collateral data. So, there can be lot of type of applications which involves some sort of a data analysis data analytics which could have been offloaded other things. So, this sort of amalgamation whether it is possible, what sort of architecture is there what are the different type of challenges that we would like to look at in this lecture maybe one or two cop consecutive lectures.

(Refer Slide Time: 04:30)



So, today we will talk about basics of mobile cloud computing what are the things. So, what we see what motivates as we are discussing, there is a huge growth of Smartphones or uses of the Smartphone and not only the phone of smart applications right. Increased capability of the mobile devices in terms of running these applications and resources, access of internet mobile devices became pretty easy like this connectivity of these mobile devices with a back end internetworking; resource and on the other hand there are some of the things resource challenges like battery life storage bandwidth these are some

of the resources challenges, if there is if the application is heavy that is more require more resource to execute this or what we say resource hungry applications right.

So, in mobile devices cloud computing on the other hand offers the advantage to the users by allowing them to use infrastructure platforms software by cloud providers as a low cost and elastically in an on demand fashion as we have seen that the cloud computing on the other hand provides the user with a low cost elastic service like. As pay as you go model and whether there is a possibility of amalgamation exactly what we are looking trying to say we need to talk about mobile cloud computing or sometimes abbreviated at MCC right. So, to have something mobile and something at your finger tips while you are move.

(Refer Slide Time: 06:04)

**MobileBackend-as-a-service**

<b>What</b>	<ul style="list-style-type: none"><li>Provides mobile application developers a way to connect their application to backend cloud storage and processing</li></ul>
<b>Why</b>	<ul style="list-style-type: none"><li>Abstract away complexities of launching and managing own infrastructure</li><li>Focus more on front-end development instead of backend functions</li></ul>
<b>When</b>	<ul style="list-style-type: none"><li>Multiple Apps, Multiple Backends, Multiple Developers</li><li>Multiple Mobile Platforms, Multiple Integration, Multiple 3rd Party Systems &amp; Tools</li></ul>
<b>How</b>	<ul style="list-style-type: none"><li>Meaningful resources for app development acceleration – 3rd party API, Device SDK's, Enterprise Connectors, Social integration, Cloud storage</li></ul>

<http://www.rapidvaluesolutions.com/whitepapers/How-MBaas-is-Shaping-up-Enterprise-Mobility-Space.html>

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, something what we trying that mobile backend as a service type of things like as we try to do everything XaaS is type of things x, x anything as a service. So, why not mobile backend as a service like, what we when you look at it? It provides mobile application developers a way to connect to their applications back in cloud provider and cloud storage and processing. Abstract away complexity of launching managing own infrastructure right that is a need. Focus more on the front end development instead of the back end functionalities right.

So, that it is more on the how this apps will be develop rather looking at that how my back end how to manage the back end and so on and so forth. So, multiple apps multiple

back end, multiple developers things are possible multiple mobile platforms multiple integration multiple third party systems and type of things. That means, there is a integration or multi party type of systems and meaningful resource for app development like third party API devices decays enterprise connected and so and so forth.

(Refer Slide Time: 07:28)

The slide has a yellow header bar with the title 'Augmenting Mobiles with Cloud Computing'. Below the title is a bulleted list of initiatives:

- Amazon Silk browser
  - Split browser
- Apple Siri
  - Speech recognition in cloud
- Apple iCloud
  - Unlimited storage and sync capabilities
- Image recognition apps on smart-phones useful in developing augmented reality apps on mobile devices
  - Augmented reality app using Google Glass

At the bottom of the slide, there is a footer bar with the IIT Kharagpur logo, the text 'NPTEL ONLINE CERTIFICATION COURSES', and a circular profile picture of a man in a yellow shirt.

So, there are different what why when how type of things, which are which there is a possibility of answering this type of situations. Now augmenting these mobiles with cloud computing there are several efforts or initiatives like amazon silk browser, which is a split browser, Apple Siri, Apple iCloud, image recognition apps on Smartphones useful in developing augmented reality apps in the mobile devices. So, all those things at the back end it talk to a cloud right. So, it may be for storage, it may be for computing, it talk to a cloud. The things are possible primarily because you have a intermediate high bandwidth scenario right, that you seamlessly things that as if the calculation is working on the mobile device.

So, that this type of delay can be metalized can be handled right. So, whenever I offload any application or any computing things to some other things a couple of things come into play right. One is that there should be a fill of that the as if the application running on the device itself so; that means, intermediate delays in offloading. So, this particular portion of the application the whole application should be minimum right. There are other things like if I want to do a dynamic offloading, then I need to do lot of things like

I need to appropriately partition the things, need to have a lot of synchronization in to play and if there are dependencies with other applications those has to be taken care.

So, this is not a straightforward scenario there are lot of lot of complexity involved in unit, and lot of timing relationships and other things come into play right.

(Refer Slide Time: 09:17)

**What is Mobile Cloud Computing?**

*Mobile cloud computing (MCC) is the combination of cloud computing, mobile computing and wireless networks to bring rich computational resources to mobile users.*

- **MCC provides mobile users with data storage and processing services in clouds**
  - ✓ Obviating the need to have a powerful device configuration (e.g. CPU speed, memory capacity etc.)
  - ✓ All resource-intensive computing can be performed in the cloud
- **Moving computing power and data storage away from the mobile devices**
  - ✓ Powerful and centralized computing platforms located in clouds
  - ✓ Accessed over the wireless connection based on a thin native client

**IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES**

<https://www.ibm.com/cloud-computing/learn-more/what-is-mobile-cloud-computing/>

So, if it can be managed then we can have scenario things like computing. So, if you would try to look at what mobile computing is as a definitions, there are lot a there are a number of definitions available across the internet, few we are trying to look at.

So, it is a combination of cloud computing mobile computing and wireless network, intermediate wireless networks to bring rich computational resources to the mobile users. So, that is one of the aspects of the thing. So, MCC provides mobile user with data storage and processing services on cloud. So, it is provide mobile users with data storage and processing service that primarily the two things. So, other thing is that moving computing power and data storage away from the mobile device. So, in some case in other cases what we do that the I detach this computing power or migrate this computing power and the storage to this cloud infrastructure.

(Refer Slide Time: 10:15)

**What is Mobile Cloud Computing?**

*Mobile cloud computing (MCC) is the combination of cloud computing, mobile computing and wireless networks to bring rich computational resources to mobile users.*

- **MCC provides mobile users with data storage and processing services in clouds**
  - ✓ *Obviating the need to have a powerful device configuration (e.g. CPU speed, memory, etc.)*
  - ✓ *All Mobile Cloud computing is the combination of cloud computing and mobile networks to bring benefits for mobile users, network operators, as well as cloud providers*
- **Moving computation to the cloud**
  - ✓ *Provides access to rich computational resources*
  - ✓ Accessed over the wireless connection based on a thin native client

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | <https://www.ibm.com/cloud-computing/learn-more/what-is-mobile-cloud-computing/>

So, this two primarily things are there and we can say it is a combination of cloud computing, mobile networks to bring benefits of the mobile users network operators as well as the cloud providers. So, it is a some sort of a what we quote unquote win situation to a have all those things.

(Refer Slide Time: 10:32)

**Why Mobile Cloud Computing?**

**Speed and flexibility**  
Mobile cloud applications can be built or revised quickly using cloud services. They can be delivered to many different devices with different operating systems

**Shared resources**  
Mobile apps that run on the cloud are not constrained by a device's storage and processing resources. Data-intensive processes can run in the cloud. User engagement can continue seamlessly from one device to another.

**Integrated data**  
Mobile cloud computing enables users to quickly and securely collect and integrate data from various sources, regardless of where it resides.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, why computing mobile computing or MCC; this is well understood first of all what we look at the speed and flexibility, mobile applications can be build on or revised

quickly using the cloud services. So, two things are there, that I can have a lot of speed up on the things right.

So, computing as cloud are more resourceful things it can be do immediately there is another thing is that if there is a apparatus and needs and something has to be there like suppose I am using a algo for finding a minimum path or minimum distance between point a and point b on a typical map, then I can if there is a new heuristics come up I change at the clouded and it is seamless for my front end application at the mobile device, whether to how to offload on the cloud. It is only the duty to offload the things the rest is calculation is there right.

So, based on that different situations or different scenarios, I can have different algos and things coming up. So, them it is something a detachment detaching that actual processing from the devices. So, it is both flexible and speedy thing shared resources I can use shared resources mobile apps that can run on cloud are not constrained by device storage and processing resources data intensive processes can run on the cloud, user engagement and continuing seamless processes can round on the devices right. So, what do we say that some of the things can need to be done on the mobile devices some can be offloaded, that type of dynamic deals and partitioning is still a major challenge, right.

An integrated data, mobile cloud computing enables users to quickly and securely collect and integrate data from various sources regardless of their where it resides right like I want to do some sort of a disasters management type of things. So, if there is a disaster management system. So, on very quickly lot of data need to be collected by different sensors, which are mobile base and need to be run to see that what sort of other what sort of mitigation techniques to be there. So, there can be lot of applications which run at the back end which is in the cloud and can be integrated seamlessly along with the different sort of heterogeneous data sets.

(Refer Slide Time: 13:04)

**Key-features of Mobile Cloud Computing**

*Mobile cloud computing delivers applications to mobile devices quickly and securely, with capabilities beyond those of local resources*

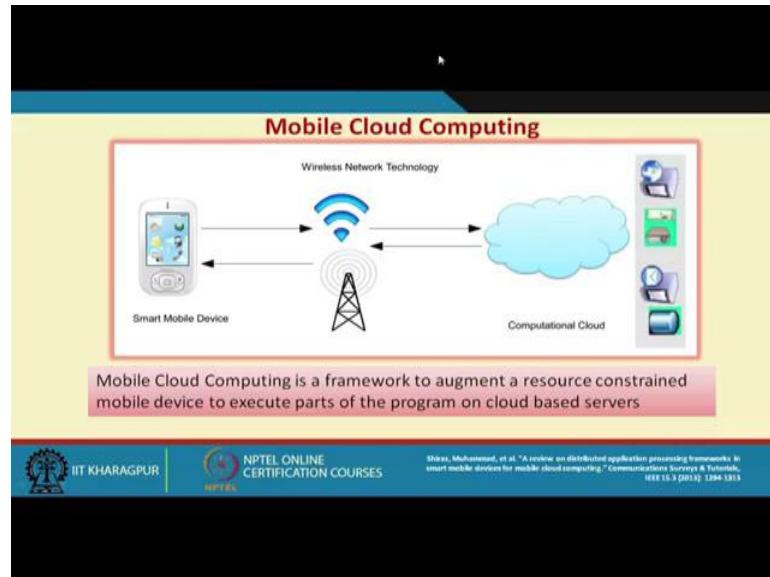
- Facilitates the quick development, delivery and management of mobile apps
- Uses fewer device resources because applications are cloud-supported
- Supports a variety of development approaches and devices
- Mobile devices connect to services delivered through an API architecture
- Improves reliability with information backed up and stored in the cloud

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, if we look at the key features, the cloud computing delivers application to mobile devices quickly and securely and capabilities beyond those of local resources, like facilitates quick development delivery and management of mobile apps, uses fewer device resources because the applications are cloud supported. So, I have less the device resources or I am less loading the device resources, it is less battery power, less heating up and you can run a lot of other apps into the thing right. Supports a variety of development and approaches where a development approaches and devices mobile is as it is doing at the at the cloud end.

So, that it can a same type of computing thing can support number of apps so; that means, easy to standardize across the things or running the reusability of some applications or reusability of algorithms is much higher. Mobile devices connect to services delivered through an API architecture improves reliability and information backed up and stored on the cloud right. So, it is not device dependent suppose you store the data at the device if the device goes down then the whole thing goes down, but is if I have a ah storage in the somewhere in the cloud with the other reliability into the place, like reliable storage facilities then my it is the my information is more reliably stores and can be handled.

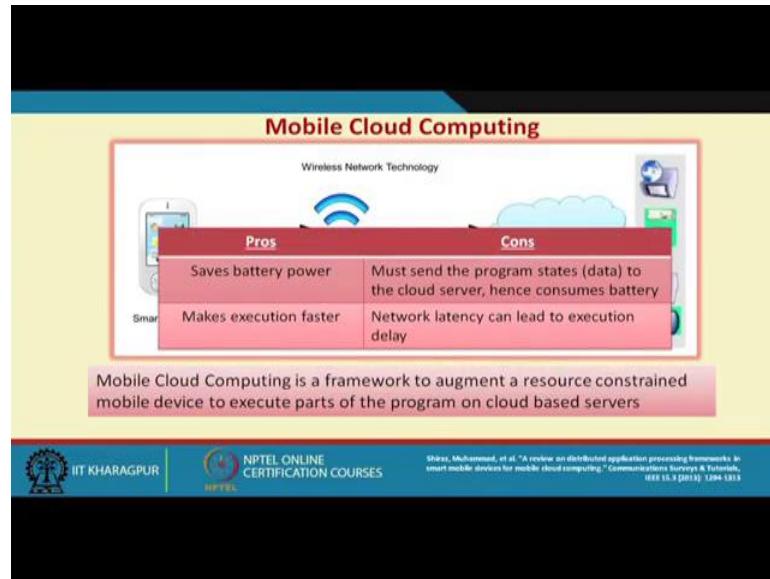
(Refer Slide Time: 14:41)



So, if we look at. So, one end is this mobile devices or smart mobile devices sometimes we abbreviate as SMD other end is a computational cloud, and in between we have a wireless network technology provided by the service provider right. So, mobile cloud computing is a framework to augment resource says resource constrained mobile devices, to execute parts of the program on the cloud based servers right.

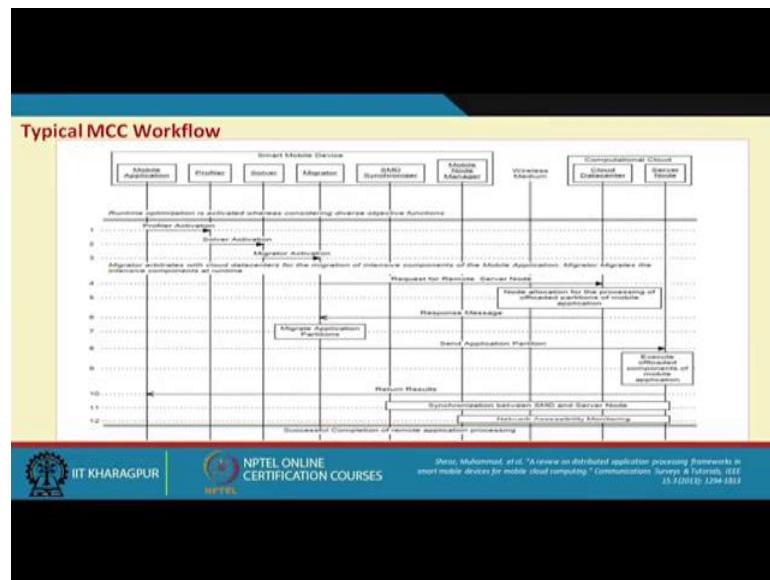
So, that is there no one can argue that these days mobile devices are much resourceful, but there is no match with the type of resources like cloud or a cloud-based servers can provide right. So, there is absolutely no match on the things so; that means, you can run now more powerful or more resource hungry applications into the device.

(Refer Slide Time: 15:34)



So, there are pros like saves battery power as you are using less computing, makes execution faster at times if you have a good bandwidth and type of things. There are definitely the flip side of it must send the program state data or the state or state and data and the cloud server hence consume battery right. So, it is additional loading of transferring data network latency can lead with excuse and delay.

(Refer Slide Time: 16:09)



So, if there are later on latency. So, there can be execution delay into the system. So, this is a overall big picture a flowchart. So, if you see that is one part is related to the mobile

devices, other part is more on the cloud computing things and how this workflow will go on. So, there are mobile apps. So, that should be a profiler which profiles this different that particular app and devices there is a solver of which takes care of that how this if I partition these and how it need to be profiled to be sent to the things, then how to solving will be there that is a SMD synchronization SMD stand for smart mobile devices, device and synchronizer that because if we do partition and execute in the on the cloud.

Then we need to take care of the how this synchronization process will go on and there is more while load manager right as to takes care that it says that data to the cloud captures the data, and how things are managed on the cloud and we have that cloud data sender and server nodes which are already we have seen we have discussed a lot on those things that.

So, it for the cloud it is a some sort of a third-party application or what we say there some third party a user requests coming to the things need to be executed within a particular queues within ah queues, and served with some SLA's. And need to be served right and in between we have a wireless media which plays a important role though it is neither computing or giving directly doing anything with the app, but it plays a important role in the sense that it takes care of this latency right or you need to take care of this latency.

(Refer Slide Time: 17:57)

**Dynamic Runtime Offloading**

Dynamic runtime offloading involves the issues of

- dynamic application profiling and solver on SMD
- runtime application partitioning
- migration of intensive components
- continuous synchronization for the entire duration of runtime execution platform.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, if we look at the runtime offloading there are different tricky issues right or factors like a dynamic runtime offloading involves the issues of dynamic application profiling and solver on SMD right runtime application partitioning you need to partition the it is runtime migration of intensive come component like which are computational intensive component need to be migrated.

Continuous synchronization of the entire duration of the runtime execution platform, the long it is running we need to synchronize the things. So, if you look at these are very very very tricky issues right, these are these are not any straight forward things anything doing dynamically runtime is extremely challenging right and. So, dynamic application profiling and solver in the SMD on that the mobile device runtime application partitioning, you need to partition on the runtime, migration of intensive components and continuous synchronization of the inter duration of the runtime execution platform.

(Refer Slide Time: 19:18)

**MCC key components**

- Profiler
  - Profiler monitors application execution to collect data about the time to execute, power consumption, network traffic
- Solver
  - Solver has the task of selecting which parts of an app runs on mobile and cloud
- Synchronizer
  - Task of synchronizer modules is to collect results of split execution and combine, and make the execution details transparent to the user

**IIT KHARAGPUR** | **NPTEL ONLINE CERTIFICATION COURSES**

So, this these are there are major issues we need to be addressed or which are being addressed when you develop this type of mobile computing cloud computing apps. There are several components some of the key components one is this profiler. So, which monitors a application execution to collect data about the time to execute power consumption and network traffic. Solver has the task of selecting which part of an app runs on mobile and cloud. So, it need to looked at that which you need to be use running

at the mobile end and we you need to be uploaded at the cloud end at the one of the main duty of the solver.

Synchronize at tasks of synchronizing module synchronizer module is to collect results of split execution and combine and make the execution details transparent to the user. So, that is important right you have partitioned the things runtime and this split execution things need to be again stitched together and give a user a some feeling that as if it is application run was running seamlessly on the device and without any much delay and type of things. So, this is also a very major challenge in handling this.

(Refer Slide Time: 20:30)

The slide has a black header and footer. The main content area is yellow. The title 'Key Requirements for MCC' is in red. The requirements are listed in a bulleted list.

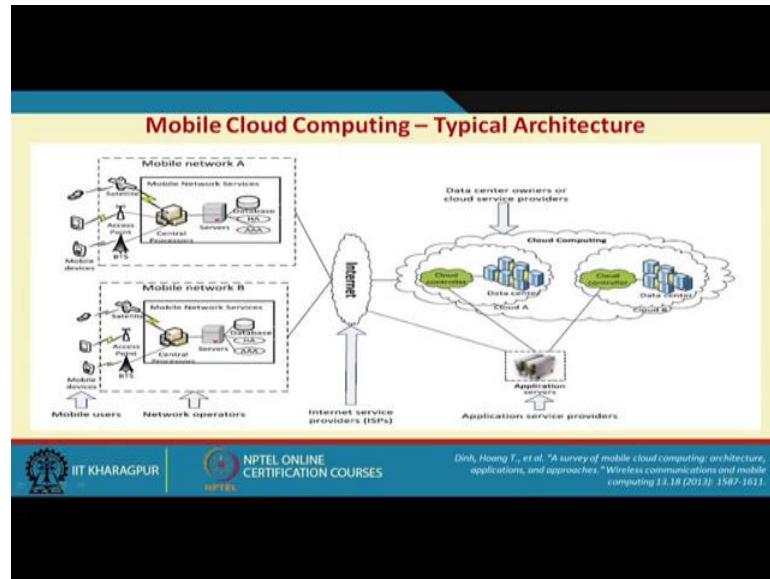
**Key Requirements for MCC**

- **Simple APIs** offering access to mobile services, and requiring no specific knowledge of underlying network technologies
- **Web Interface**
- **Internet access** to remotely stored applications in the cloud

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

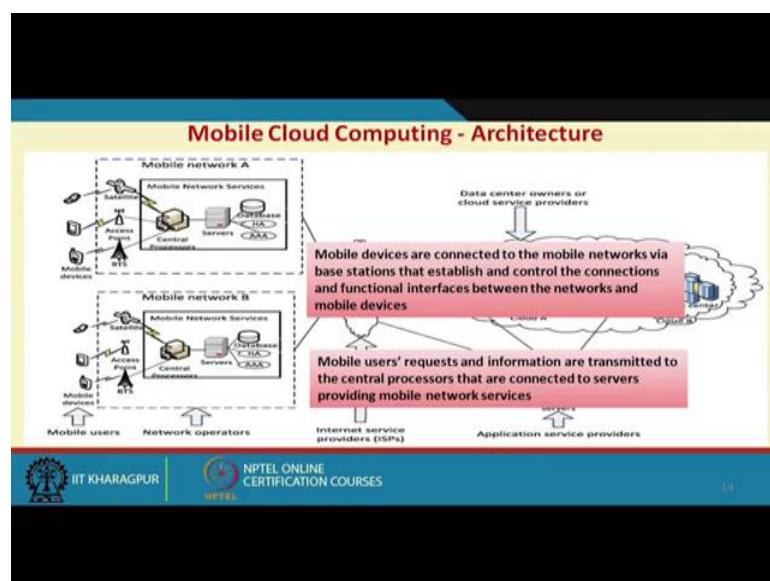
So, there are other components like requirements what we see that simple APIs, API should not be very cumbersome. So, that user is inclined to use it. So, simple API is offering access to mobile devices and requiring no specific knowledge of underlining network technologies right. So, it should be the user should not be burdened with the knowledge of network, technologies and knowing all those things wave interface should we should have a appropriate wave interface internet access to remotely stored applications in the cloud right. So, you should have Internet access to remotely stored application in the cloud, that that is there.

(Refer Slide Time: 21:15)



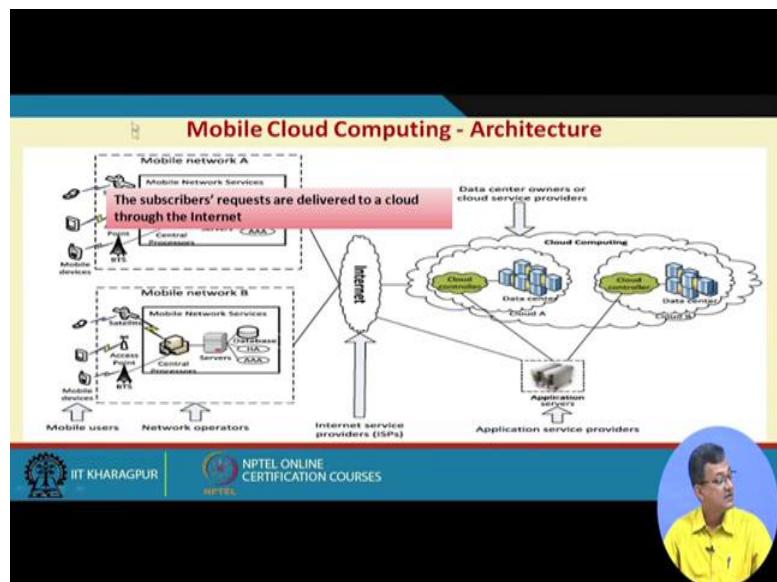
So, if we look at a typical architecture. So, one side we have mobile devices, one side this cloud computing environment and there are different service provider, which provides the network backbone and there are internet services to connect to this cloud right. So, therefore, the cloud it is more of a giving appropriate user interface were connecting this mobile devices or the applications, which are offloaded by the mobile devices onto the cloud. So, there are mobile user and network operators internet service provider and application service provider, so this which need to be appropriately working in sync right.

(Refer Slide Time: 22:00)



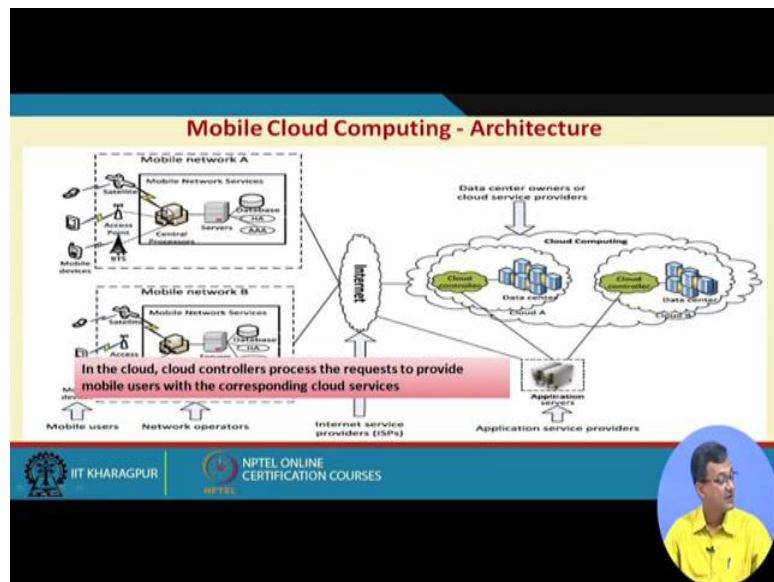
So, mobile devices are connected to the mobile networks via base stations and establish and control the connection of functional interface between the network and the mobile devices mobile user requires the information. So, one is that more devices connected to the base station, and mobile user requires the information are transmitted to the central possessing and are collected by the server providing those things, so that has to be are transmitted. So, there is a transmission of the data.

(Refer Slide Time: 22:30)



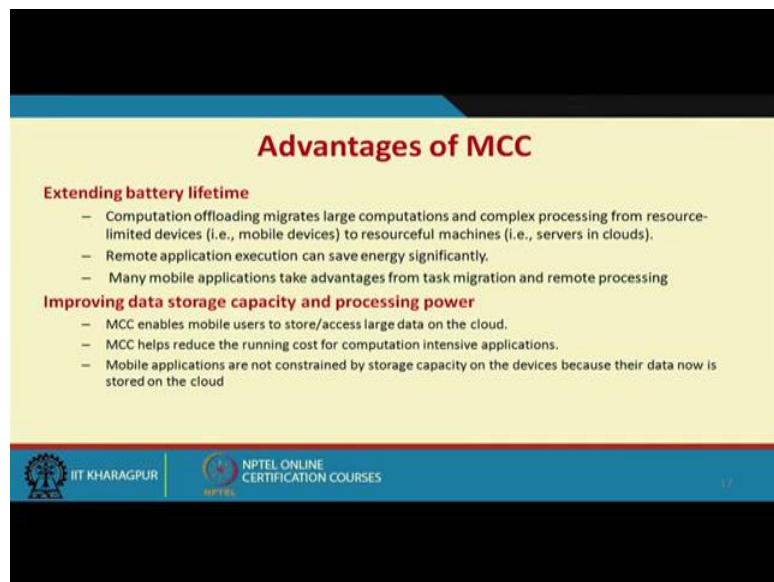
The subscriber requests are delivered to the cloud through the Internet right the subscriber should be transparent to that.

(Refer Slide Time: 22:38)



In the cloud, cloud controllers process the request to provide the mobile user with the corresponding cloud services. So, based on the request that cloud controller process the thing so that, it can give the mobile user the requested services.

(Refer Slide Time: 22:55)



So, there are several advantages already we have looked into few of them and one is that extending battery life time. So, as you are off loading some computing thing, and which is if it is resource hungry, then in a sense you are saving energies and battery life time. Computation offloading migrates large computations and complex processing from

resource limited a device that is mobile devices to resource full machines that is servers in clouds right. Remote application execution can save energy sub secure significantly, many mobile application takes advantage from task migration and remote processing right. So, those are like helping saving battery life time improving data storage capacity and processing power right.

So, if it is a as it is stored on the cloud in with a huge amount of data storage capability. So, it improves that data storage capacity right I can go on continuously go on take data say, I am going on taking environmental data. So, and offloading it. So, that is no loading on my own devices. So, that is a there is a advantage of the things and of course, I have a huge processing power at the backend right provided my network latency is within the permissible limit right. Improve reliability and availability right.

(Refer Slide Time: 24:31)

**Advantages of MCC (contd...)**

**Improving Reliability and Availability**

- Keeping data and application in the clouds reduces the chance of loss on the mobile devices.
- MCC can be designed as a comprehensive data security model for both service providers and users:
  - Protect copyrighted digital contents in clouds.
  - Provide security services such as virus scanning, malicious code detection, authentication for mobile users.
- With data and services in the clouds, they are always(almost) available even when the users are moving.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, that keeping data and application in the cloud reduces chance of loss of the data along with the mobile devices and thinks if there is a loss things are there. MCC can be designed as a comprehensive data security model for both service provider and users though there are lot of security issues, but the can be designed to look into like put it copyrighted digital content in the cloud or provide security services such as virus scanning malicious code detection etcetera. With data and services in the cloud then they there are always available within the when the users are moving right.

So, there is another thing is that, if the data and services are offloaded to the cloud. So, they are you omnipresent right, I can have always wherever I am moving whether this device if I have a sync with the other device or other things are accessing. So, it is always available in nothing, some sort of having a centralized vision of services or processing services and data can be looked into.

(Refer Slide Time: 25:54)

**Advantages of MCC**

- Dynamic provisioning
- Scalability
- Multi-tenancy
  - Service providers can share the resources and costs to support a variety of applications and large no. of users.
- Ease of Integration
  - Multiple services from different providers can be integrated easily through the cloud and the Internet to meet the users' demands.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, there are other advantages like a dynamic provisioning. So, I can dynamically provision the things, scalability issues like as whatever is provided by the cloud that is infinite scalable scalability, multi tenancy like service provider can share the resources and cost to support variety of application in large number of users like as there is a property of the cloud of multi tenant. Ease of integration, multiple services from different providers can be integrated easily through the cloud and the internet to meet the users demand right.

(Refer Slide Time: 26:39)

**Mobile Cloud Computing – Challenges**

**MCC Security Issues**

Protecting user privacy and data/application secrecy from adversaries is key to establish and maintain consumers' trust in the mobile platform, especially in MCC.

MCC security issues have two main categories:

- Security for mobile users
- Securing data on clouds

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, that I can have standardized in one place, I can connect different devices and type of things in through the cloud and I can have integration of the things. There are several challenges one major challenges as we see in other cases also is security issues right. So, protecting user privacy day privacy and data application secrecy from adversaries is a key to establish and maintain consumer trust on mobile platform, especially in case of mobile cloud computing right.

So, how my data processing things are protected from other adversaries or attackers are is important, otherwise at least when a work on the devices I somewhat have a much larger trust on the devices, but if I once I offload, I do not know that what is happening. And most of the cases I am not subscribing to the cloud itself, I am subscribing to the service provider and the whole paradigm right it is offloading to the cloud, it may so happen that the whole framework may select different cloud, based on the availability pricing and different other factors, right.

So, it is sometimes becomes tricky issue on the security point. So, MCC security issues have two major categories like security of mobile users and securing data of cloud. So, privacy also plays the important role because in doing so, it is not only the data of the mobile user, but also my concerned maybe that my identity, my mobility, my GPS footprints are being tracked, right.

(Refer Slide Time: 28:15)

**Mobile Cloud Computing – Challenges**

**Security and Privacy for Mobile Users**

- Mobile devices are exposed to numerous security threats like malicious codes and their vulnerability.
- GPS can cause privacy issues for subscribers.
- Security for mobile applications:
  - Installing and running security software are the simplest ways to detect security threats.
  - Mobile devices are resource constrained, protecting them from the threats is more difficult than that for resourceful devices.
- Location based services (LBS) faces a privacy issue on mobile users' provide private information such as their current location.
- Problem becomes even worse if an adversary knows user's important information.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, there are several issues; security and privacy of the mobile user that is why is a major issue. So, mobile devices can expose numerous security threats like malicious code and these, GPS can cause privacy issue of the subscribers. So, there are location based services faces privacy issue of the mobile user provide private information such as their current location etcetera.

Problem become worse if an adversary knows the users important information right; and security of the mobile users approach to move threat detection capabilities on the cloud. So, there are different types of approaches.

(Refer Slide Time: 28:41)

**Mobile Cloud Computing – Challenges**

**Security for Mobile Users**

- Approaches to move the threat detection capabilities to clouds.
- Host agent runs on mobile devices to inspect the file activity on a system. If an identified file is not available in a cache of previous analyzed files, this file will be sent to the in cloud network service for verification.
- Attack detection for a smartphone is performed on a remote server in the cloud.
- The smartphone records only a minimal execution trace, and transmits it to the security server in the cloud.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES



Host agent runs the mobile devices to inspect the file activity on the system, attack detection on a Smartphone is performed on a remote server on the cloud, it can be there like if there is a connectivity between the thing. So, there can be adversely affect the mobile user. The Smartphone records only a minimum execution trace and transmit it to the security server in the cloud right. So, that is another problem. So, mobile cloud computing there are other type other challenges like context aware mobile cloud services right.

(Refer Slide Time: 29:22)

**Mobile Cloud Computing – Challenges**

**Context-aware Mobile Cloud Services**

- It is important to fulfill mobile users' satisfaction by monitoring their preferences and providing appropriate services to each of the users.
- Context-aware mobile cloud services try to utilize the local contexts (e.g., data types, network status, device environments, and user preferences) to improve the quality of service (QoS).

H. H. Lo and S. D. Kim, "A Conceptual Framework for Provisioning Context-aware Mobile Cloud Services", in Proceedings of IEEE International Conference on Cloud Computing (CLOUD), pp. 466, August 2010.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, this is context our aware services of the mobile users are there a lot of research is going on. Like it is important to fulfil mobile user satisfaction, by monitoring their preferences and providing appropriate services to each of the users right. So, different users after all is human being, has different way of looking at thing. So, I need to categorize that what sort of users are which type of things like some are maybe looking at more streaming multimedia things, some are more used to having some sort of a mail and data services and type of things right some of the users right may be running some scientific things like say data analysis type of thing.

So, this type of different user has different type of need at the back end. So, that services is based on that or based on the context need to be looked at right. So, context aware mobile cloud services try to utilize the local context, that is the data types network status device environment, user preferences to improve the quality of services right.

(Refer Slide Time: 30:43)

**Mobile Cloud Computing – Challenges**

**Network Access Management:**

- An efficient network access management not only improves link performance but also optimizes bandwidth usage

**Quality of Service:**

- How to ensure QoS is still a big issue, especially on network delay.
- CloneCloud and Cloudlets are expected to reduce the network delay.
- The idea is to clone the entire set of data and applications from the smartphone onto the cloud and to selectively execute some operations on the clones, reintegrating the results back into the smartphone

**Pricing:**

- MCC involves both mobile service provider (MSP) and cloud service provider (CSP) with different services management, customers management, methods of payment and prices.
- Business model including pricing and revenue sharing has to be carefully developed for MCC.

**IIT KHARAGPUR** | **NPTEL ONLINE CERTIFICATION COURSES**

Network access management this another challenge; like a if we see a network access management only to improve link performance and also optimizes bandwidth usage. So, what is this network latency is plays a major bottleneck. So, it is not only improving the bandwidth network giving latency, but how to optimize those uses after all these are resources which are being used by several users right. Quality of service is another aspects pricing is a tricky issue right.

So, there are number of parties now involved, it is not only the cloud provider it is a mobile service provider also. So, how this price will be there, what should be the business model and how to make use of the price, where a user will pay and what benefit is get, what is the SLAs and how things are served those are very tricky and very important issues and challenges.

(Refer Slide Time: 31:42)

**Mobile Cloud Computing – Challenges**

**Standard Interface:**

- Interoperability becomes an important issue when mobile users need to interact with the cloud.
- Compatibility among devices for web interface could be an issue.
- Standard protocol, signaling, and interface between mobile users and cloud would be required.

**Service Convergence:**

- Services will be differentiated according to the types, cost, availability and quality.
- New scheme is needed in which the mobile users can utilize multiple cloud in a unified fashion.
- Automatic discover and compose services for user.
- Sky computing is a model where resources from multiple clouds providers are leveraged to create a large scale distributed infrastructure.
- Service integration (i.e., convergence) would need to be explored.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES



There are other challenges like standard interfaces. So, that you can provide a challenge a standard interfaces or different applications, different sort of devices can work on it. Service convergence, services will be differentiated according to the type cost availability and quality. New scheme is needed in which the mobile user can utilize multiple cloud in a unified fashion. So, that is what we mean by service conversion. Service integrated that convergence would need to be explored in order to achieve all those things right. So, these are some of the important aspects of the things.

So, what we see definitely there is a need and it is a increasing need right day to day things are increasing that you want to run more stronger application or resource hungry application on the mobile devices. So, we need at the back end something a good infrastructure, which can work on behalf of my devices right. So, cloud is one of the definitely one of the option one of the good option of doing that and. So, what we required this there should be a seamless integration of this devices with the cloud service

providers and in between we have this network service provider or the who provides the mobile services and things.

So, one side mobile devices, one side cloud and intermediate this network services including internetworking, which can put all together to all of them together and so that I can have a enriched applications or more resource full applications running on devices and doing lot of other application a lot of job for variety of users.

So, with this we will conclude today.

Thank you.

**Cloud Computing**  
**Prof. Soumya Kanti Ghosh**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**

**Lecture - 32**  
**Mobile Cloud Computing – II**

Hello. We will continue our discussion on Cloud Computing; rather we will continue our discussion on Mobile Cloud Computing. So, what we are discussing about mobile cloud computing said that the mobile is a mobile devices these days are omnipresent right. There is a huge glorification of smart mobile devices which has fairly good capability of doing lot of applications or running lot of applications. As there is a as the capability increases, the need for running more computational and data intensive applications also there is a increase in the those type of applications.

Now, in though there is a increase in this resources of the mobile devices, but what we still feel that some of the cases had it been a background, backend, some more computing services that could have been better like we execute some part here, some part offload, some part of the computation or data on the backend devices where what we have seen that cloud may play a important role. But there are several issues into the things right that any transfer of data and processes involves costing. Costing in terms of execution time specifically that if the delay is more than something, it is it becomes unviable also we see that is a question of reliability when once your application runs goes to another place to run.

So, if there is there are a lot of dependency involves like intermediate network or the wireless network, the availability of the resources at the other end or failure or fault in the other end and of course, the fault in the device at the a mobile device itself. So, whereas, when you offload all those things come into play right then we have also seen there are issues of security right. So, far the application running your own device, you have someone on the process and the data over it right, but whenever the application running in some other premises or some other devices or cloud, which is maintained and owned by some third party, then your there is a concerned about privacy and other security issues the regarding those data and a processes.

So, this becomes pretty tricky issue and nevertheless mobile cloud computing is becoming pretty popular because of several reasons. First of all you can now have think of a large application running on or rather controlled by a small devices, there can be things that the your process may be using somebody elses data. So, you run somewhere where the data is there and the process is there or the mobile device acts as a collection of data which is contributing to some central repository, some process at this device or some other device can do. So, there is there are issues of in exchanging of data interoperability issue and sharing of this sort of devices right.

Or what we see a evolution of this that internet of devices come into play where cloud plays as a central role of managing the so right. So, we will look at a some of the aspects today regarding a mobile cloud computing in continuous in our with our previous lecture of mobile cloud computing, and rather we will try to refer some of the research works which will give us a opportunity or give us a more better view of how what this what are the different challenges and how to handle them and so and so forth.

(Refer Slide Time: 04:24)

**Mobile Cloud Computing (MCC) - Key challenges**

- MCC requires dynamic partitioning of an application to optimize
  - Energy saving
  - Execution time
- Requires a software (middleware) that decides at app launch which parts of the application must execute on the mobile device, and which parts must execute on cloud
  - A classic optimization problem

 IIT KHARAGPUR |  NPTEL ONLINE CERTIFICATION COURSES

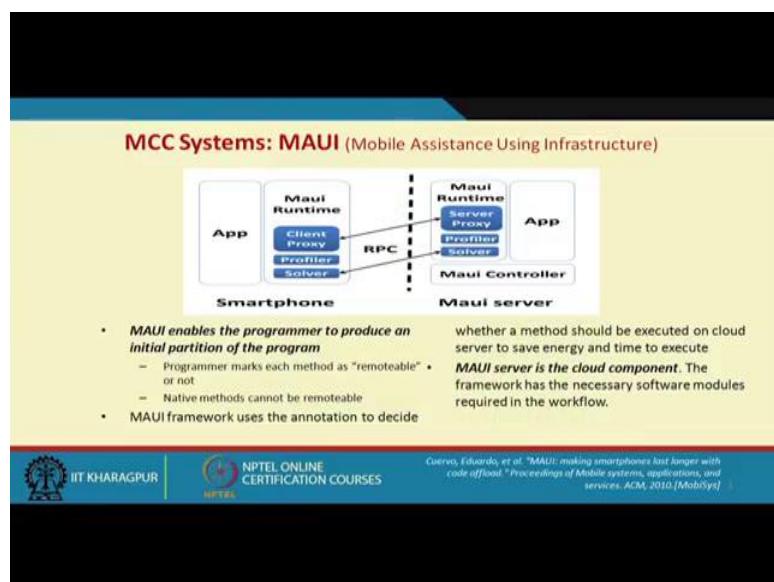
So, we will look at a mobile cloud computing, as we have seen some of the key challenges are that that requires dynamic partition of a application to optimize energy saving execution side right. So, two things as we mentioned earlier also, one is energy saving and to optimize the execution time right. If the time is beyond a limit then there may be that then that process may not be viable to run on the things also requires

software or something middleware that decides that the app launch which parts of the application will be executed locally and which part of the application need to be executed on the externally or remotely or in the cloud right.

So, it is a classic optimization problem right. So, we can do this sort of means dividing this application into different parts, some of the parts or the component of the application need to be run locally and some of the remotely, this can be done statically like I have a priori knowledge of this sort of things can be done or in some cases I may not know the a priori that which need to be locally or locally or remotely, primarily it is if it is dependent on the data and other type of things right. So, what will happen I need to have a dynamic these and all the things.

And as we understand whenever we need this type of thing synchronize a sense scheduling and the resource management come into play in a big way right. So, all those things here and when you are specifically doing the devices and say infrastructure, which are owned and maintained by different authorities then it is a more tricky. So, like mobile devices is something and cloud is maintained by other third party, intermediate you have service provider we are who are again have a different type of authority. So, all those things makes a very complicated issue need to be solved.

(Refer Slide Time: 06:36)



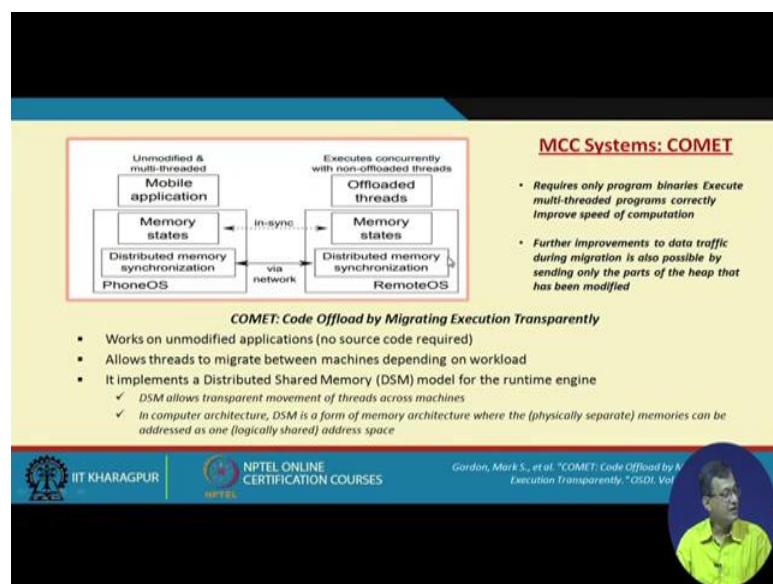
So, this is one of the work is as referred in the in this paper, there is mobile assistance using infrastructure right. So, it is one form of mcc. So, MAUI enables the programmer

to produce an initial partition of the program. So, programmer marks the each method as remoteable or not that means, whether a remote or local, native methods cannot be remoteable which are need to be done on the on the device itself MAU framework uses the annotate to decide whether the method should be executed on cloud server to save energy and time to execute the thing right.

So, there are a couple of times the time to execute on the cloud there is a time to transfer this process and data to the cloud and get back the data if as if the need demands and MAU server is a cloud component. So, there is a MAU client which is on the mobile side there is a MAU server which is a cloud component, the framework has necessary software modules required for the workflow. So, that there is a client server type of architecture where what they propose and can be able to synchronize.

So, if you see; there is the app and there is a profiler and the solver profiler basically profiling of the application and profile along with the server partitions the application these gates executed here, and with RPC calls and then it in a two and four things Lan server communicate to the profile as sync among itself when the apps run in a some part on the app and some part on the some part on the device and some or what we say smart mobile divider, SMD and some part on the cloud right.

(Refer Slide Time: 08:27)



So, there is another work which is what is called comet quad, code offloading by migrating execution transparently right. So, works on unmodified application no source

code required that is on the app itself. Allows threads to migrate between the machine depending on workload, they based on the things that threads can migrate, it implement a distributed memory shared memory or DSM model to runtime engine right those who have gone through computer architecture or advanced computer architecture and various type of things, they understand that this DSM model. A DSM allows transparently movement of the threads across machine. In computer architecture DSM is a form of memory architecture where physically separate memories can address as one logically shared address space, there is the core of the or there is what we say very quick view on the what this DSM says.

So, likewise there is a phone OS and remote OS. So, it need to maintain this memory states and there is a distributed memory synchronization and the executed this processes other codes are executed or offloaded by migrating execution transparently. So, it is not that the neither the source code is required nor that has nor the a user end there is known that this things are going on at the backend. So, what it requires only program binaries right execute multi thread programs correctly. So, it does not require source code only the binaries; furthermore improvement on the data traffic during migration is also possible by sending only parts of the heap that has been modified.

(Refer Slide Time: 10:26)

**Key Problems to Solve**

- At its core, MCC framework must solve how to partition a program for execution on heterogeneous computing resources
- This is a classic "Task Partitioning Problem"
- Widely studied in processor resource scheduling as "job scheduling problem"

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES



So, it is not transmitting the whole heap, but the part of the heap which has been modified. So, what we see the key problem to address or solve is a at the at its core MCC

framework, must solve how to partition a program right or execution on heterogeneous sources. Thus not only partitioning, it need to be executed on heterogeneous resources right it one on the mobile device side which may be some mobile related OS when you are running on the cloud side that is some other kind of environment. So, it is a heterogeneous environment how things need to be run need to be looked into.

So, it is a task partitioning problem as we have seen in other computer architecture and organizations, widely studied in processor scheduling as a job scheduling problem also right we do have this sort of challenges out here. So, need to be looked into from that angle. So, task partitioning problems. So, what we have typically in a task partitioning problem in MCC a call graph representing the applications method or call sequence right.

(Refer Slide Time: 11:22)

**Task Partitioning Problem in MCC**

**Input:**

- A call graph representing an application's method call sequence
- Attributes for each node in the graph denotes
  - (a) energy consumed to execute the method on the mobile device,
  - (b) energy consumed to transfer the program states to a remote server

**Output:**

- Partition the methods into two sets – one set marks the methods to execute on the mobile device, and the second set marks the methods to execute on cloud

**Goals and Constraints:**

1. Energy consumed must be minimized
2. There is a limit on the execution time of the application
3. Other constraints could be – some methods must be executed on mobile device, total monetary cost, etc.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, it is a call graph, attributes of each node in the graph in denotes a energy consumed to execute the method on the mobile device and energy consumed to transfer the program states to the remote server. So, one is the energy consumed in the mobile device and transferred the things.

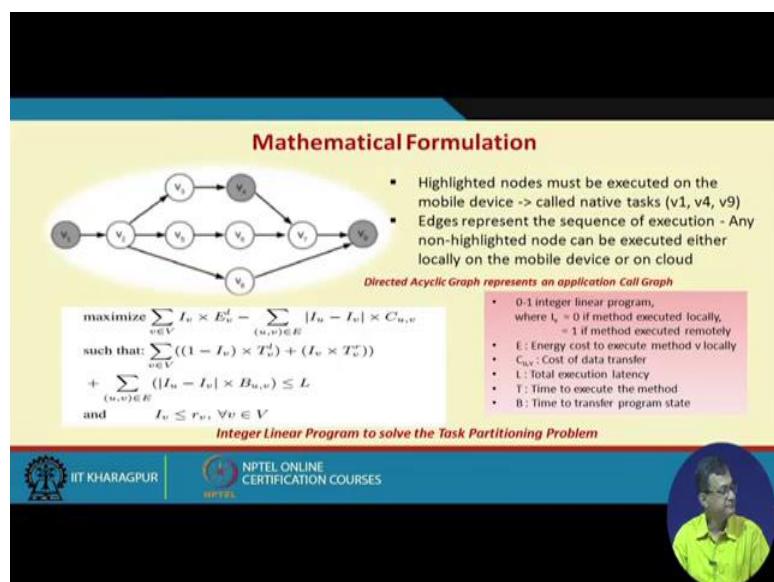
What we are considering that the other end that is the cloud end, the it is pretty resourceful and efficient. So, the energy consumes in is in executing that in the cloud end is negligible in compared to the energy consumption at this end at the device end. Rather the device does not much clear about that, but you need to clear about that executing its own locally the portion to be executed, and the transfer cost of this the same memory

state and processes and so and so forth as things required right and data memory state and so on. So, output what we get as a partition the methods into two sets one set of the matter executes at the mobile device, and the second state set it goes to the execute on the cloud.

So, go if we look at the goal and constraint, energy consume must be minimized. So, I need to minimize the energy consumption, there is a limit on the execution time of the application. So, it should within the threshold say time capital T, other constant could be some methods must executed on the mobile device total monetary cost etcetera. So, two we have couple of a constant. That means, we need to goal is to reduce the energy consumption, time should be within that particular threshold value. And there can be other constant like some of the application some part of the application may need to be executed in the device itself like I may not be able to offload it.

And I may look at the total costing of the thing the total cost should not execute some something some portion more than that, so these things are necessary.

(Refer Slide Time: 13:54)



So, if we look at that little mathematically. So, we have a execution graph right out of that what we say that these are v1, v4, v9 in this case are native task; that means, needs to be executed in the device itself right I cannot offload it. So, there can be so, highlighted nodes here must be executed device, edges represent the sequence of execution that how the sequence of execution will be there. Any non highlighted node

can be executed either locally or on the mobile device right can be locally or on the mobile device.

So, if it is offloaded to the local device, this first four senses that how much energy it could have saved right or in other sense had it been worked on the local how much energy could have been taken. So, that has been saved if I offload it to the remote, but I need to also look at that where the costing of transfer this data right. So, there is a cost involved or there is a energy cost involved energy expenditure involved in transferring the data from the device to the remote device right that has to be incurred by the mobile device right. So, that are needs to be there.

So, if we need to maximize this. So, if it is if this result is a positive or the maximum of the maximum if you can maximize, then I have a substantial energy saved otherwise cannot be there the there is no benefit of doing that right and of course, it should be within a total execution latency should be less than L right we should not have this whatever we do need to be less than the particular latency L right. So, that it should be within that particular things and I need to look at that a dependency also like execution of one need to be executed, if there is a dependency in the graph that need to be executed out here.

(Refer Slide Time: 16:13)

**Static and Dynamic Partitioning**

- **Static Partitioning**
  - When an application is launched, invoke an ILP solver which will tell where each method should be executed
  - There are also heuristics to find solutions faster
- **Dynamic or Adaptive Partitioning**
  - For a long running program, the environmental conditions can vary
  - Depending on the input, the energy consumption of a method can vary

**IIT KHARAGPUR** | **NPTEL ONLINE CERTIFICATION COURSES**

So, as we are discussing there are two type of partitioning one is static and dynamic as we have seen. Static in case of a static when the application is launched the invoke a ILP

solver and which tells where each method should be executed. So, it says that statically says that when it is launched. So, there is there are also heuristic to find a solution first term that there are we can deploy. In case of a dynamic or adaptive partitioning, what we do for long running program environmental conditions may vary right if it is executing for a long time environmental condition may vary like a requirement of the data and even the availability of the remote resources transferring bandwidth all those things can vary.

Depending on the input the energy consumption of a method may vary so; that means, in other sense I need to take a call that which partition need to be executed locally and remotely because of that varying nature of the things there are hosting may go on changing right. So, that is another challenge out here.

(Refer Slide Time: 17:26)

**Mobile Cloud Computing – Challenges/ Issues**

**Mobile communication issues**

- *Low bandwidth:* One of the biggest issues, because the radio resource for wireless networks is much more scarce than wired networks
- *Service availability:* Mobile users may not be able to connect to the cloud to obtain a service due to traffic congestion, network failures, mobile signal strength problems
- *Heterogeneity:* Handling wireless connectivity with highly heterogeneous networks to satisfy MCC requirements (always-on connectivity, on-demand scalability, energy efficiency) is a difficult problem

**Computing issues (Computation offloading)**

- One of the main features of MCC
- Offloading is not always effective in saving energy
- It is critical to determine whether to offload and which portions of the service codes to offload

 IIT KHARAGPUR |  NPTEL ONLINE CERTIFICATION COURSES



So, there are as we see there are when we do this type of partitioning and transfer the job executes something locally, something remotely or everything within a time limit, a particular threshold values, threshold time and we want to maximize the energy saving right.

So, these are the different constraint and in order to achieve this sort of issues. So, there are several challenges and issues need to be addressed. So, mobile communication issues that is low bandwidth one of the biggest issue, because of the radio resource of the wireless network is much more scarce than the wired networks. So, the it is the availability band availability of this bandwidth mobile bandwidth or wireless bandwidth

at times is pretty less in compared to a wired type of things or. So, that is not only the less it also varies on if it is a long running execute executive means long running process then it may vary over time.

So, getting a guaranteed bandwidth is a challenge service availability mobile user may not be able to connect to the cloud to obtain a service due to traffic congestion network failure mobile signal strength and so and so forth. Issue of a heterogeneity handling wireless connectivity with highly heterogeneous network to satisfy mcc requirement always on a connectivity on demand scalability, energy efficiency is a difficult problem. So, there are heterogeneous heterogeneity involved in it and there are these are different challenges. So, these are more issues with the communication issues, there are a few computation issues or computing offloading related issues; one of the this is one of the major feature of the things like you are able to offload it.

Offloading it not always effective in saving energy as we have seen that if those parameters are not made then it may not be that energy saving, may not be what we say within that quote unquote profitable margin right. So, it may be costly stuff; it is critical to determine whether to offload and which portion of the source code to offload. So, its it is a critical challenge, especially as that there are different type of application with different type of requirement etcetera and its it is a major challenge to decide that where when to offload, how much to offload and so on and so forth.

(Refer Slide Time: 20:11)

**CODE OFFLOADING USING CLOUDLET**

- **CLOUDLET:**
  - ✓ "a trusted, resource-rich computer or cluster of computers that is well-connected to the Internet and is available for use by nearby mobile devices."
- **Code Offloading :**
  - ✓ Offloading the code to the remote server and executing it.
  - ✓ This architecture decreases latency by using a single-hop network and potentially lowers battery consumption by using Wi-Fi or short-range radio instead of broadband wireless which typically consumes more energy.

↳

 IIT KHARAGPUR |  NPTEL ONLINE CERTIFICATION COURSES

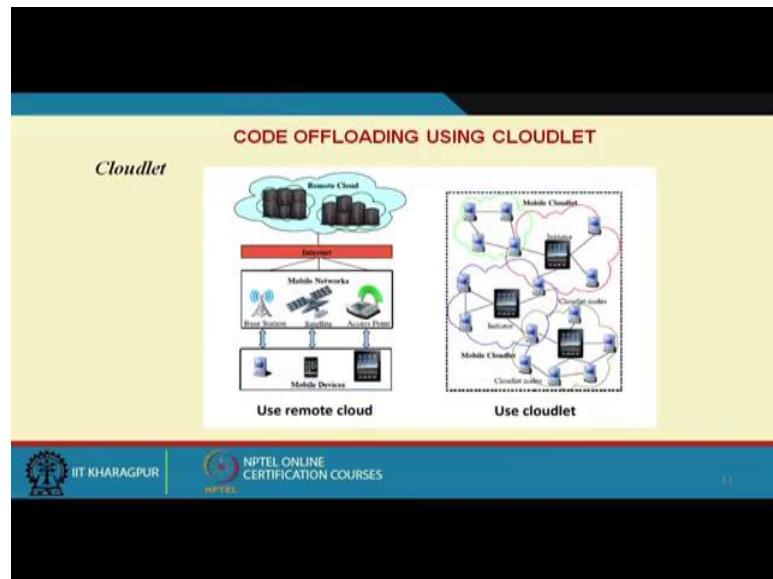


So, there are other trials or other type of a initiative one is that code offloading using cloudlet right. So, if you look at cloudlet, it is a trusted resource rich computer or cluster computers that is well connected to the internet and is available for use by nearby mobile devices. So, it is what we say we can look at as a smaller version of cloud, but it is resource rich, but it is more within the locality of the mobile devices or nearby regions right.

So, offloading a code to the remote server and executing it, when you look at the court code offloading part. So, this architecture decreases latency by using single hop network and potentially lowers battery consumption using Wi-Fi, short range radio broadband wireless etcetera. So, what we try to see that first of all of offloading and secondly, when we make multiple hops then the cost increases. So, what you try to see that whether I can have presence of smaller clouds, which can take care of this sort of application, this sort of a executing this portion of the application.

And if a single hop then we can have something more or what we say that we can have more the better energy saving and better response time, with having those resources much closer to the device.

(Refer Slide Time: 21:53)



So, cloud code offloading using cloudlet. So, we have a mobile devices, mobile networks and the remote cloud and there are if you look at there are different small small different set of systems or mobile cloudlet, and cloudlet nodes which takes care of these devices

right or what we say that goal is to reduce the latency, in reaching the cloud service servers use a servers that are closer to the mobile devices.

(Refer Slide Time: 22:21)

CODE OFFLOADING USING CLOUDLET

*Cloudlet*

- Goal is to reduce the latency in reaching the cloud servers Use servers that are closer to the mobile devices → use cloudlet
- A cloudlet is a new architectural element that arises from the convergence of mobile computing and cloud computing.
- It represents the middle tier of a 3-tier hierarchy  
*mobile device --- cloudlet --- cloud*

Use remote cloud      Use cloudlet

IIT KHARAGPUR      NPTEL ONLINE CERTIFICATION COURSES

That means going for instead of the full place cloud, we can have smaller version of the clouds or cloudlet. A cloudlet is a new architectural element or that arises from the convergence of mobile computing and cloud computing right.

It represents a middle tire of a 3 tier architecture like mobile devices cloudlet and cloud. So, that in other sense what we are trying to do in order to reduce the latency, and a better energy savings possible better energy saving in one side the cloud is there, one side a devices we made a we make a intermediate a thing what we say cloudlet right; so which takes care of the things. Now when to offload is again a serious question.

(Refer Slide Time: 23:17)

**When to Offload ?**

Amount of energy saved is :  
$$P_c \times \frac{C}{M} - P_i \times \frac{C}{S} - P_{tr} \times \frac{D}{B}$$

S: Speed of cloud to compute C instructions  
M: Speed of mobile to compute C instructions  
D: Data need to transmit  
B: Bandwidth of the wireless Internet  
 $P_c$ : Energy cost per second when the mobile phone is doing computing  
 $P_i$ : Energy cost per second when the mobile phone is idle.  
 $P_{tr}$ : Energy cost per second when the mobile is transmitting the data.

Suppose the server is F times faster—  
 $S = F \times M$ .

We can rewrite the formula as  
$$\frac{C}{M} \times (P_c - \frac{P_i}{F}) - P_{tr} \times \frac{D}{B}$$

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, if we look at a very straightforward formulation of the things. So, if we look at that equation that. So, what we say that  $P_c$  is the energy cost per second when the mobile phone is doing computing; that means, mobile phone is computing  $P_i$  is that energy cause the mobile phone is idle.

$P_{tr}$  is the per second a cost per second when the mobile phone is transmitting data. So, a it a mobile device can be either computing or idle or transmitting data. So, this three component are  $P_c$ ,  $P_i$  or  $P_{tr}$  or when energy when transmitting data. Now if you consider c as the number of instruction, and m is the speed of the mobile compute the c instruction. So, what we see that if we see that  $P_c \times \frac{C}{M}$  if it is offloaded  $P_c \times \frac{C}{M} - P_i \frac{C}{S}$ .

So, what is s the speed of cloud to compute the C instruction.

So, if it is s is the p Ps s is the speed of the cloud then I we can see that idle  $P_i$  is the idle time, C there is number of instruction by the speed. So, this P by C by S and the transmitting P trd by B D means the total data to transmit and B is the bandwidth available. So, if we offload. So, this is my overall Ballpark figure right this should be a positive number. So, it says that I save energy right. So, this was my energy consumption if I offload this consumption are not there and then then this is this is save of the saving of the energy or if I say the server or the cloud server is f time faster, then I can look at  $S = F \times M$  right.

So; that means, your speed of cloud to compute C instruction and M is the speed of mobile phone to compute C instruction if it is f times faster, and we can see that S equal to F into M right or in other we can rewrite the program or rewrite that particular formula or substitute this s equal to F by M into the formula to look to basically arrive this thing

$\frac{C}{M} \times \left( P_c - \frac{P_t}{F} \right) - P_{tr} \times \frac{D}{B}$ . This is a very simple arithmetic you just substitute and if you

take appropriate common. So, we come to a figure called in this sort of scenario right

$$\frac{C}{M} = \left( P_c - \frac{P_i}{F} \right) - P_{tr}.$$

(Refer Slide Time: 26:40)

**When to Offload? (contd..)**

- Energy is saved when the formula produces a positive number. The formula is positive if D/B is sufficiently small compared with C/M and F is sufficiently large.
- Cloud computing can potentially save energy for mobile users.
- Not all applications are energy efficient when migrated to the cloud.
- Cloud computing services would be significantly different from cloud services for desktops because they must offer energy savings.
- The services should consider the energy overhead for privacy, security, reliability, and data communication before offloading.

The amount of energy saved is :  
 $P_c \times \frac{C}{M} - P_i \times \frac{C}{S} - P_{tr} \times \frac{D}{B}$

We can rewrite the formula as  
 $\frac{C}{M} \times (P_c - \frac{P_i}{F}) - P_{tr} \times \frac{D}{B}$

**IIT KHARAGPUR** | **NPTEL ONLINE CERTIFICATION COURSES**

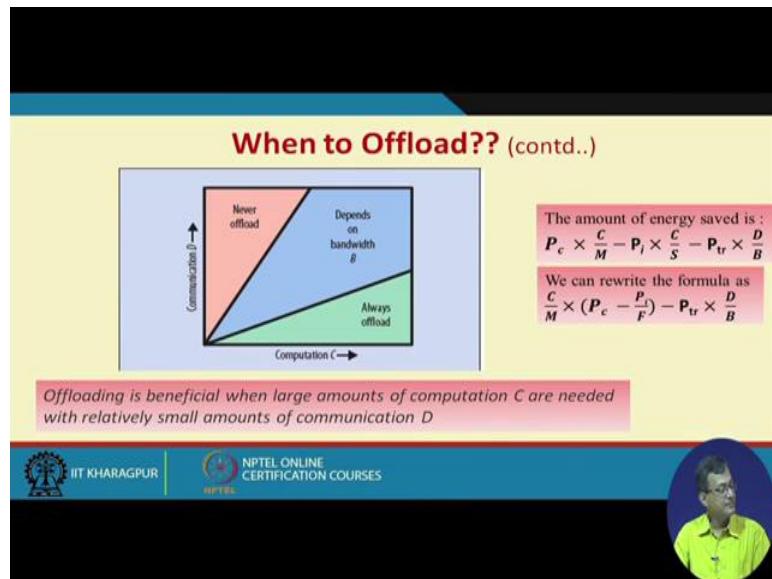
So, what we see from here that energy is saved when formula produces a positive number definitely if it is a positive then only energy is saved the formula is positive if D by A, D by B is sufficiently small compared to C by M right. So, one side is D by B should be sufficiently small compared to C by M. So, that I have a larger number on the left side. So, this is what we say that this is one of the condition cloud computing. So, what cloud computing can potentially save energy from the mobile devices, note that all application in the energy efficient when migrated to the cloud are may not be energy efficient right.

So, from this formula already say that first of all D by B should be must lesser less than C by M and F should be sufficiently higher which makes sense right. F is the how many times the cloud server is faster than the mobile device; it is likely that the cloud server

should be sufficiently higher otherwise there is no point in offloading and getting executed at externally right if it is you know. So, if that conditions are satisfied then what we see that we get a positive number, and it is it is obvious that all the applications may not be always beneficial when you offload, cloud computing service should be significantly different from the cloud services for desktop because they must offer energy savings in this case.

The services should consider the energy overhead for privacy, security, reliability, data communication etcetera which we have not directly considered in these cases.

(Refer Slide Time: 28:37)



So, it is ideally it should take care of other type of features which in while calculating the things. So, if we look into different other way of looking at us same a formulae formula is that offloading is beneficial when large amount of computation C are needed with relatively small amount of data communication right.

So, we require large amount of computation right with relatively less transmission of data right. Here also as we have seen that C should be sufficiently large with D should be sufficiently small. Keeping like F must be much faster and do we see that it is a thinks and we can see that in different scenario how things will be there right.

(Refer Slide Time: 29:24)

**Computation Offloading Approaches**

- Partition a program based on estimation of energy consumption before execution
- Optimal program partitioning for offloading is dynamically calculated based on the trade-off between the communication and computation costs at run time.
- Offloading scheme based on profiling information about computation time and data sharing at the level of procedure calls.
  - A cost graph is constructed and a branch-and-bound algorithm is applied to minimize the total energy consumption of computation and the total data communication cost.

Z. Li, C. Wang, and R. Xu, "Computation offloading to save energy on handheld devices: a partition scheme," in Proc. 2001 Int'l Conf on Compilers, architecture, and synthesis for embedded systems (CASES), pp. 238-246, Nov 2001.  
K. Kumar and Y. Lu, "Cloud Computing for Mobile Users: Can Offloading Computation Save Energy," IEEE Computer, vol. 43, no. 4, April 2010

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, partly computation overloading there are several approaches, you can see several research papers also partitioning a program based on estimation of energy consumption before execution. So, what we say that pre partition estimation of the thing, optimal program partitioning for offloading is dynamically calculated based on trade off between communication and computational cost at the at run time.

Offloading scheme based on profiling information about computation time data sharing at the level of procedure calls right. So, at the level of procedure calls we can do. So, a cost graph can be generated or constructed and branch and bound algorithm to applied to minimize the total energy consumption of computation and total data requirement. So, what we are looking at that it is not vanilla offloading some of the things to the things. So, there are a lot of calculation involved as we come back to the our previous premise that one is that your energy savings should be higher, need to be maximized whereas, your latency should be preserved.

And we see that there are several other aspects of the components are there like security, reliability, fault in the server system etcetera and there are issues which are not directly under not and may not be a directly under the control of the cloud, neither on the mobile devices which is what we say that intermediate wireless network that also plays a things. So, that is there are issues of heterogeneity need to be addressed and need to be looked into a in a totality it is not that isolated I offload something, get some executed and type

of things need to be looked at in totalities. So, that the user at the mobile end or the user get a seamless feeling or better resource utilization or better energy savings at that things.

Thank you.

**Cloud Computing**  
**Prof. Soumya Kanti Ghosh**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**

**Lecture - 33**  
**Fog Computing – I**

Hello. So, we will be continuing our discussion on cloud computing. So, what we have seen that in case of cloud computing, what we are trying to do we are trying to offload our computing and computing processes and data to the cloud right. So, that it is maintained by a third party and the on the other end the customer or the consumer or the user more concentrate on the business processes process.

So, that is the basic objective of or model of the things right. There are a lot of technical technicalities at the backend, but nevertheless we are offloading the thing. So, what for that what we need a very strong backbone right or a strong backbone network which should be always up and able to transfer data on a large volume. As we see as along with the development and being most of the things are digitally enabled we are what we are getting a huge volume of data.

In other sense a huge volume of data maybe need to be transmitted from this customer end or consumer end to this cloud service provider being executed the results in some cases are transmitted back or transmitted in other places right. The major issue is this huge volume or transfer of data and what we have what we see in recent development with number of activities specially internet of things, coming up and more only and huge variety of sensors in place.

So, we have lot of multimedia data which need to be transmitted right. And that is one part of the story that we require a huge backbone and type of things as for as the cloud is concerned what we consider it is a huge computing power much higher than what we what the devices can do and it is some sort of infinite computing power is there.

On the other hand what we see that a huge volume of data are being generated and being transmitted over the network. Again what we see that the devices starting from as we discussed about mobile devices, smart mobile devices or other type of devices even

intermediate network devices, what they are becoming is more powerful in terms of computation and more resourceful things ok.

Or in other sense we are not all the times exploiting the resources available in the things sic say consider a particular sense sensor node or a local sync node of a sensor which are collecting the data and transmitting to the in the upward path maybe to the cloud. So, this this could have been done some processing at the end like I can say that suppose in this particular room or a particular lab, I have say 10 temperature sensors right.

So, what is my basically business model? This temperature should be varying may between say 18 to 22 degree centigrade that is the operating range of this temperature.

Now, what we are doing this all these 10 or sensors, are sending the data to the up in the server maybe in a cloud that is calculating, whether the within the limit and type of things and I can have the say 10 such labs. So, there are 100 such data are going on and if the temperature is varying somewhere it is sending error.

Now, if you consider this particular a enclosed lab or single things otherwise I have could taken, I taken a local ds and whether my temperature is in the lab by the sync node of the sensors which are collecting this data of this particular room and it takes a call that whether higher or upper, and then sense that say it is some statistical data or what we say some aggregated data to the sensor, it may be the average or it may be average with other standard deviation etcetera to the things.

In other sense it is not in other than sensing this transmitting this 10 sensors data I am sending one average data or and which has my purpose. Even you can say that the if the if my sync node is intelligent enough, it can take a call that whether the temperature is within this operating range yes or outside thing some 01 or yes no type of things and then transmit this.

In other sense this is taking a some part of computing at the things. So, with the intermediate devices becoming more intelligent, whether there is a possibility of pushing the computing from logically centralized cloud to somewhere more down the line right towards the edge of the things right, that is exactly what we are trying to discuss today is what we say this sort of computing is fog or from cloud to fog right. So, cloud is the whole thing and fog computing.

(Refer Slide Time: 05:46)

### Cloud Computing : Challenges

- Processing of huge data in a datacenter.
- Datacenter may be privately hosted by the organization (private cloud setup) or publicly available by paying rent (public cloud).
- All the necessary information has to be uploaded to the cloud for processing and extracting knowledge from it.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, as we see the challenges or the data what the cloud computing todays is doing the processing of huge data in the datacenters, may be privately hosted or publicly available by paying rent that is it can be a public cloud or a private cloud, all necessary information has to be uploaded or transmitted to the cloud for processing and extracting knowledge of it right.

(Refer Slide Time: 06:20)

### Cloud Computing – Typical Characteristics

- **Dynamic scalability:** Application can handle increasing load by getting more resources.
- **No Infrastructure Management by User:** Infrastructure is managed by cloud provider, not by end-user or application developer.
- **Metered Service:** Pay-as-you-go model. No capital expenditure for public cloud.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, the whole data as we are discussing need to be transmitted to the cloud.

Now, also we have seen the typical characteristics of cloud for which we are the today's world is inclined towards is that dynamic scalability, I can scale up or scale down based on my need. So, another is that no infrastructure management or practically minimal infrastructure management at the user end. So, if I offload everything that computing etcetera on the cloud. So, I require very less infrastructure management at my user end and secondly, and finally, what we have a metered service right pay as you go model.

So, these three things that dynamic scalability minimal management or all my infrastructure management pushing it to the cloud and metered service pay as you go model these are primary features of the cloud which makes it popular there are several other things which are there, but never the less these are the three things which are the driving force. So, whatever we do we do not want to lose out of the things if we compromise on those type of features then the very motivation to going towards cloud may be challenged.

(Refer Slide Time: 07:38)

**Issues with “Cloud-only” Computing**

- Communication takes a long time due to human-smartphone interaction.
- Datacenters are centralized, so all the data from different regions can cause congestion in core network.
- Such a task requires very low response time, to prevent further crashes or traffic jam.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Now, there are issues with cloud only computing. So, what we say that only the cloud is computing register sitting duck or maybe the issues, especially in today's applications which variety of sensors, variety of real time operations and lot of redundant data right there are lot of data which are redundant like if I am sending temperature things it may not mean may be meaningless to sense all the sensors data, which are more or less same information unless there is a different in some sensor data I may not want to send the

data all are reporting between around 20 degree centigrade, it does not require a cloud to take a call it could have been done as a much lower level. So, that or in other sense I have a huge amount of digital data to be transmitted.

So, communication takes place takes a long time due to hum human smart for interaction and type of things, if the state still datacenters are centralized right datacenters and in some portion. So, all the data from different region can cause congestion in the core right. So, being transmitted things especially in case of exigencies where a lot of volumes of data suddenly pushed into the thing right in case of say some disaster or some huge amount of in flux due to some event, this is a lot of volume of data suddenly in flux. So, there is a huge volume of data to be transmitted and there can be congestion and such a task requires very low response time to prevent further crashes etcetera.

So, if I have this sort of things which has a some sort of accident, some accident prevention mechanisms can into should be activated. So, where we require a very low response time, immediately need to be act acted. So, waiting for that cloud to take a call revert back and all those things may take lot of time. So, that is another problem.

(Refer Slide Time: 09:40)

**Fog Computing**

- Fog computing, also known as fogging/edge computing, it is a model in which data, processing and applications are concentrated in devices at the network edge rather than existing almost entirely in the cloud.
- The term "Fog Computing" was introduced by the Cisco Systems as new model to ease wireless data transfer to distributed devices in the Internet of Things (IoT) network paradigm
- CISCO's vision of fog computing is to enable applications on billions of connected devices to run directly at the network edge.
  - Users can develop, manage and run software applications on Cisco framework of networked devices, including hardened routers and switches.
  - Cisco brings the open source Linux and network operating system together in a single networked device

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, the emergence of a concept called fog computing. So, on the cloud we are to the we are talking about fog that is little bit bringing tao down to the ground or in other sense we are pushing this computing thing from the centralized datacenter or the cloud datacenters to this edges right or intermediate or the edges of the network edge of the network.

So, fog computing also known as fogging and edge computing though some people have little other views of that edge computing, but nevertheless it is a fogging or edge computing it is a model in which data process applications are concentrated in devices at the network edge rather than existing almost entirely on the cloud. So, now, not only the cloud at the centralized things the data application and processes are distributed between the edge right which comes effect of some way of distributing this whole processing whether things. What it helps us? It helps us in reducing the data load in the communication, I can have a local decision and which is not needed for the global type of things they say smart traffic light management system.

The traffic light management system in Kolkata is nothing to do with the traffic light management, this system in Delhi apparently right for day to day traffic management right. So, I could have done it locally or even I can say that a region of a particular city may have only aggregated data which need to be transmitted at the higher level for traffic management right. So, that basic intermediate management could be done locally.

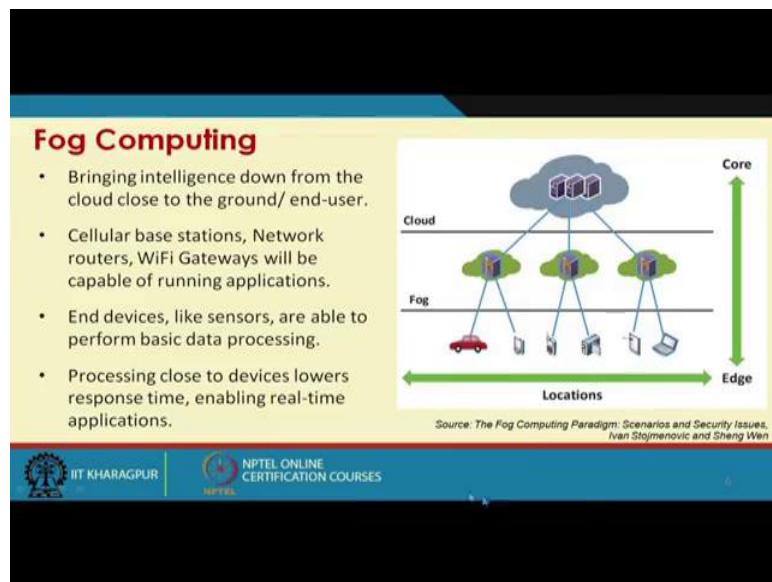
So, those things could be done in a concept of what we say fogging or fog computing. The term fog computing was first introduced by CISCO as a new model to ease wireless data transfer to distributed devices in the internet of things network paradigm. So, as IoT is becoming omnipresent or IoT is becoming everywhere it is there internet of things. So, it is huge volume of data devices which mass computing capability or resources much higher resources can do a bit of a job which could have been solved at a lower level.

So, CISCO'S vision if we look at that fog computing is to enable application, on billions of connected devices to run directly on the network edge; since CISCO is primarily a network driven organization. So, it has a huge number of devices across the world and those devices are somewhat managed etcetera managed by a some sort of a homogeneity is therefore, because upon the one make and there are resourceful devices which could have done some sort of computing things, and I can even run applications on the devices and doing so on and so forth right.

So, user can develop manage run software application of CISCO framework of network devices, including hardened routers switches etcetera. CISCO brings say open source Linux and network operating system together in single network devices. So, it

helped to do not only computing, but it was if you want to do computing you need to give some sort of a platform to run the applications for the computing things right. So, those things are they are in the devices and this is possible because of there are resources available at different layer of the network towards the edge.

(Refer Slide Time: 13:39)



So, if we look at a view. So, this cloud are at the top it is still there and it should be there, are intermediate devices which were now helping only a so far was only transmitting the data, now can they do some sort of a computing what we say fogs fog computing and there are end user devices which are spread over different locations, starting form say smart vehicles or which can communicate, devices servers smart cameras and anything which can do write any device which can capture detailed data compute and transmit right.

So, bringing intelligence down from cloud closer to the end user of the edge of the network that is one of the things. Cellular base station network routers Wi-Fi gateways will be capable of running these applications right. So, there are because whenever I communication we have cellular networks, Wi-Fi router into place and if those are having surplus resources and they are able to do that. So, that a my application can run say I want to run a application for monitoring the environment of different labs, starting from temperature to humidity may be some sort of a what sort of air pollution or air content etcetera.

So, this sort of things can be done; end devices like sensors are able to perform basic data processing right. So, the sensors can do a basic data processing. Processing close to the devices lowers the response time enabling real time applications right. So, whenever we process close to the devices. So, the response time reduces that is obvious and I can do lot of real time processing of the things right. So, I can do a real time processing of a of a say of any applications like I do a application based on that what we say dynamic signalling mechanism of a traffic light based on the traffic on the road.

So, the cameras which are on the road capturing that how many what is the traffic flow based on that I the traffic light signalling may change if that is the need of this traffic management. So, that is local right local to a particular portion, local to a region, local to a city right. So, that definition of locality may vary from application to application, but what we require that your devices like that traffic light device etcetera should be able to run this application which can take a call right. So, those are things nevertheless this is this is about the fog.

(Refer Slide Time: 16:33)

**Fog Computing**

- Fog computing enables some of transactions and resources at the edge of the cloud, rather than establishing channels for cloud storage and utilization.
- Fog computing reduces the need for bandwidth by not sending every bit of information over cloud channels, and instead aggregating it at certain access points.
- This kind of distributed strategy, may help in lowering cost and improve efficiencies.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, if we look at fog computing enable some transactions and resources at the edge of the cloud, rather than establishing channels for the cloud storage and you utilization. So, rather than just transmitting, it do some sort of a transaction processing or application running on the things; fog computing reserves reduces the need of bandwidth by not

sending every bit of information to the cloud channels over the cloud channel, instead aggregating at a certain axis point.

So, it aggregates and send the aggregate data this kind of distributed strategy may help in lowering cost and improve efficiency. So, this sort of it is a distributed strategy and this type of distributed phenomena may help us in lowering the overall cost not only in terms of monetary, if the cost of transmission in terms of time etcetera and I can we can do efficiency right I can run several applications which can be real time and type of things.

So, this motivation is obvious already whatever we have discussed the motivation, the fog computing a paradigm that extends cloud and its services to the edge of the network. Fog provides data, compute, storage, application services to the end user if you see it says some sort of a small form of a instance of the cloud for that local type of things right. So, it is doing some sort of a computing or giving some sort of a cloud service at that time at that portion of that region recent. And we and there is another side of the things because we have several series of developments one is the smart grid other is the smart traffic lighting in cities, specially cities connected vehicles or strong regular networks which is coming up and also the software defined network.

(Refer Slide Time: 18:29)

**Fog Computing - Motivation**

- Fog Computing is a paradigm that extends Cloud and its services to the edge of the network
- Fog provides data, compute, storage and application services to the end-user
- Recent developments: Smart Grid, Smart Traffic light, Connected Vehicles, Software defined network

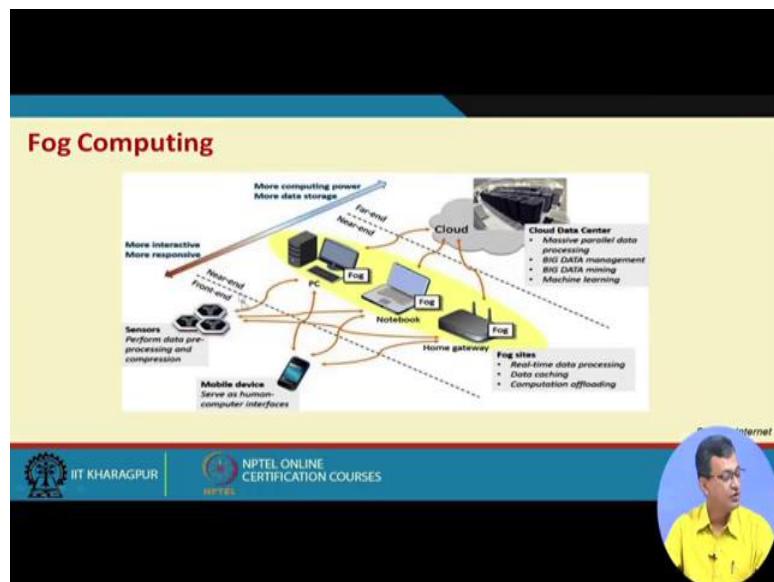
IT Kharagpur | NPTEL ONLINE CERTIFICATION COURSES

So, these are the different aspects which are itself is a topic to work at, but the smart grid smart traffic lighting, smart vehicles, as software defined network and so on and so forth they are becoming pretty popular, and in turn they generate huge volume of data right

everyone is generating huge volume of data, which are being transmitted at the higher up in the layer for doing that.

So, all this different aspects has motivated or what it has pushed the push the processing towards doing it at the edges or intermediate layer rather than pushing everything to the cloud. So, this is what we look at the fog.

(Refer Slide Time: 19:32)



So, this is the same thing what we discussed so, we have one in this cloud. So, which has a datacenter with huge capability massive parallel data processing, big demand, big data mining machine learning algorithms etcetera, which is they are and should be their; intermediate layer which is more near to this edge or the devices. So, they are can act as a fog. So, they are they can be there can be fog sides with real time data processing data caching computation of offloading and those type of things.

So, these are not. So, powerful at that, but as such they are intermediate devices we are which are used for transmitting data. So, that these are this can be used at the end or the at the front end or the edge or the last mind what we say, what we have the sensors we are connecting different type of data, perform data pre-processing and compression, mobile device serve as a human computer interfaces like this these are the different type of things which are transmitting out here and in turn transmitting to the things.

So, these some sort of communication yes if we see that both way arrow, can be taken a call at this end itself right without transmitting the whole data at the things. It may be some sort of aggregated reporting and type of things or aggregating the data and taking putting it to the cloud for running some intelligent algorithm and machine learning based algorithm and type of things.

So, we have more interactive and more responsive end, to more computing power and more storage end at the other end.

(Refer Slide Time: 21:34)

**Fog Computing Enablers**

- **Virtualization**: Virtual machines can be used in edge devices.
- **Containers**: Reduces the overhead of resource management by using light-weight virtualizations. Example: *Docker* containers.
- **Service Oriented Architecture**: Service-oriented architecture (SOA) is a style of software design where services are provided to the other components by application components, through a communication protocol over a network.
- **Software Defined Networking**: Software defined networking (SDN) is an approach to using open protocols, such as OpenFlow, to apply globally aware software control at the edges of the network to access network switches and routers that typically would use closed and proprietary firmware.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, instead of just putting a channel to transmit everything to the cloud and compute and come back we are doing some intermittent the provisioning of intermediate processing for to serve the application based. So, this is definitely a major motivation.

And what we try to look at that has as we have seen that the typical properties of cloud that having here infinite scalability, theoretically or quote unquote infinite scalability or off loading or having infrastructure no need of maintaining infrastructure at the client end or meter services those need to be preserved or respected right in case of a fog and those are definitely are still there as what we are discussed.

So, there are fog computing there are several enablers as those are true for our cloud computing also, one is that virtualization. So, virtual machines can be used as the edge devices right. So, there are there can be virtual machines containers or containers

services are reduces the overhead of resource management by using lightweight virtualization or what we say container based application or services is. One of the popular container is Docker container right. So, it the idea is it docks into that particular things and run on the thing. So, you do not have to that dependencies is carries along with the thing right. So, it is a again a separate topic it possible we will discuss sometime, but that is this docking or container services are becoming very popular. So, that is another enabling technology out here.

Such this oriented architecture as we have discussed which is a enabling technology for cloud also is here also that SOA is a style of software design where services are provided to the other components by application component a component through a communication protocol over a thing. So, you have a service oriented architecture which three major component of service provider, service comma consumer and service registry so that heterogeneous loosely coupled parties can talk to each other right.

So, in SOA architecture is one of the driving enabling technology, and also what we are looking seeing at is the software defined network right SDN. So, SDN is an approach of using open protocols like for example, open flow to apply globally aware software control at the edges of the network to access network switches routers that are typically would use closed and proprietary from one. So, this is another technology which is becoming pretty popular or already popular in software defined network, and which is enabling technology for fog also.

(Refer Slide Time: 24:45)

**Fog Computing - not a replacement of Cloud Computing**

- Fog/edge devices are there to help the Cloud datacenter to better response time for real-time applications. Handshaking among Fog and Cloud computing is needed.
- Broadly, benefits of Fog computing are:
  - Low latency and location awareness
  - Widespread geographical distribution
  - Mobility
  - Very large number of nodes
  - Predominant role of wireless access
  - Strong presence of streaming and real time applications
  - Heterogeneity

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, with this several enabling technology fog is becoming a reality, and being deployed and used in several cases.

So, in looking at we should not see that fog as a replacement of cloud, it is not a replacement of count nor neither a competitor in that sense right. It is basically offloading some of these workload from the cloud to this edge devices because the resources are available because there are applications which are real time and needs more quick responses and overall process may be cost effective and efficient.

So, fog edge devices are there to help cloud datacenters, to better response time for real time applications right. Handshaking amount fog and cloud is needed; appropriate handshaking or synchronization between these fog and cloud is very much needed. Broadly benefits of fog computing can be that low latency and location awareness. So, it is aware that which location is operating, widespread geographical distribution especially with the sensors etcetera. So, it has thing mobility there is another important things like nowadays devices are we have lot of mobile devices right. So, the distance from the cloud or the intermediate devices which a device say end device passing through the intermediate devices will change based on the mobility of the things.

Now, this require a resynchronization reestablishment of the path; had it been in locally somewhere it may increase the computing and response time. So, low latency and location awareness widespread these mobility, very large number of nodes with as we are

discussing with sensors and things predominant role of wireless access right huge volume of wireless accesses strong presence of streaming and real time applications. So, these days we are having a huge streaming and real time applications, and which requires quick response time huge volume of data and type of things need to be processed quickly and may not require all data to be transmitted right.

So, this huge volume of data can be locally processed and aggregated data can be transmitted so that the overall response time improves in a considerable way, strong presence of streaming and real time and heterogeneity, different sort of devices different type of the mix and inner and the heterogeneous. So, I can have it some sort of a fog sort of intermediate framework which basically talked to some devices which may be different from other devices. Like I have a group of sensors, I have a their sink node which talks to the sensor also do this aggregation which may be different in another other set of sensors which has a different sink note, but nevertheless when they do this aggregated data that is in a more standardized format. So, I can handle heterogeneous devices.

(Refer Slide Time: 27:53)

**FOG Advantages ?**

- Fog can be distinguished from Cloud by its proximity to end-users.
- Dense geographical distribution and its support for mobility.
- It provides low latency, location awareness, and improves quality-of- services (QoS) and real time applications.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, advantage is already we have already discussed. So, can be distinguished from cloud by proximity to the end user that is one of the advantage or over this free service cloud, dense geographical distribution and it is support for mobility right. So, we can have instead of scintillate I can have a lot of distribution. It provides low latency low

awareness and improves quality of service and real time applications right. So, there is a chance of better performing the things rather we try to look at it is not isolated fog, but fog plot cloud a as a whole can give a better service to these consumers right in terms of cost, in terms of scalability, in terms of your efficiency and type of thing specially applications where we have high quality of services and real time services streaming videos and type of things.

(Refer Slide Time: 28:52)

**Security Issues**

- Major security issues are authentication at different levels of gateways as well as in the Fog nodes
- Man-in-the-Middle-Attack
- Privacy Issues
- In case of smart grids, the smart meters installed in the consumer's home. Each smart meter and smart appliance has an IP address. A malicious user can either tamper with its own smart meter, report false readings, or spoof IP addresses.*

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | NPTEL

There are of course, some issues related to security. So, one is that as devices are dispersed right. So, maintainability of the security protocols at different fog devices is a serious challenge right. So, it is at different location now had it been cloud you have a provider at a particular centralized things, you can put lot of security mechanism in the place, but if you once you distribute over the form then you have to maintain so many things on the on different edge devices.

So, it is not only the data processing etcetera. So, there can be security issues. So, an man in the middle attack type of things can happen. So, as devices are disparts differ a as this computing data are being there in the in different edge devices. So, there is a issue of things of man in the middle attack can be there are issues of privacy issues as a as same that it is being processed at different edges and whether the data leakage is there.

Then whether you know about the things like if I consider smart grid or connected vehicles. So, if you do the intermediate ports processing whether you are basically

tracking the vehicle or looking at the processing of the consumption of a individual house or home, and type of utilization those can be there. Like in case of a smart grid smart meter installed at the consumer home, each smart meter and smart appliance as an IP address a malicious users can either tamper with it is own smart meter, report false reading or spoof IP addresses and so on and so forth.

So, whatever it comes with typical network security related issues may also come may also be problem out here. So, may be a challenge. So, there are definitely security issues there are security issues in the cloud, but this extend that to much more things as you have different devices are activated.

So, what we sees today that this fog is just not extension of the cloud, it is a necessity based on the different application, huge volume of data and the devices intermediate devices becoming more resourceful right and they are able to capable to do this type of calculations computation and. Secondly, in case in order to in doing so, I may not be doing all these high profile computation, but I can we can basically do some sort of a aggregation of the information and sending only the aggregated information, which lowers the basic bandwidth requirement intermediate bandwidth requirement also lowers the data load at the cloud end. So, what we see it is a technology which is a need of the hour and especially with IOTs and other things coming in a big way.

So, with this we will stop today.

Thank you.

**Cloud Computing**  
**Prof. Soumya Kanti Ghosh**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**

**Lecture – 34**  
**Fog Computing – II**

Hello. So, we will continue our discussion on cloud computing, rather we will continue our discussion on fog computing. So, what we have seen in the in our previous lecture or previous discussion on fog that there is a need of several application to instead of pushing all the data services and applications to the cloud instead whether you can do at a much lower level, right, there is some of the requirements are due to the bandwidth limitations like or reducing the bandwidth overload and in some of the things are real time applications, right, you need to do some real time applications which instead of pushing everything to the cloud getting the feedback process etcetera to end of the edge of the network may take much time, right.

So, in order to handle this, we need to push or there is a requirement of pushing some of this functionalities to the edge of the network or at intermediate level, right and also we have seen that all not all cases we require everything to be pushed into cloud, right, like specially applications like connected vehicles or you are that streetlight or traffic light management where it is more localized the phenomena is localized, right. So, it is more deal with the objects which are in nearby spaces.

So, it is no; there is not much requirement pushing all the data to the things. Again with the huge proliferation of sensors and for that matter IoTs, there is a huge volume of data generated where instead of sending the raw data to the cloud for processing, there we can do a pre processing or what we say some sort of aggregation of the information and push it to the cloud for further processing. So, overall in looking all those things, there is a need to bring this data services application little down from the cloud what we say fogging or fog computing in some cases; also people say as a edge computing or having a distributed phenomena of the things.

There are few characteristics; what we have looked into of cloud which need to be served here also like scalability infinite scaling or scaling pay as you go model or metered services and that like making the infrastructure free type of situation of the things and

several other characteristics or cloud provide what was the median motivation of moving towards cloud need to be supported at by the fog also, right. So, that is that is the need of the things and is not like that all the a given a particular application a everything should be put on the edge of the things there may be we can do a partially at the at the cloud end and partially at the at intermediate or edge of the network, right.

So, this will bring different sort of challenges what we will try to look at in today's talk that what are the different type of things; what are the fog; fog devices are not so resource reach, right, like they are devices like intermediate routers or at time the sensing devices or the sync of that particular sensor deployment sync node of the versus deployment which is not actually. So, cool at the cloud not no whereas, as this was cool at the backend high resource cloud. So, there is a resource management comes in a big way of success of this type of a phenomena, right.

(Refer Slide Time: 04:44)

## FOG Computing

- Cloud computing has been able to help in realizing the potential of IoT devices by providing scalability, resource provisioning as well as providing data intelligence from the large amount of data.
- But, the cloud has few limitations in the context of real-time latency (response required in seconds) sensitive applications.
- Fog computing has been coined in order to serve the real-time latency sensitive applications faster.
- Fog computing leverages the local knowledge of the data that is available to the fog node and draws insights from the data by providing faster response.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

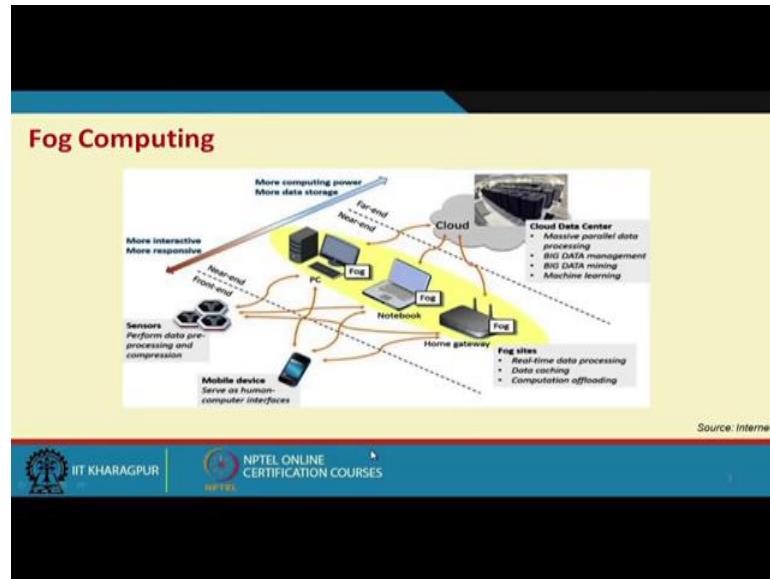
So, we will talk about fog computing. So, what we see that cloud computing has been able to help in realizing the potential of IoT devices or IoTs by providing scalable, scalability resource provisioning as well as providing data intelligence from the large amount of data. So, one is the scalability resource provisioning and the other end it can basically do some sort of a knowledge mining from the data right making data to from data to knowledge transformation sort of things, right that is at the backend, we are machine learning another type of algorithms which can run.

But the cloud has few limitations like specifically in the context of real time latency, right, response required in seconds or milliseconds or microseconds sensitive applications like in case of its a application accident say; if there is a collision of the car and if there is this cars are intelligent car what we say that having which are regular ad hoc network or sort of things; if we need to push this policy and related information to the cloud, get it refined and find out the location etcetera in from the nearby cloud that maybe by that time, there may be few more or many more applicants a accident could have happened.

So, it could have been done in a very localized manner where instead of taking it to that that cloud and doing the processing I could have done at the localized manner impossible. So, this sort of real time applications; where this accident management or some other type of applications what we see that there can be a possible phenomenon doing that so, fog computing has been coined in order to serve real time latency sensitive applications faster, right, it has been coined to serve real time faster.

Fog computing leverages the local knowledge of the data that is available to the fog and draws insights from the data by providing faster applications. So, first of all is a order to serve real time latency or real time applications or which are time sensitive applications and also it as it leverages; the local knowledge of the data more at the local level available to the node and draws insight from the data by faster response, right. So, that is that exactly it tries to look at that fog sort of environment.

(Refer Slide Time: 07:22)



Now, this picture we have seen in the earlier our discussion on fog like we have one in cloud one in these sensors and mobile devices another; what we say contributing to were contributing to the data intermediate devices can constitute this fog, right.

So, this is this may not be specially installed for that right. So, we are having say routers or gateways some systems and other devices which are basically communicating to which are basically used for intermediate communication of the sensed data to the cloud, right. So, transmitting they are more as a some sort of a working as a store and forward from there why whether we can do store process forward and take some all locally right some of the things need to be forwarded at the other end for higher level of things or maybe aggregation over a over a larger geographical space when different sensors coming from the things nevertheless we can have a localized phenomena what we look at the using the data at the local level.

So, what we see that more computing power; more storage is at the end. So, the applications which are more computing intensive can be push to the other end whereas, more interactive or more responsive can be at the lower end. So, these are the 2 side of the things. So, intermediate; we have fog or even in some cases, these inter frontend devices can do some sort of a calculation and aggregate the data and same to the things.

(Refer Slide Time: 09:10)

The slide features a title 'Fog Computing and Cloud Computing' at the top. Below it is a table comparing 'Cloud computing' and 'Fog computing' across seven requirements. The table has three columns: 'Requirement', 'Cloud computing', and 'Fog computing'. The requirements listed are Latency, Delay jitter, Location of server nodes, Distance between the client and server, Security, Attack on data enroute, and Location awareness. The source of the information is cited as 'Source: Internet'.

Requirement	Cloud computing	Fog computing
Latency	high	low
Delay jitter	High	Very low
Location of server nodes	With in internet	At the edge of local n/w
Distance between the client and server	Multiple hops	One hop
Security	Undefined	Can be defined
Attack on data enroute	High probability	Very Less probability
Location awareness	No	Yes

Source: Internet

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, if you look at if we try to compare or compare and contrast between cloud and fog as we have discussed in earlier lecture that these are not competitive means in the sense they are not replacing one by another. So, what it is trying to do; it is mode of a; in a companion mode, right, some of the things which can be done better in cloud and should be done there and fog and they should do in a proper orchestration of the things a much requirement of the orchestration.

So, if we take a look at requirement of cloud and fog. So, if we say latency. So, is higher in cloud or lower in fog, right. So, latency is a higher delay jitter definitely the then say high are usually it is high and it is very low in the fog or low in the fog, location of server nodes within internet, right whereas, at the edge of the local network, right. So, where the server nodes are located in case of a cloud it is in the Internet, right you really do not know that where the services are given you talk about Amazon or Google or Microsoft or IBM or anything. So, sports or anything; so, what we do? We basically connect through their portal or link and we really as the individual do not know that where your application is being done; it is not like that that you cannot know; you can know, but nevertheless, they do on their resource management and provisioning and type of thing. So, that is what we say in the internet whereas, if you look at that fog type of scenario.

So, it is more on the edge in the local network like if I am doing say aggregation of the temperature sensing and taking a design of this particular room, with which is having say 10 odd sensors, then what I am doing is basically locally and I know that the server in this particular room is working for that things or the server which is serving for this particular room is working on the things, right at times that is useful because if you see some congestion or some problem in the things you can address the things at times there are other challenges, it may be a security loophole also because you know, if the temperature at being sensed by the servers and it is sending to things and if this particular lab or room in housing important other systems, then I can basically attack that server and say that do some manufacturing, right, even the temperature or means environmental sensing is giving some allowed, I say that everything is going on fine and at that type of thing.

So, there are downfall and type of things required and. Secondly, it needs to be resourceful to cater to the type of applications which I am trying to do. Distance between client and the server usually in case of cloud computing is multi hop, you are using this standard networking to go to the thing usually in case of a cloud, it is a one hop may in case of a fog, it is a one hop, right. So, they are at a one hop distance from the client to the server, right.

Security in case of a cloud computing 1; 1 argue is it is undefined; undefined in the sense that as the user and I do not have much control over the things, right. So, it is in that sense, it is undefined whereas, here it can be defined like you have a local things you may have some control over the thing. So, you use the may have some control over these particular devices or the organization can have control over the devices and can try to ensure some security like if I say the traffic light. So, of a particular city, then the traffic light management in that server of a particular zone is under the traffic authority of that city, right. So, they have a control had it been on the totally on the cloud. So, you do not know that what the data or applications are doing that is that is based on that service provider.

Attack on data enrooter; right. So, in the enroot data whether the attack in case of cloud if there are multiple hops. So, there is a chance of much become getting much compromised where in case of a fog, if it is a single hop, then as it is a single hop, then the getting compromised things are less, right. So, you have a little more control over the

thing location awareness in case of a cloud computing is minimal whereas, edge fog computing is location aware, it is primarily what we are doing is location aware type of things.

(Refer Slide Time: 14:24)

**Fog Computing and Cloud Computing**

Requirement	Cloud computing	Fog computing
Geographical distribution	Centralized	Distributed
No. of server nodes	Few	Very large
Support for Mobility	Limited	Supported
Real time interactions	Supported	Supported
Type of last mile connectivity	Leased line	Wireless

Source: Internet

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, there are other things like geographical distribution in case of a cloud it is more of a centralized feeling. So, it is having a logically centralized things, in case of a fog it is distributed right number of server nodes in case of cloud are few because the that is at the clouding and usually servers are extremely resourceful though number of nodes required for the publications or type of things are very few in case of a fog; as it is edge as you go down the hierarchy, that the number of nodes increases much more. So, they are the very large number of urban nodes. Support for mobility in case of a cloud is limited right if you move from one or mobile application that it need to be switched etcetera to the different intermediate devices.

Now, it is carried by say one path once you move to the other path whereas, in case of a fog some sort of support is there in the mobility, because it takes a local deals and other things, right, it knows a priori things are that is where under which control and need to be transfer at a much lower level.

Real time applications real time interactions do supported by cloud and of course, it is supported by fog that is one of the major motivation to move towards fog computing scenarios type of last mile connectivity is usually disliked, right. So, or cable line in case

of a cloud, right whereas, in case of a fog is usually wireless, right that there is no hard and fast type of things, but these are you usual standard processes in reality, right.

So, there are pros and cons; it may so happen that as we are discussing fog. So, it is little bit more supportive towards the fog why it is there and, but nevertheless to both has importance and proper a synchronization orchestration between this fog and cloud should make the whole thing a reality.

(Refer Slide Time: 16:50)

**Fog Computing Use-cases**

- **Emergency Evacuation Systems:** Real-time information about currently affected areas of building and exit route planning.
- **Natural Disaster Management:** Real-time notification about landslides, flash floods to potentially affected areas.
- Large sensor deployments generate a lot of data, which can be pre-processed, summarized and then sent to the cloud to reduce congestion in network.
- **Internet of Things (IoT)** based big-data applications: Connected Vehicle, Smart Cities, Wireless Sensors and Actuators Networks(WSANs) etc.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, there are several use cases, right or several scenarios where fogs are; fog will be very much applicable, one is that emergency evacuation system for any catastrophe or disaster real time information about currently affected areas or buildings and exit route planning etcetera. So, if it is a large building with several route paths for exit; if there is some catastrophe like fire or earthquake or something then the default paths may be getting locked right or may not be routable.

So, what you need to do? You need to do a based on that availability of the exit at every level or every individual rooms, etcetera, we need to path plan the path. So, we need to have it need to be dynamically replan or reroute it and where may be some local DSM is much more helpful natural disasters management. So, real time notification about landslides flash flood to potential affected area. So, that is one requirement. So, when we are having natural disaster management. So, real time notification things and there which are sometimes pretty localized for a particular region of interest where the things are

going on and may be useful way if we have this location aware information at a fog level.

So, large sensor deployments generate a lot of data which can be preprocessed, summarized and then to send to the cloud to reduce congestion in the intermediate network. So, that is another requirement of fog, it may not be natural disasters management or natural disasters or hazard, but in other sense what we have that huge deployment of sensors that can have produce lot of data and it takes if you send the all the raw data to the cloud it takes a lot of bandwidth and lead to congestion.

So, which can be reduced by moving, pushing aggregated data and of course, the Internet of things, right based on big data applications like connected vehicle smart cities wireless sensor network, actuators, networks and those and its said the; so, these are all different-different; what we aspects of are the scenarios of internet things which again push lot of data and in the whole framework and all every time that you are; this may not be important like that maybe a local DSM is more important than taking a global DSM and type of things.

(Refer Slide Time: 19:45)

The slide has a yellow header bar with the title "Applicability" in red. Below the title is a bulleted list of six applications: Smart Traffic Lights, Connected Vehicles, Smart Grids, Wireless Sensors, Internet of Things, and Software Defined Network. At the bottom of the slide, there is a footer bar with the IIT Kharagpur logo and the text "NPTEL ONLINE CERTIFICATION COURSES".

Applicability
• Smart Traffic Lights
• Connected Vehicles
• Smart Grids
• Wireless Sensors
• Internet of Things
• Software Defined Network

And if we look at the applicability, there are few here there are hundred more which can be there. So, smart traffic lighting maybe one application, connected vehicles, smart grids of course, sensor network, internet of things and software defined network. They provide this backbone of these applications to work on, all right.

(Refer Slide Time: 20:07)

**Connected Vehicle (CV)**

- The Connected Vehicle deployment displays a rich scenario of connectivity and interactions: cars to cars, cars to access points (Wi-Fi, 3G, LTE, roadside units [RSUs], smart traffic lights), and access points to access points.
- Fog has a number of attributes that make it the ideal platform for CV in providing services, like infotainment, safety, traffic support, and analytics: geo-distribution (throughout cities and along roads), mobility and location awareness, low latency, heterogeneity, and support for real-time interactions.

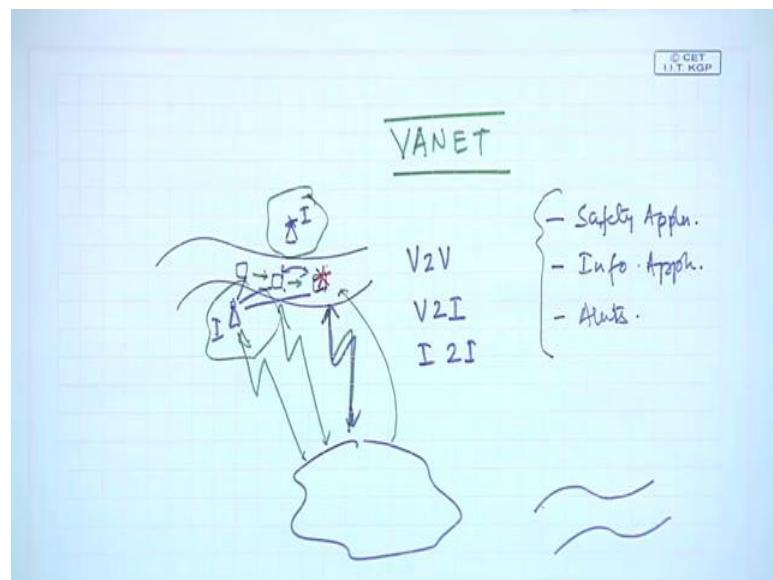
Source: *Fog Computing and Its Role in the Internet of Things*, Flavio Bonomi, Rodolfo Millo, Jiang Zhu, Sateesh Addepalli

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, if we look at the connected vehicle deployment displays rich scenario of connectivity car to car; car to access points and type of things.

So, what we look at when we talk about vehicular infrastructure we have on the road different moving cars, right.

(Refer Slide Time: 20:46)



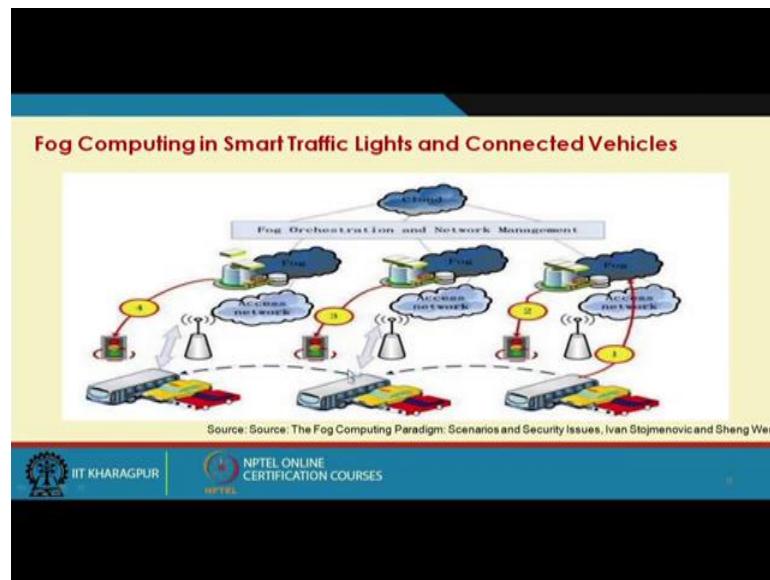
So, these are different vehicle and of course, different infrastructures. So, we have connectivity between this vehicle to vehicle or what we say V2V or V2I, I2I. So, these are different infrastructure which are hovering around and giving connectivity to these

different vehicles. So, this sort of things are different type of application one is safety related applications, another is say information on infotainment information related application and there are other things like Alerts and other type of things based on the title.

So, those are the information which will be there and in c there may be a overall cloud infrastructure where all could have communicated right and it takes a diesel and sends back information type of things now say there is a clash in this particular vehicle, right, if there is a clash then other vehicles which are approaching this vehicle now get information via this instead it could have been done locally, right, if I can set up a some sort of a fog around this sort of things, then I could have taken a local diesel because the accident here may not be nothing to do with some road going somewhere in some cities etcetera right or even in the same city some other part of the thing.

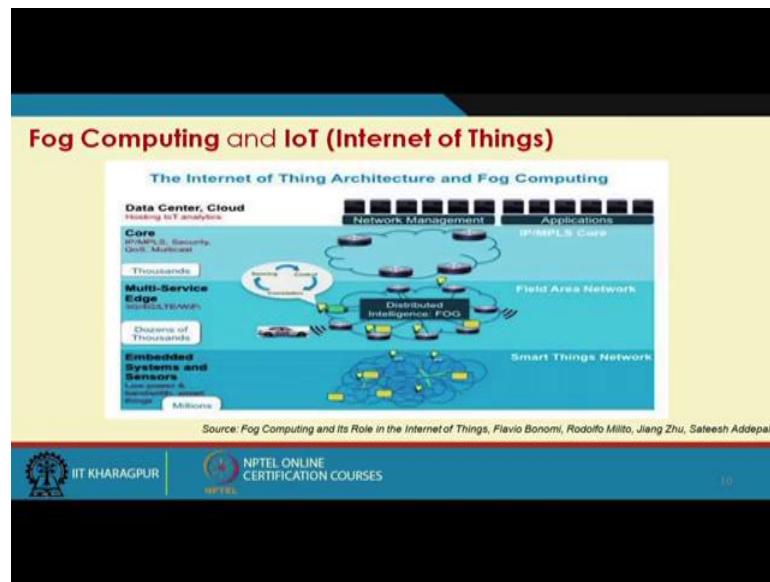
Now, here this type of connected vehicle phenomena or what we say a concept which is coming up or it is already they are VANET vehicular ad hoc network, right. So, to make it successful this fog maybe one of the applications, right. So, again smart city lighting as we see that if there is a traffic congestion etcetera that within the thing within that particular localized things sitting. So fog has a number of attributes that make idle platform for connected vehicle in providing services like infotainment safety traffic support and analytics like geo distribution, mobility, location awareness, low latency, heterogeneity and so and so forth. So, there is a lot of applications are there.

(Refer Slide Time: 23:15)



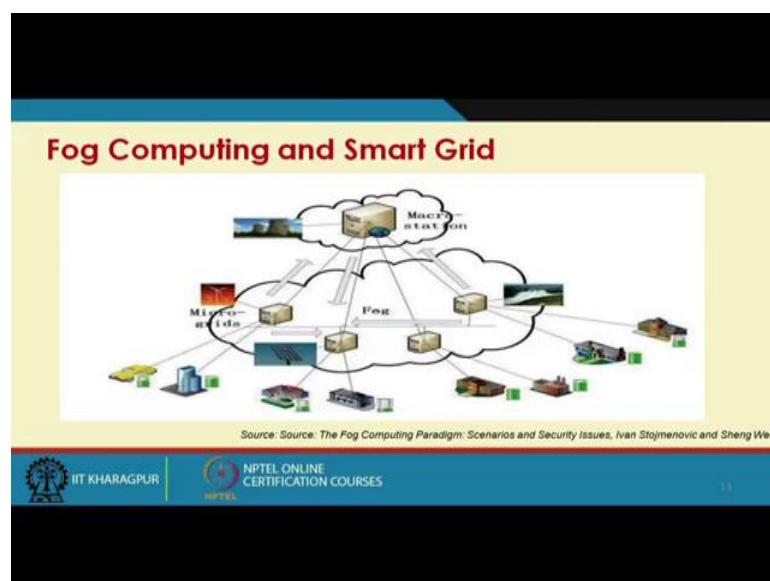
So, like what we see that there are different vehicles running at the backend. So, there are the access points or what we say that infrastructure along or sometimes known at road side unit or RSUs. So, these units are there; there are other traffic lights. So, based on this congestion etcetera this traffic light can be intelligent like it, it gives the timing for stop and go type of things may vary based on the congestion level. So, and at the backend we have that cloud which takes larger analytics problems which requires larger resources and we require a fog orchestration and network management layer which takes care of this synchronization or orchestration of the fog and also orchestration with the backend cloud.

(Refer Slide Time: 24:17)



So, it is a system. So, what we try to see that this fog take a local decision whereas, large analytics can be pushed to the cloud environment, right. So, this is one of the scenarios there of course, if we look at a more of Internet of things. So, at the lower end we have embedded systems and sensors then multi service edge which takes the distributed in inclusions where fog can played a role and then we have a core network which is which is used to push the things at the upper cloud, right. So, this is in generic way of looking at the internet of things; where we see that a fog layer may help in reduce latency and providing better services.

(Refer Slide Time: 24:51)



And though in our country may not be highly proliferated or it is not in a big way use, but it is going to come is that a smart grid, right. So, every home will have a smart meter and based on the utilization of the things at a from the home level to from at the house unit of the house to as say region level and a largest state level the overall management can be done. So, I have a smart grid which not only give power to the homes and offices and installations also take a feedback and takes a call based on the things that how the power utilizations are there; it is connected to the power plant and the overall power management across the region across the country or across a larger geographical space can be managed by these sort of things.

So, here also fog plays a important role if I this if you are looking as all things can be pushed into the cloud and take a call, right; however, suppose I consider IIT Kharagpur, if the homes are having smart meters then I could have taken a local decision that what is the overall utilization of the power and type of things and I send aggregated information to the backend cloud, right. So, which takes a more takes of this aggregated information and take to analytics that over the over larger time span or over days months etcetera how things varies and take a call that what to provisioning of electricity based on the things.

So, this type of things are useful where we have that in devices or in consumer of electricity, then we have a fog infrastructure where micro bleeds and other things are there.

(Refer Slide Time: 26:55)

## Fog Challenges

- Fog computing systems suffer from the issue of proper resource allocation among the applications while ensuring the end-to-end latency of the services.
- Resource management of the fog computing network has to be addressed so that the system throughput increases ensuring high availability as well as scalability.
- Security of Applications/Services/Data

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

And then we have a micro station and push it to the larger in sort of a cloud environment and as we see that all; what we say lucrative or golden side of the form there are few definitely challenges or there are good amount of challenges in or to have a realization first of all these devices are not that resource full as cloud or cloud servers, etcetera.

So, that can give; what we say support to a minimize things. So, if it is such application is there which require much larger thing? So, you need to divide that biggest and accordingly, right. So, it may be among the fog devices some portion on these fog devices and some of the cloud and whenever we do this there is a lot of need of synchronization orchestration of the things, right because not only now the data are divided your application or the process is also divided. So, somewhere this aggregation of this data and processes need to be there. So, that is one of the challenges there are other challenges of the resource management in the fog itself like suppose if one of the fog device is overloaded whether I can migrate this application on the things whether I migrate this on the life like executing things can be migrated to the things.

So, these are serious challenges if we if we need to look at the things. So, let us see some of the things. So, fog computing system suffers from issue of proper resource allocation among applications while ensuring end to end latency of the services right. So, what we want to do that end to end latency and challenges are face resource management of the fog computing network has to be addressed so, that the system throughput increases

ensuring high availability as well as scalability. So, the basic phenomenon of the fog and finally, as these are distributed over different geographical space may be at with different authorities, then what about the security of this application and data and type of things whether that becomes a source for things; so, security aspects we discussed last lecture or last discussion on fog that this is a serious challenge.

(Refer Slide Time: 29:14)

**Resource Management of Fog network**

- Utilization of idle fog nodes for better throughput
- More parallel operations
- Handling load balancing
- Meeting the delay requirements of real-time applications
- Provisioning crash fault-tolerance
- More scalable system

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, resource management in fog network; so, has different aspect they utilization of idle fog nodes for better throughput whether there is possible there is some of the fog nodes and it is basically skewed some nodes are more loaded than others more parallel operations how to generate more parallel operations handling load balancing meeting the delay requirements of real time applications, right. So, if you have real time applications; how it can be provisions properly provisioning crash fault tolerant and type of things, right. So, like I can say that if the fog node goes down what will happen to those applications and data and which are running on that fog nodes how to handle those how to migrate those data whether I require a prior in application that even if goes down the other will take up and all those things require a resource management and incurs cost and so and so forth.

More scalable systems; so, at the scalability is our core of the whole thing right scalability is one of the major aspects of cloud we service fog computing; so, how to have better scalable systems.

(Refer Slide Time: 30:22)

**Resource Management – Challenges**

- Data may not be available at the executing fog node. Therefore, data fetching is needed from the required sensor or data source.
- The executing node might become unresponsive due to heavy workload, which compromises the latency.
- Choosing a new node in case of micro-service execution migration so that the response time gets reduced.
- Due to unavailability of an executing node, there is a need to migrate the partially processed persistent data to a new node. (State migration)

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, they are if we look at little more nitty-gritty data may not be available at the executing fog node. So, the application is there the node may not be there therefore, data fetching is needed from the required sensor or data source. So, that is a phase cyclist their executing node might become unresponsive due to heavy workload which compromises of the latency there may be a issue choosing a new node in case of a micro service execution migration.

So, that the response time gets reduced even if I have a way of migration if I am a micro service is running on a particular node or a smaller or cut down version of the service or chopped service or a partition service is running if I am if even if I am able to migrate that I see that this node is going down; what should be where I should migrate how to find a node at a thing.

So, some algorithms would run and type of things and some sort of managements would come into play due to a unavailability of executing node there is a need to migrate partially processed persistent data to a new node. So, it is half-cup thing need to be migrated. So, that is another need final result has to be transferred to the client or actuator within less amount of time in order in doing.

(Refer Slide Time: 31:31)

**Resource Management – Challenges (contd...)**

- Due to unavailability of an executing node, there is a need to migrate the partially processed persistent data to a new node. (State migration)
- Final result has to be transferred to the client or actuator within very less amount of time.
- Deploying application components in different fog computing nodes ensuring latency requirement of the components.
- Multiple applications may collocate in the same fog node. Therefore, the data of one application may get compromised by another application. Data security and integrity of individual applications by resource isolation has to be ensured.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, I should not lose out on the time deploying application components in different fog nodes ensuring latency requirements of the components. Multiple applications may collocate in the same fog node, right, therefore, the data of one application may get compromised by the other.

So, it may happen that number of applications more than one application in the same data security and integrity of individual application and resource application has to be ensured, right. So, what we look at the multi-tenancy problem in the cloud that sort of problem can be there in the fog and fog gives less resource.

(Refer Slide Time: 32:19)

**Resource Management – Approaches**

- Execution migration to the nearest node from the mobile client.
- Minimizing the carbon footprint for video streaming service in fog computing.
- Emphasis on resource prediction, resource estimation and reservation, advance reservation as well as pricing for new and existing IoT customers.
- Docker as an edge computing platform. Docker may facilitate fast deployment, elasticity and good performance over virtual machine based edge computing platform.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, that may the problem may be more escalated and there are several approaches people try to follow like executing migration on the nearest node which is available or to the thing. So, the nearest node which is free may not be the most suitable, but the available node type of things minimizing carbon footprint or video thing my major objective is that to reduce that energy or carbon footprint emphasis on resource prediction whether I can have a the approaches, prediction, resource estimation, reservation, advanced reservation as well as pricing of the new IoT application; so, that we can do a priori estimation of the things.

There is another service which we say Docker as an edge computing platform to deploying Docker may facilitate fast deployment elasticity good performance.

(Refer Slide Time: 33:06)

**Resource Management – Approaches (contd...)**

- Resource management based on the fluctuating relinquish probability of the customers, service price, service type and variance of the relinquish probability.
- Studying the base station association, task distribution, and virtual machine placement for cost-efficient fog based medical cyber-physical systems. The problem can be formulated into a mixed-integer non-linear linear program and then they linearize it into a mixed integer linear programming (LP). LP-based two-phase heuristic algorithm has been developed to address the computation complexity.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, there are resource management based on fluctuating relinquished probability of the customer research prices etcetera people follow that there are other things like studying the base station association tasks distribution virtual machine placement and so on for and formulating a LP formulation to optimize the thing applying heuristics algorithm to a approach that problem.

So, these are the different type of approaches which people are trying to do and you can see; these are there are lot of research motivation here, right; there are lot of research going on and those who are interested; this is a field where which can be looked into or you can work on those type of aspects this is a ongoing another upcoming area to look at.

(Refer Slide Time: 34:05)

**Fog - Security Issues**

- Major security issues are authentication at different levels of gateways as well as in the Fog nodes
- Man-in-the-Middle-Attack
- Privacy Issues
- In case of smart grids, the smart meters installed in the consumer's home. Each smart meter and smart appliance has an IP address. A malicious user can either tamper with its own smart meter, report false readings, or spoof IP addresses.*

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

And security issues already we have discussed I am not repeating that. So, security what we try to emphasize here security is also a major challenge right because low resource you cannot run say resource pool or what we say resource hungry security applications what we say which or security measures which are resource hungry. So, you need to be need to be appropriately sized to this fog type of fog devices to run on.

So, security is also a major issue it is not only that the data I can be compromised the fog is devices may be a platform to simulate to simulate or launch attacks right. So, because it is distributed now things are distributed less control over this centralized cloud and the ISP. So, there is a there is a chance of this being exploited. So, need to be looked into things are there how secured or how robust this fog devices are also a major challenge, people are working on with this we let us stop today.

So, what we discussed that that the importance of fog, amalgamation of the fog and cloud that it is not like that a through throughout one of the technology are things are there rather proper synchronization and orchestration between them is a is the real way to have a successful implementation of this framework.

Thank you.

**Cloud Computing**  
**Prof. Soumya Kanti Ghosh**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**

**Lecture – 35**  
**Use Case: Geospatial Cloud**

Hello. So, we will continue our discussion on cloud computing and today we will talk about a typical use case of spatial cloud or more specifically geospatial cloud, right. So, what we mean by this geospatial data and information and type of things. So, what we see these days that this spatial information about earth surface and both with coordinated system, right with location based information are becoming extremely popular we are we all of us are more or less used to or using day in day out these location based services or navigation services for finding trajectories from one location to other and type of things, right.

So, all these are possible because there is a there are spatial data which are being maintained by various organizations, right as data is becoming more heavy with and multiple organization a storing this data. So, there is a need that how to integrate this data for decision makings purpose, right one navigation maybe one of the things there are several other decision making things specially different development program of government and government agencies or state agencies and type of things which need require some sort of a ground level information to be integrated, right.

So, one way of looking at it that I store the data in a central place and then query on the look at the things, right, but if there are multiple stakeholders of the data or multiple providers of this data, then these making it is a central repository sometimes become a challenge, right.

First of all the data itself is cost intensive if there are expertise and there are manpower there are fund involving a collect in collection and maintenance of the data. So, once you put the data in some other place then the whole thing goes for a different type of management right. Secondly, that the volume of data is enormous. So, this sort of transferring of the data over things may be extremely difficult, right

So, what is what one of the solution what this spatial science or spatial management and things are looking at whether we can migrate to a cloud, right? It is some sort of a cloud sort of infrastructure which can be provided on the on this platform, right. So, this is a lot of work going on across the globe and we would like to see that what is the feasibility what sort of things are they in here and IIT Kharagpur also we have a small group working of those; this sort of activity.

So, we will try to see that what is the feasibility of this sort of domains spatial domain going to the cloud, right.

(Refer Slide Time: 03:31)

## CLOUD ?

- ▶ **On-demand self service**
  - ▶ Use resources as and when needed
  - ▶ Minimal human interaction between user and CSP
- ▶ **Ubiquitous Network Access**
  - ▶ Services accessible over Internet using Web applications
- ▶ **Resource Pooling**
  - ▶ Large and flexible resource pooling to meet the consumers' need
  - ▶ Allocating resources efficiently and optimally for execution of applications
- ▶ **Location Independence**
  - ▶ Resources may be located at geographically dispersed locations
- ▶ **Rapid Elasticity**
  - ▶ Dynamic scaling up and down of resources
- ▶ **Measured Services (*pay-as-you-use*)**
  - ▶ Customers charged based on measured usage of the cloud resources

CSE, IIT Kharagpur

So, if you look at. So, this is all we are discussing for long like what cloud keeps a on demand service ubiquitous access resource pooling location independence rapid elasticity measured services right and so, this is one part of the cloud side.

(Refer Slide Time: 03:47)

## Geographic Information

- ▶ Information explicitly linked to locations on the earth's surface
- ▶ Geographic information can be static or dynamic
  - ▶ Static: does not change position
    - ▶ Locations, such as city/town, lake, park
  - ▶ Dynamic: changes over time
    - ▶ Population of a city
- ▶ Geographic information vary in scale
  - ▶ Information can range from meters to the globe
  - ▶ Scale vs. detail and ecological fallacies



CSE, IIT Kharagpur

If you look at the geographic information side or the information spatial information or geospatial information side; so, information explicitly linked to the location on earth's surface, we are primarily looking at the spatial data which is looking specifically on the earth surface, right. So, it is a code driven thing. So, everything every data has a location incidentally if you see there was a report that more than ninety percent of our data are somewhere other related to some location, right.

Suppose if I see our student data here. So, along with lot of other things it says that which hall he or she is staying which department he or she is studying right where he is or his or her hometown. So, all those things has a PoI a location on the earth surface. So, somewhere other with this spatial information is inherent or it is there in all our day to day life not only that in our most of our databases also right we may not be using it, but it is there if there is a pin code it says that which region it is referring to; right.

So, it is explicitly earth surface geographic information can be static or dynamic like static we does not change with time or changes over a long period of time like it may not be like a boundary of a state or a boundary of a district or a or sometimes in park lake etcetera they are not changing je day in the out are there are can be dynamic like some of the say traffic movement or population of a particular city which varies over time right there are lot of influx during data in people are going out during night time and type of things if I look at this particular building where we are sitting here at now and so, it has

the if you look at the building wise it is static right or we can say pseudo static it is not changing day in day out, but if you look at that food fall on the building that is changing right or even electricity need of the things may be changing.

So, some of the data can be static some of the things can be dynamic and there is a typical thing this geographic data varies over in scale right. So, information can range from meters to globe right some of the data if I want to point, it needs say up to a meter scale right find out that what is the location of a lamppost right or, but it is can be higher the basically the map of a country or map of a district or map of a state may be much larger scale as; so, and scale versus detail and ecological fallacies and those things are there. So, with the scale things goes on changing right like if I look at a building from a higher from the sky it is a point, but if I close down it becomes a polygon and the with the scale the overall nature may change. So, these are some of the characteristics of the data, right.

(Refer Slide Time: 06:55)

## Geospatial Information

- ▶ Legal (cadastral; zoning laws)
- ▶ Political (county lines; school districts)
- ▶ Cultural (language; ethnicity; religion)
- ▶ Climatic (temperature; precipitation)
- ▶ Topographic (elevation; slope angle; slope aspect)
- ▶ Biotic (biodiversity; species ranges)
- ▶ Medical (disease; birth rate, life expectancy)
- ▶ Economic (median income; resource wealth)
- ▶ Infrastructure (roads; water; telecommunications)
- ▶ Social (education; neighborhood influences)

CSE, IIT Kharagpur

And if you look at that common geographic data what we which we are accustomed with there can be some legal data that is location at these boundaries etcetera political maps, cultural, climatic, some economic, etcetera.

(Refer Slide Time: 07:14)

## Geospatial data source

- ▶ Social surveys
- ▶ Natural surveys (i.e. SOI maps)
- ▶ Remotely sensed (air photos, satellite imagery)
- ▶ Reporting networks (weather stations)
- ▶ Field data collection (GPS data or map marking associated with some attribute of interest)

CSE, IIT Kharagpur

So, there can be n number of such structure and there can be different agencies who are collecting the data or who are maintaining the data that can be some sort of social survey there are national survey like we have survey of India maps, etcetera there can be some remotely sense sensor bits things or sensor data like it can I can have remotely sensed data like air photos photographs taken from the air or satellite images even reporting by weather stations collected by GPS and map marking associated with some attribute interest there can be pollution sensors across the things which gives data with locations where things are there; there can be air quality control pollution sensor temperature sensor and other type of atmospheric sensors switch gives data.

So, these are different sources of data incident with these data are maintained by different organization by their proprietary nature and when I want to query of one more than once such data that I may need to see our or interoperate between the data, right where this we need to see that where this cloud will be make some sense where using it.

(Refer Slide Time: 08:26)

## Geographic Information Systems (GIS)

- ▶ A computer system for capturing, storing, querying, analyzing, and displaying geospatial data. (Chang, 2006)
- ▶ Geographic information systems are tools that allow for the processing of spatial data into information, generally information tied explicitly to, and used to make decisions about, some portion of the earth (Demers, 2002).



CSE, IIT Kharagpur

So, if we look at some of the formal definitions like for capturing, it is a capturing storing querying analyzing and displaying geospatial data geographic information system or tools or platform which allows you to spatial data information related explicitly or implicitly for that surface.

(Refer Slide Time: 08:48)

## Components of a GIS

- ▶ Computer hardware
- ▶ Software
- ▶ Data management and analysis procedures (this could be considered part of the software)
- ▶ Spatial data
- ▶ People needed to operate the GIS



CSE, IIT Kharagpur

So, we are not going into deeper things just to show that why what sort of data just to have a idea of the and it can have hardware required hardware software data management is a major part spatial data which is has a different type of characteristics

both volume wise meaning wise and type of things and you require a special category of people who can operate on this data, right.

(Refer Slide Time: 09:16)

### Geospatial Information System - Challenges

- ▶ Data intensive
- ▶ Computation Intensive
- ▶ Variable Load on the GIS server demands dynamic scaling in/out of resources
- ▶ GIS requires high level of reliability and performance
- ▶ Uses Network intensive web services



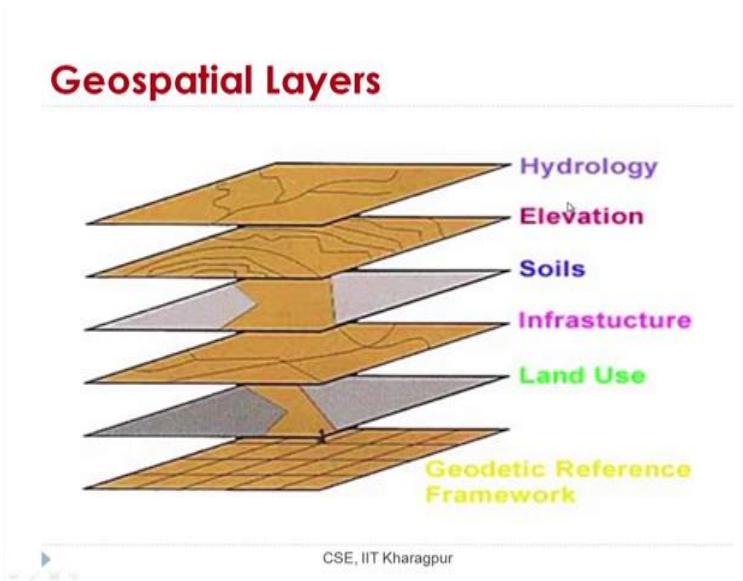
CSE, IIT Kharagpur

So, that is what we say people who can work on those type of things. So, these are the different component and what are the major challenges which are faced by this sort of huge spatial data it is data intensive; it is computation intensive, if I want to find a shortest path from a to b where the number of roads etcetera are like if I look at the representative as a road network as a graph. So, huge volume of the things of this graph then it is extremely computationally intensive not only that if I want to find out a some sort of a overlaying type of things then also it is computationally intensive. So, variable loads on the GIS server demands dynamic scaling in and scaling out of the resources and another interesting thing; it is not like that always you are doing what. So, data is there whenever the query comes you are doing and that it may also varies depending on the things.

So, we may require scaling in and scaling out of resources, right. So, that will be advantageous that when I have huge computation I scale bring in resources when I do not require; I release those resources type of things. So, there is a chance of having this sort of utility based computing or cloud sort of computing here and uses number of time uses network intensive web services as we have discussed web services are services which are based on this service oriented architecture and instead of data driven we have a service

driven. So, we are talking about spatial web services which take care of spatial data type of things. So, it is extremely network intensive because the data at very place maintained by different organization and you need network intensive thing.

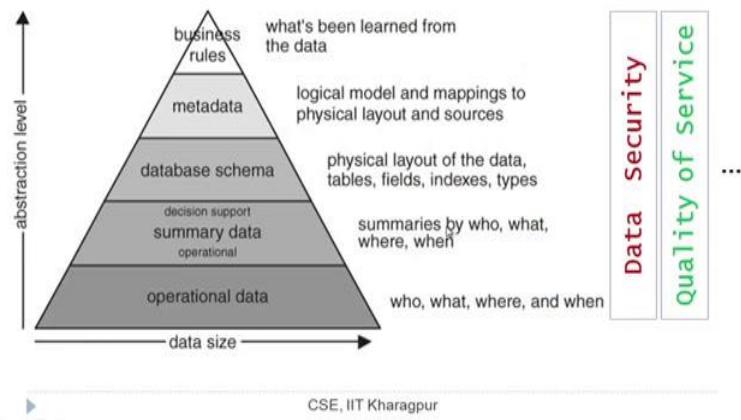
(Refer Slide Time: 11:02)



So, if we look at a typical batch of land. So, if you see the spatial data there can be hydrological data there can be elevation data soil data infrastructure same place and we require a referencing system, right, the means based or the well known referencing system what we understand from our school days is the lat-long, it can be other referencing system which basically try to pin the data to a particular location.

(Refer Slide Time: 11:28)

## Generic Architecture of Data



And if this is not only true for spatial data is true for any sort of data repository like we have operational data then summarizes making schema then metadata information of the schema which contains not only this schema information; lot of other things and there are business rules which are which the finally, the end user wants right that I want to know from location Kharagpur; IIT Kharagpur to going to say IIT Chennai what should be the root or IIT Kharagpur to some place in Kolkata. So, I am not bothered about that how where data is stored how they are maintained I am I am more bothered about that what should be my path and what are the PoIs or the point of interest across the path and what is; how I should move and what may be the expected time to reach that things what are the congestion level, etcetera.

So, I am more interested in the service rather than the actual data itself right. So, these are required. So, that is my that is my business rule that I want to find the shortest path I may want to find a particular corridor of the things like if I make a say if one man makes a canal. So, what should be its catchment area how much it will resolve? So, it says that if there is a canal is. So, much flowing water is there then I want to see the means if the if the rule is that it says that both side 2 kilometer buffer will be the catchment area where the that water can be served. So, out of this 2 kilometer how many population has been served how much agree means land agricultural land has been served can be the way or means can be my calculation.

(Refer Slide Time: 13:32)

## Heterogeneity Issue

- ▶ **GIS layers** are often developed by **diverse departments** relying on a mix of software and information systems
- ▶ **Each department** uses its individual system to **increase efficiency**, but sharing data and applications across the enterprise is a near impossible
- ▶ Issues to be resolved
  - ▶ Making *data description* homogeneous
  - ▶ Standard encoding for data
  - ▶ Standard mechanism for data sharing

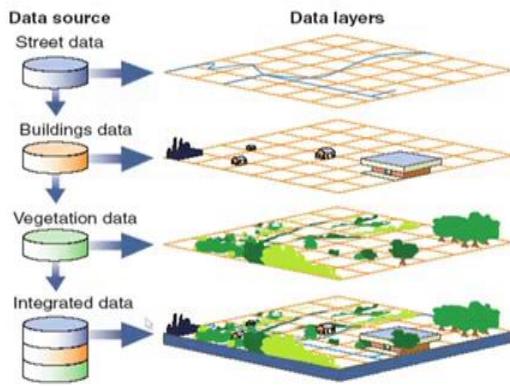
CSE, IIT Kharagpur

So, that is my business. So, the end user is more at the top of the thing and this is true for all sort of data and we have some of the verticals like data security which is a into end there are quality of service how reliable there can be management services and different other verticals which are which goes from the different aspects and what we try to see in case of specially in case of this sort of spatial data of geospatial data, we want to have some sort of a homogeneous way of looking at it like that different organization keeping that data when I query; I require them in the same type of format same type of means both syntactic and semantic wise some sort of a homogeneity otherwise I cannot query on those data, right.

So, some references system is different, then I cannot quarry on the data if the scales are different then and. So, I cannot query data, right. So, that we need to resolve issues like data description that were homogeneous I should understand standard encoding of the data standard mechanism of the data sharing. So, all this what it says that some sort of a computing centralized computing facility which are like cloud type of things which can be helpful for the things, right.

(Refer Slide Time: 14:30)

## Homogeneity (Needs to be achieved !)

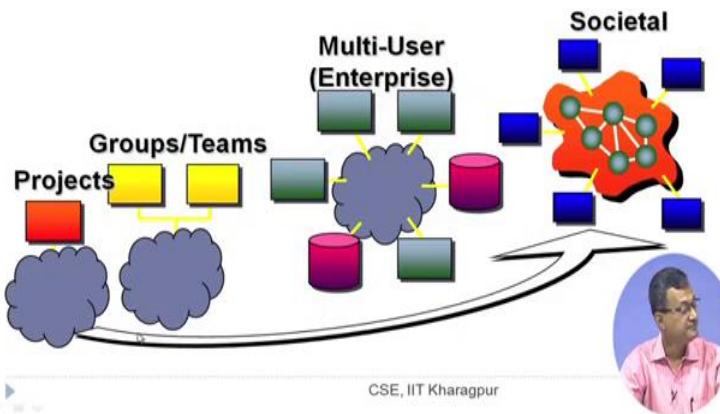


CSE, IIT Kharagpur

So, though I can have these separate type of data like; it may be some data of the street building vegetation, but some sort of a homogeneous integrated data I am looking for because I need to query on this data.

(Refer Slide Time: 14:44)

## GIS Users - Trend



CSE, IIT Kharagpur

And we have seen that the trend is now more of a starting from the project mode; now of a more societal mode; that means, data is available to everybody based on because it end of the date is taxpayers money.

So, based on the considering policy securities, etcetera, but it should be available somewhere ubiquitously.

(Refer Slide Time: 15:10)

## Spatial Data Infrastructure (SDI)

- ▶ “Infrastructure” implies that there should be some sort of coordination for policy formulation and implementation
- ▶ “The SDI provides a basis for spatial data discovery, evaluation, and application for users and providers within all levels of Government, the Commercial sector, the non-profit sector, Academia and by Citizens in general.”

--The SDI Cookbook

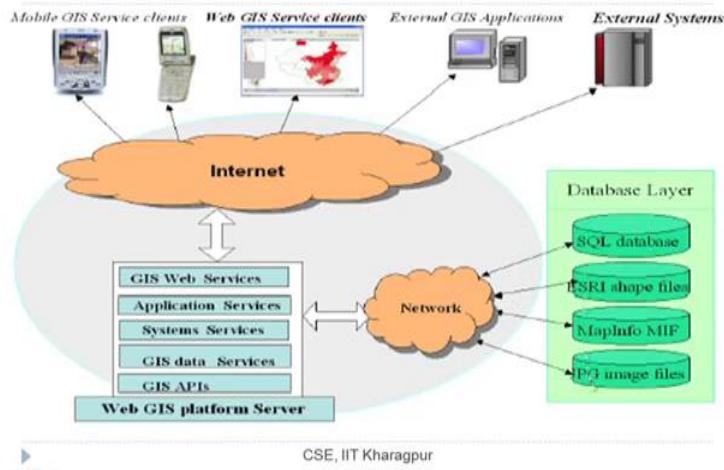
CSE, IIT Kharagpur

So, that may be a cloud may be a solution. Another concept came up or another very popular in advance country or also coming becoming popular or becoming a need in our country also, it is making a spatial data infrastructure. So, looking the data as an infrastructure; so, whenever I want to query on the data I hook into this infrastructure right. So, it is a infrastructure. So, some sort of analogous I can we can look at that our telecom infrastructure, right, you have a there are different-different stakeholders and there have different type of say policies or plans etcetera, but I have a overall communication infrastructure somewhere other if I hook into the things like if I have a Sim card with particular plan then I can access voice data and multimedia type of things right based on again IP as I go.

So, but as such that telecommunication is infrastructure, right. So, whether I can have a spatial data energy infrastructure if it is there, then I can do lot of things like I can do say planning for like what it; what comment does for say development planning thing for developing a particular area sitting on this school to some disasters management also right otherwise you need to always collect data put somewhere and do this.

(Refer Slide Time: 16:38)

## Interoperable GIS – Service driven



So, this sort of infrastructure is the need not only it is coming; becoming true for our country also slowly. So, in other sense what we need there can be different way of accessing the data, there can be different instances of repositories. So, what we harvest here is more of a web services or spatial web services which can take requests from here and talk to the back in databases and serve that it, right.

(Refer Slide Time: 17:05)

## Need for Geospatial Cloud

- ▶ “Huge” volume of Data and Metadata
- ▶ Need of Services and Service Orchestration
- ▶ Evolving Standards and Policies
- ▶ Need for Geospatial Cloud

CSE, IIT Kharagpur

So, what we see that huge volume of data and metadata need of services and service orchestration because the if the if my business rule may not be a just find a shortest

path I can have a things which has a series of services or the processing service to be initiated evolving standard and policies need for geo this may this basically lead to a need for a spatial cloud, right or geospatial cloud.

(Refer Slide Time: 17:32)

## Need of Geospatial Cloud

- ▶ Private and public organization wants to share their spatial data
  - Different requirement of geospatial data space and network bandwidth
- ▶ Get benefits by accessing others' spatial services
- ▶ Less infrastructure and spatial web service expertise needed
  - Easy to port spatial service image to multiple virtual machines
- ▶ Organizations lack this type of expertise
- ▶ GIS decisions are made easier
  - Integrate latest databases
  - Merge disparate systems
  - Exchange information internally and externally



CSE, IIT Kharagpur

So, there are as already we have discussed this point must as to look at. So, private public organization wants to share their spatial data. So, different requirements for spatial data space network bandwidth, etcetera based on the data volume get benefits from other things sharing the things, less infrastructure and spatial service expertise needed, right, there is another important thing, right.

So, if I want to store; if I am a organization of collecting data and storing. So, I require a infrastructure to store there the data are volume us data are requires a different type of bandwidth etcetera and a I need a processing thing and not only that for that I require separate expertise on manpower do that. So, if I have and as we understand that every infrastructure goes off or in the sense has need to be renewed every 3 to 5 years. So, it is sustainability of these things become say extremely difficult. So, instead if I hire infrastructure for my purpose that can be a more beneficial thing, right.

So, sometimes our organization lacks who are more specialized in spatial data lacks in maintaining the infrastructure and GIS decisions are used ubiquitously in various government same government and private things and if we look at this spatial cloud geospatial cloud.

(Refer Slide Time: 18:58)

## Need of Geospatial Cloud (contd...)

- ▶ It supports shared resource pooling which is useful for participating organizations with common or shared goals
- ▶ Choice of various deployment, service and business models to best suit organization goals
- ▶ Managed services prevent data and work loss from frequent outages, minimizing financial risks, while increasing efficiency
- ▶ Cloud infrastructure provides an efficient platform to share spatial data
- ▶ Provide controls in sharing of data with high security provision of cloud.
- ▶ Organizations can acquire the web service space as per needed with nominal cost.



CSE, IIT Kharagpur

So, it support shared resource pooling which is useful for participating organization commons or shared goals as we are discussing a couple of slide back that it may be bumpy the requirement; that means, I need to scale up scale down on my resources. So, it is a say shared resource pooling will be there and choice of various deployment service business model to base suit organization requirement managed service prevent data and work loss from the frequent outage etcetera like.

So, if it is having more; many service and which may minimize my financial constraint etcetera provide controls in sharing data with high security provision of the cloud as cloud also looks at the security features and looking aggressively on the security features. So, and security always come with a lot of cost. So, all we can leverage on the things and it can be hum much what we say economical to use those type of things and organization can acquire web service space as per their need. So, if I need only this sort of spatial services, I only harvest those services, right, I or if I need this type of data I only look for this data.

(Refer Slide Time: 20:18)

## Cloud Computing

NIST's (National Institute of Standards and Technology) definition:

- ▶ “*Cloud computing is a model for enabling convenient, on demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.*”

CSE, IIT Kharagpur

So, it is customized based on the need from the infrastructure right. This is the NIST definition I just repeated.

(Refer Slide Time: 20:23)

## Cloud Advantage

- ▶ Scalability on demand
  - ▶ Better resource utilization
- ▶ Minimizing IT resource management
  - ▶ Managing resources (servers, storage devices, network devices, softwares, applications, IT personnel, etc.) difficult for non-IT companies
  - ▶ Outsourcing to cloud
- ▶ Improving business processes
  - ▶ Focus on business process
  - ▶ Sharing of data between an organization and its clients

CSE, IIT Kharagpur

So, already we know and cloud advantages also are known to us that scalability on demand minimizes IT resources improving business processes.

(Refer Slide Time: 20:33)

## Cloud Advantage (contd)

- ▶ Minimizing start-up costs
  - ▶ Small scale companies and startups can reduce CAPEX (Capital Expenditure)
- ▶ Consumption based billing
  - ▶ Pay-as-you-use model
- ▶ Economy of scale
  - ▶ Multiplexing of same resource among several tenants
- ▶ Green computing
  - ▶ Reducing carbon footprints

CSE, IIT Kharagpur

And all these advantages minimize startup cost consumption base building or pay as you go billing pay as you go model economy of scale, green computing these are all is need for these spatial data infrastructure or spatial data management and type of thing. So, geospatial data infrastructure.

So; that means, it is it is somewhere it is suited for this sort of things, right.

(Refer Slide Time: 21:00)

## Cloud Actors

- ▶ Cloud Service Provider (CSP) or Broker
  - ▶ Provides with the infrastructure, or the platform, or the service
- ▶ Customer
  - ▶ May be a single user or an organization
- ▶ Negotiator (optional)
  - ▶ Negotiates agreements between a broker and a customer
  - ▶ Publishes the services offered on behalf of the broker
- ▶ SLA Manager/Security Auditor (Not present in current clouds)

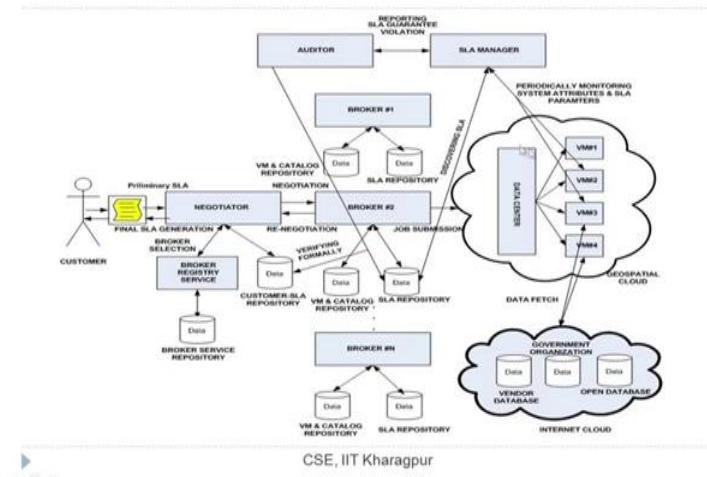


CSE, IIT Kharagpur

And this also we have seen that there are who are the cloud actors like one we have cloud service CSP or broker, customer is the other end negotiator is can be optional that negotiate between cloud and SLA manager of security auditor.

(Refer Slide Time: 21:18)

### Typical Geospatial Cloud Architecture



So, if you look considering all those things and if we look at typical cloud architecture for geospatial data this architecture may be true for any type of other services, but if we look at the geospatial data where the geospatial or spatial services are put into place. So, these overall things come into play. So, there is a once a customer or a user or a service consumer comes into play.

So, there is a negotiator which there can be number of brokers where which broke where which basically broker acting as agent for different data or different sets of data there can be the same data may be handled in 2 repositories may so happen and we have different data centers which gives VMs and there are different organizations which launch their data and services on this VM and these are being negotiated and connected with the end user they are other than these we require SLA manager to see that that this is SLAs or survey level agreements are not violated or things and there can be a auditor or auditor to see that whether this sort of policies in terms of SLAs, there can be security auditors security policies etcetera are being maintained in the this type of situation.

So, I require a brokering service or who broke for the things there are different data centers which hosts this spatial data of various organizations and there can be SLA there

will be SLA manager and auditor which takes care of the overall management things, there can be other component of the thing other different components there can be separate auditing services, etcetera, but this is the overall broad infrastructure.

(Refer Slide Time: 23:19)

### Cloud as Service Provider

- Collection of Enterprise GIS (eGIS) Instances
  - **Resource Service** – resource allocation, manipulation of VM and network properties, monitoring of system components and virtual resources
  - **Data Service** – maintains persistent user and system data to provide a configurable user environment
  - **Interface Service** – user visible interfaces, handling authentication and other management tools.

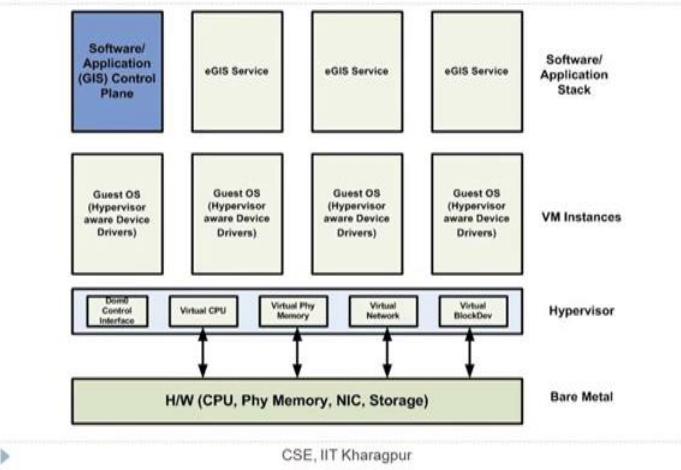
CSE, IIT Kharagpur

So, if I look at some sort of interface GIS or what we say some sort of a realization of this SDI which gives which gives spatial services what do you mean by this interface GIS is that it itself does not have any data per se, but it basically allows the consumer or the customer to connect with different service provider.

So, it is some sort of a cloud or some sort of a middle tire which connects those things, right. So, resource service resource allocation manipulation data services interface services all can be provided.

(Refer Slide Time: 24:02)

## Geospatial Cloud



CSE, IIT Kharagpur

And we can see that a typical structure we have different instances of this there can be. So, this is my backbone hardware over there we have this hypervisor which emulate different virtual machines there can there in guest OS over that I can have different type of instances of this spatial cloud, right or instances of these spatial web services which are there. So, I can say that if it is say that spatial web services for a particular organization may particular state may be particular district can be instantiated on the things right and then I work on the thing like if you if we try to look at a some sort of analogy may not be very good analogy.

So, if I say that if I when I am using somewhat processing service on the cloud say Google docs. So, what we are doing we are basically instantiating my domain on that particular Google doc right somebody else is again instantiated things, but at the back end, it is the same it is the basically the Google infrastructure we are using where the hypervisor and it is basically software as a service. So, we are at a much higher level, but it is instantiating my thing. So, here whether I can have spatial service instantiated for my particular work.

(Refer Slide Time: 25:28)

## Geospatial Cloud Model

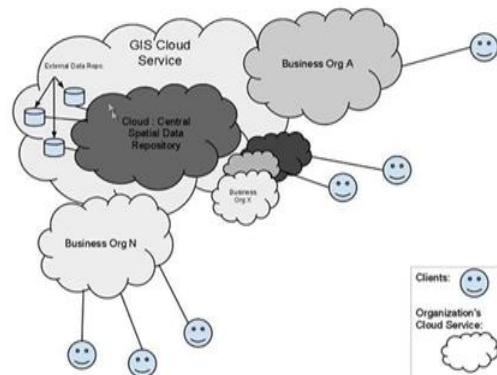
- ▶ Web service is the key technology to provide geospatial services.
- ▶ Need to integrate data from heterogeneous back-end data services.
- ▶ Data services can be inside and/or outside the cloud environment.
- ▶ Data services inside cloud can be run through PaaS service model.
- ▶ Using PaaS makes load balancing, distributed replica and dynamic scaling transparent.

CSE, IIT Kharagpur

So, if we look at the geospatial cloud model. So, web services is here also key technology to provide spatial services need to integrate data and heterogeneous back end data services service can be set data service can be inside or outside the cloud there can be that some of the services can be outside can be run through PaaS service model using PaaS makes load balancing etcetera we can make scaling transparent.

(Refer Slide Time: 25:55)

## Geospatial Cloud – Typical Scenerio



CSE, IIT Kharagpur

So, what way of looking at it that a typical scenario where different organization are there and I have a one coordinating cloud which has the central repository which takes care that where the organization.

(Refer Slide Time: 26:10)

### **Geospatial Cloud**

- ▶ Need to integrate data in an unified format.
- ▶ Performance Metrics: computation power, network bandwidth.
  
- ▶ Data sources:
  - Central Data Repository within the cloud.
  - External Data Repository providing data as WFS,WMS services.

CSE, IIT Kharagpur



How they are working and type of things, right where the data is there. So, that we have already discussed need to integrate data in a unified format or homogeneity it needs to be instead or interoperated instead performance metric computation power network things etcetera need to be judged and there can be various data sources.

(Refer Slide Time: 26:29)

**Experimental GeoSpatial-Cloud  
@IITKgp**

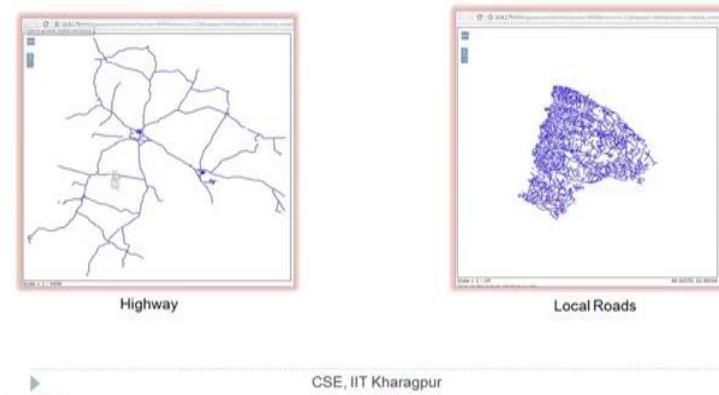
CSE, IIT Kharagpur

We just saw some snapshot that we made it experimental geospatial cloud I should say in our; at IIT Kgp on our own experimental cloud that Meghamala where we have seen that which were and so on open stack open source cloud platform and it is a in house cloud used its a private cloud to IIT Kharagpur and over that we have did some experiment.

So, as such we are not showing that some performance, but showing that what sort of things may be there.

(Refer Slide Time: 27:01)

### Service Integration for Query in Cloud (Case Study 1)

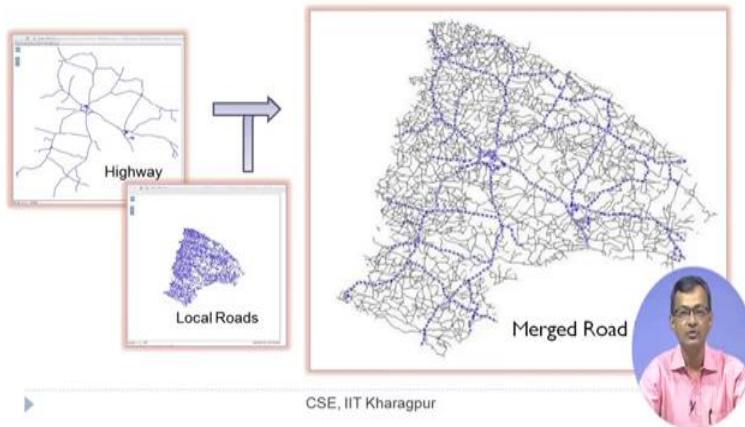


CSE, IIT Kharagpur

So, this is a particular area is a area of a Bankura districts which is a district in West Bengal. So, if I have a highway which is supplied by say highway national highway authority and on the other hand if I have a local road network which is a more handle by the state or the district authority then how to query something which involve this now if you the though the image may not be clear this are running into separate VM, it is something 1 dot 74 here, 1 dot 75 on the fly, I want to have a processing service which integrate, right.

(Refer Slide Time: 27:37)

### Service Integration for Query in Cloud

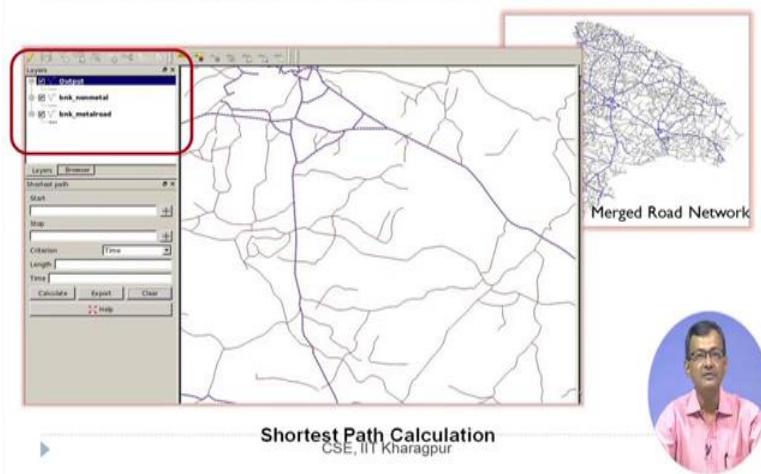


CSE, IIT Kharagpur

So, taking this 2, I need I basically have a merged road network right or in other sense I have a 2 services of the road, I am pulling and having a merged road or having the national highway what that and there we can basically.

(Refer Slide Time: 27:52)

### Service Integration for Query in Cloud

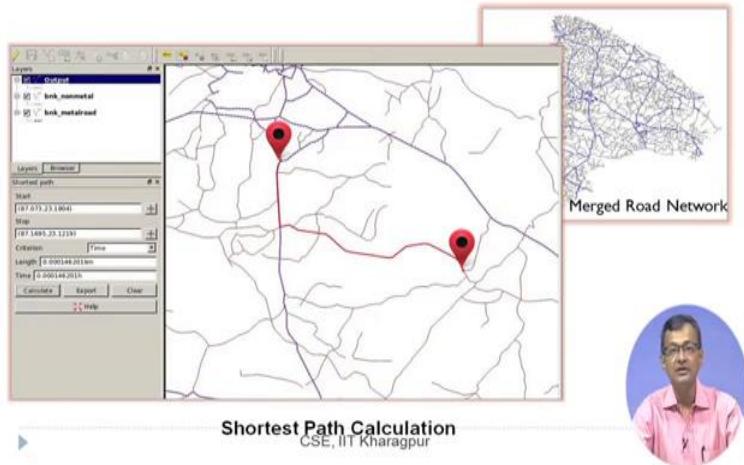


Shortest Path Calculation  
CSE, IIT Kharagpur

Now, query for a shortest path on these things, right which takes care both these national highway and the state highway like if it is my 2 location and then I want to find out the shortest path it does on the on this combine things.

(Refer Slide Time: 28:02)

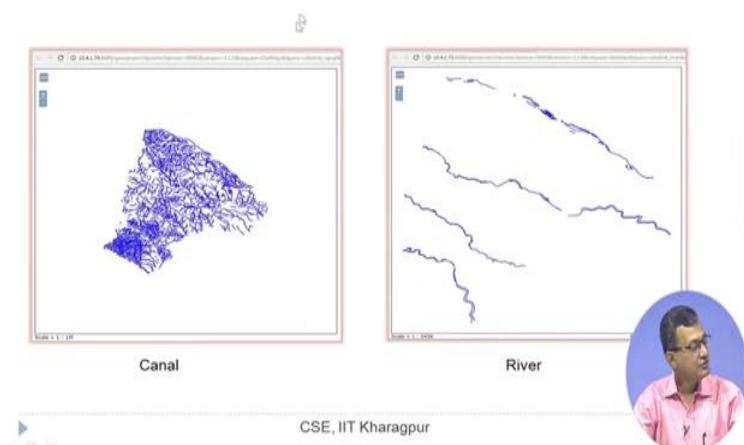
### Service Integration for Query in Cloud



So, though the application may be trivial if not trivial it may begin well known application, but see 2 separate repositories we can pull on a service mode and which are running on 2 different VMs and pull on a service mode and basically my business rule is finding the shortest path, right, I am not neither interested in the whole data of West Bengal or whole data of India of road data, I am interested in that shortest path and which are the locations.

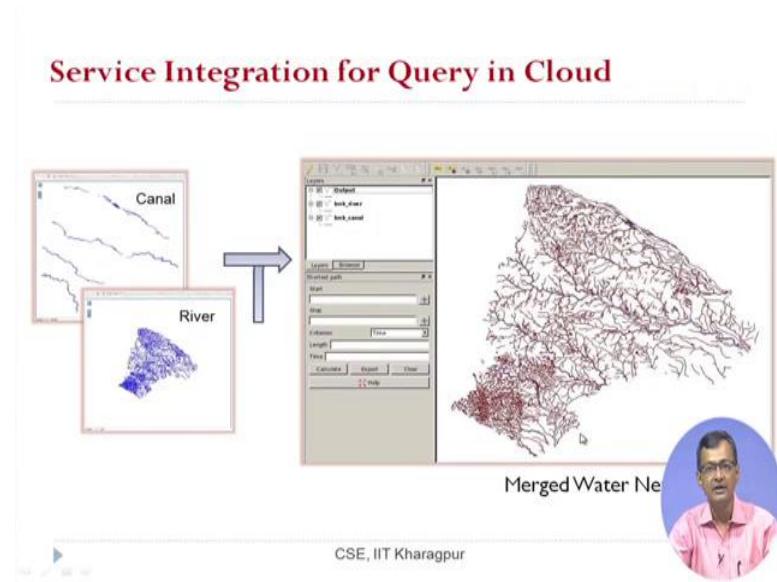
(Refer Slide Time: 28:44)

### Service Integration for Query in Cloud (Case Study 2)



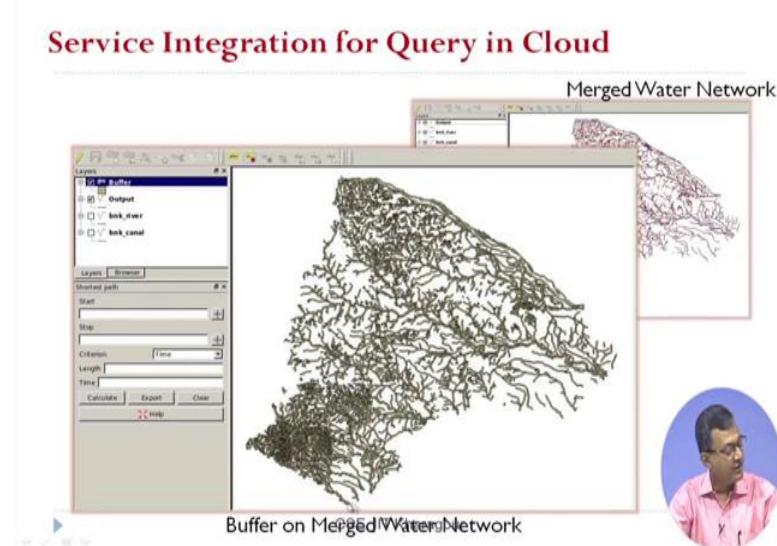
So, this is one sort of thing. So, other thing I can have again 2 sort of data where I have one a canal type of things on again same Bankura district and a river data which can be in a different thing.

(Refer Slide Time: 28:55)



Now, I want to have a merged water network which integrate this canal along with the river again these 2 are in the separate VMs or separate cloud instances I have a merged data and I want to buffer on this merged data, right.

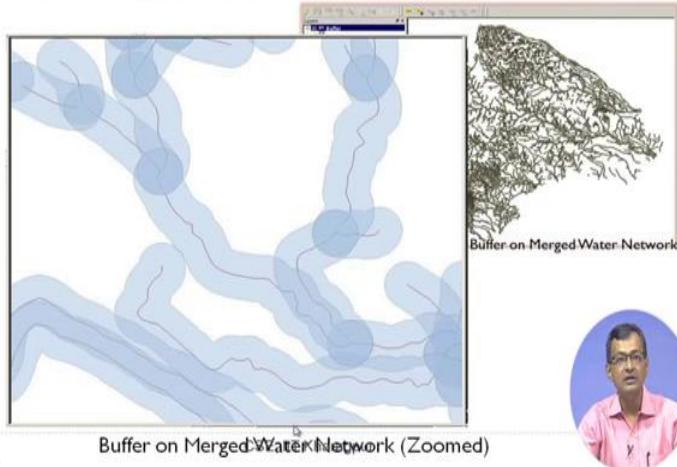
(Refer Slide Time: 29:09)



So, try to see that during heavy rain or flooding; what are the things.

(Refer Slide Time: 29:13)

### Service Integration for Query in Cloud



Which if the business rule is that within 500 meter it is vulnerable, then; what is the vulnerable area of this particular buffer zone. So, again; I have 2 separate data instances which are integrated on the fly and gives this query, right. So, here we have experimented on a experimental geospatial cloud, we have installed over our Meghamala thing.

(Refer Slide Time: 29:39)

### Challenges in Geospatial Cloud



CSE, IIT Kharagpur

So, there are several challenges here other than which are typically something are typically for the spatial data which are not there in other type of what we say non spatial data type of things.

(Refer Slide Time: 29:55)

## Challenges in Geospatial Cloud

- Implementation of Spatial Databases.
- Scaling of Spatial Databases
- Need to be Multi-Tenant
- Policy management among the tenants.
- Geographically situated Backups
- Security of Data

CSE, IIT Kharagpur

So, implementation of spatial database is one of the challenges, it requires a lot of not only space need to understand the data semantics or scaling of the spatial databases right need for multi tenant thing policy management among these different tenants geographically situated backups if you want to keep data security some of the channels are also there for other type of things some are typically for this type of spatial.

(Refer Slide Time: 30:24)

## Interoperability Issue

- ▶ Exchanging and processing of geospatial Information requires interoperability on different levels:
  - ▶ **Data Level Interoperability** ensures the ability to "consume" the information
  - ▶ **Service Level Interoperability** ensures the ability to exchange / obtain the information to be "consumed"
  - ▶ **Security Level Interoperability** ensures the ability to the above in a reliable and trustworthy fashion
- ▶ Implementation of all levels can be done by using standards from the OGC and other bodies



CSE, IIT Kharagpur

And also we have a major challenge of interoperability, right, data level interoperability like one data should match with other, while we are doing; say buffering type of things and things it can be a service level inerrability like 2 services can talk to each other security level in interoperability like; when I have different policies on different data sets then what should be the security level inter interoperability.

(Refer Slide Time: 30:49)

## Geo-Cloud – Major Security Concern

- ▶ Multi-tenancy
- ▶ Lack of complete control - data, applications, services



CSE, IIT Kharagpur

So, major security concern is the multi tenancy is already for other things and lack of complete control of data application services.

(Refer Slide Time: 30:59)

## Concerns (contd...)

- ▶ Which assets to be deployed in the cloud?
  - ▶ Identify: data, applications/functions/processes
- ▶ What is the value of these assets?
  - ▶ Determine how important the data or function is to the organization
- ▶ What are the different ways these assets can be compromised?
  - ▶ Becomes widely public & widely distributed
  - ▶ An employee of the cloud provider accessed the assets
  - ▶ The processes or functions were manipulated by an outsider
  - ▶ The info/data was unexpectedly changed
  - ▶ The asset were unavailable for a period of time

CSE, IIT Kharagpur

There are several other concerns which assets to be deployed on cloud what is the value of these assets what are the different ways this asset can be compromised etcetera. So, there are different types of other concerns which we need to look at in; if we look at the overall spectrum. So, this spatial or geospatial domain finds a immediate applicability or a huge applicability on this sort of a cloud thing. So, one of the good use cases which are which not only in our country in different organization in different countries how to make this heterogeneous data talk to each other for different type of citizen centric development planning and disastrous management is becoming a popular aspects of special cloud and this is a could use case.

Thank you.

**Cloud Computing**  
**Prof. Soumya Kanti Ghosh**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**

**Lecture – 36**  
**Introduction To DOCKER Container**

Hello, so welcome to the course on cloud computing. So, so far what we were discussing is about that how this cloud computing evolved, what are the different advantages and challenges in cloud computing? What are the different associated technologies to this cloud computing aspects? So core architecture and what are the different other type of technologies which came up with those things.

So, as we have seen it may not be a totally new technology or new innovation as such, but it is trying to look at all these innovations, all these new approaches and coming up the things. Another new development what we will look today is a container technology. So, what we are finding; one way as we have seen that there are challenges in infrastructure build up, there are challenges in platform build up, there are challenges in software build up.

That is why we try to migrate from our in house installation to some service providers installation. So, typically as we have discussed on XaaS type of services; infrastructure platform and software, these are the different type of services. Another problem which is coming up nowadays is; or rather with the development of software and other developments like there is a redeployment or reconfiguration or recompilation of the software whenever we want to ship one package from one environment to other.

This is one of the major motivation towards these; these days we are having devices, which are pretty resourceful; say your mobile device or some notepad or laptop or even these days smart watches. So, all those things are resourceful and able to run different applications. So, one way is looking at that sort of things; other way another aspects of it is; we are having variety of applications development across the things. So, this type of portability of one application from one environment to another is becoming a major challenge.

So, there were the container technology came into picture; where we will be able to bundle up the application along with these associated dependencies into a container and push it as a single container to the other type of environment; I mean the amount container fits, your applications start working on that.

So, cloud in such cases can be a platform to host this container; though there are some literature or in some forum people say this container technologies is something a contended to the cloud. But what we see these days not exactly a container to or a something which is fighting with cloud; rather we can cloud can act as a platform to host these container and basically cloud can enhance its capability; by adapting this container technology.

One such container which has become a very popular is the Docker; a open source container which has become a Docker. So, what will see; we will see a brief introduction to this Docker container. It will help us in understanding this how this portability of different systems and software will be there and also help us in looking at this container technology with a background of a cloud computing.

(Refer Slide Time: 04:24)

## Docker

- Docker is a container management service (initial release: March 2013)
- Main features of Docker are *develop, ship and run anywhere*.
- Docker aims at facilitating developers to easily develop applications, ship them into containers which can then be deployed anywhere.
- It has become the buzzword for modern world development, especially in the face of Agile-based projects.

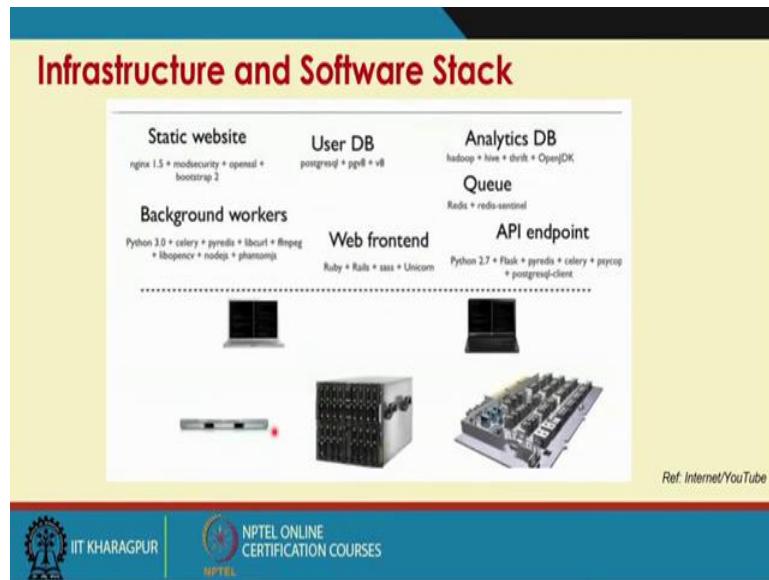
Ref: <https://www.tutorialspoint.com/docker/>

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, if you look at that Docker; it is somewhere initial release where March 2013. So, it is a container management service; main feature of Docker are develop, ship run anywhere. So, this key words if we can see like; develop; ship it and run anywhere. So, irrespective of where it is, it should be able to run; this is basic philosophy of Docker which tells

everything like you develop once and ship to any environment and run to this. It can be any your desktop environment, it can be a android environment, it can be a IOS environment, it can be a say any cloud environment, it can be open stack, it can be azure, it can be blue mix or Amazon; anything.

(Refer Slide Time: 05:47)



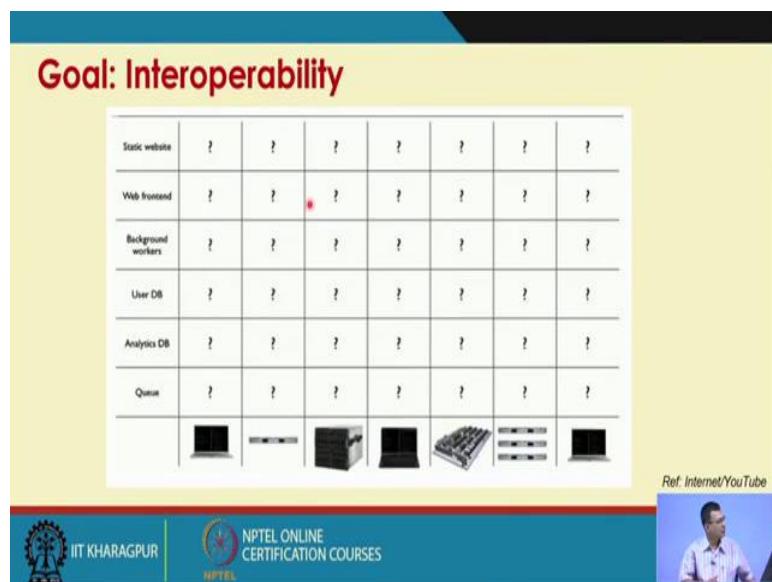
So, that is exactly the container things trying to say. So, the Docker aims at facilitating developer to easily develop applications, ship them into containers which can then be deployed anywhere. It has become the buzz word for modern world development; especially in the face of agile based projects. So, this is the buzzword these days to look at the these type of things. We will try to make a quick analogy of the things and try to understand what is there in this type of things.

Now, if you look at how our computing paradigm develop. So, it was somewhere a system like it maybe a somewhere; something like a laptop or even desktop or some specialized board or a chassis with blade servers and any type of things, even if I can look at some devices like network devices, where we basically build some applications for them. Now, if I am running something on a particular laptop and then want to run it on a server thing, then in most of the cases I need to recompile the things, I need to align the things or redevelop; not that redevelop the whole thing, but I need to redevelop a major portion of the things or recompile the thing for the server. I want to do for another type of system; I need to again redevelop the things.

So, what is happening that there is a lot of man power uses into the things. Whereas, on the other side; so one side there is a large means or there is a everyday we are coming or means periodically you are coming with new systems and software's and ways and type of things. On the other hand we have several applications which are coming up on the other hand. Some maybe static website; something like background worker like python and so and so forth, user DB, analytics DB, queue, API endpoints, web frontend and endless things are being developed.

Now, if you look at that for everything; if you want to have. If you want to sip a application from one to another and if we want to this sort of recompilation or reconfiguring the things it is a hell of a job.

(Refer Slide Time: 08:00)



So, what we try to; when look at that interoperability or portability; what we try to do? We try to set up a matrix. Like say static website; whether it will run on this particular device, whether you run on this device and so and so forth; it is a user DB and type of things.

So, if I put tick on this matrix then this is there. So, our major objective is that with minimal investment in terms of man power, skill, infrastructure and software and so and so forth; I should be able to have all ticks on this matrix; anything can run anywhere, type of situations. So, that is a big challenge to add here, but never the less; the whole computing world wants to achieve.

This leads to commercialization of a product in a better way, marketability of a product in a better way and rather in some cases like if you look at sensor based technologies; if the sensor applications runs only on a particular type of smart phone and or a particular type of OS and it is not running to the other things, then your capability also decreases. I could have captured through different type of mobile devices, but that portability becomes a major challenge.

So, these are some of the things which has driven that whether I can have a; somewhere some intermediate mechanisms which allows me to transfer this data.

(Refer Slide Time: 09:37)



Now, if we try to look at some analogies; look at shipping. I want to ship some product from one to another. So, one side these products are there like; there can be variety of products. Like it can be a box, it can be a car, it can be some barrel, it can be even a piano and drums and servers and etcetera. So, these are the things I want to ship from place A to B and it can be shift in varies form.

It can be in somewhere some trucks and sort of things, somewhere can be dumped into a racks, some fork lifter to lift on the things, somewhere some locomotive or trains being manufactured out of factory and see it and through crane and put on the ship for shipping from across the continent.

So, there are a variety of products and variety of techniques and technologies which are involved in shipping these products. Unless I standardize; like if you think that the type of mechanisms required for a server will differ from how you do for a piano or what you do for a car. So, unless I make a standardized way of doing this business; then things will not work.

(Refer Slide Time: 11:00)

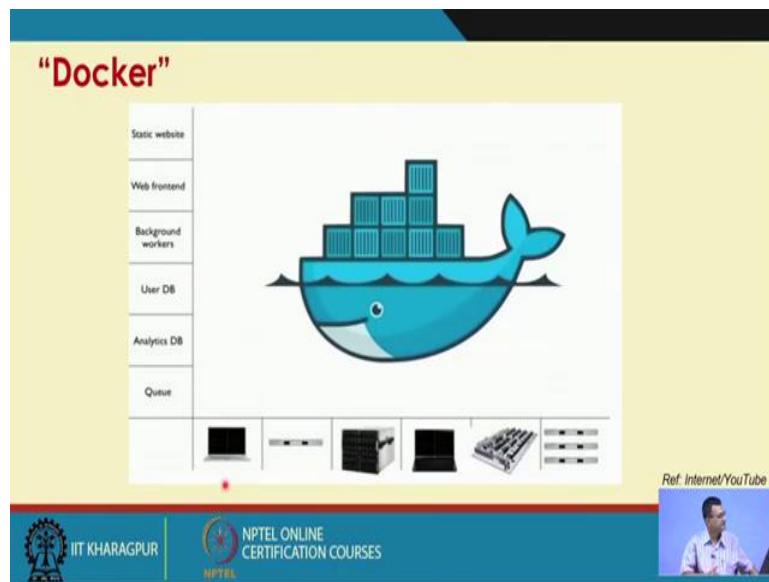


So, what they found out? They found out a concept for this sort of a docking station or this sort of a chamber where if it fits in, it can be shipped. So, what we do; any product, we put into a this sort of docking station or this type of a box. And this box has a standardized size, standardized mechanism of holding and it can be either put on rack, put on fork lifter, put on crane, put on ship; can be taken out from factory or inside the factory or put on the trucks and carry it; these are very popular.

Now, I made this shipping thing to any type of mechanism to transport; by making a intermediate mechanism call shipping containers. Whether, I can make this thing feasible for our applications, for our software applications. Now this container is a standardized size, so if somebody can hold the container; it can hold these products. So, what it does; the inside the container is more on the product which dictates, outside the container the infrastructure which dictates. This outside the environment of the container is more inclined towards this infrastructure by which it is shipped or where it is stored and type of things.

Inside the container is the way where the product is stored; like piano may have a different requirement of environment than car than a server systems. So, the same type of conceptually, the same type of philosophy is being carried to this; when we talk about container platform, when we talk about Docker.

(Refer Slide Time: 13:11)



So, same thing what we say it is a Docker service. We have different applications and there are different devices and if I can dock it in somewhere other; which is standardized and if these devices are able to handle this container or Docker container then my application will run.

So, what inside that container is there; based on the applications all these related libraries, binaries, dependences etcetera are contained in this thing; externally where which infrastructure it is running that is important. So, this sort of philosophy helped in as we staring that you develop and ship to anywhere; develop, build, ship anywhere.

So, if this type of technology is becoming extremely popular and becoming a defect standard for this and what we feel that is necessary to discuss. Because most of the cases what we are doing in a cloud, we are trying to run different applications within this cloud. Like, if you look at say I want to run in a particular banking application or we talk about some special web service applications.

So, what it tries to do? It tries to build a container; particular container class in a sense or a club of services which need to be shipped or needs to be run at different environment. Let it be Azure cloud, let it be open stack cloud, let it be any other sort of cloud like IBM Blue Mix or Amazon or Google cloud platform any type of platform. My basic requirement is there; I should not develop, redevelop for every environment and as well it should run, if the resource permits on my smart phone or my desktops systems or server and so and so forth.

So, this sort of portability of the things; this container brings into picture.

(Refer Slide Time: 15:40)

## Docker - Features

- Docker has the ability to reduce the size of development by providing a smaller footprint of the operating system via containers.
- With containers, it becomes easier for software teams, such as development, QA and Operations to work seamlessly across applications.
- One can deploy Docker containers anywhere, on any physical and virtual machines and even on the cloud.
- Since Docker containers are pretty lightweight, they are very easily scalable.

Ref: <https://www.tutorialspoint.com/docker/>

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, if we look at again that the basic features; so, Docker has the ability to reduce the size of development by providing a smaller footprint of the operating system via containers; so, in the container you have a smaller footprint of the OS. With containers, it becomes easier for software teams; such as development, QA teams operations to work seamlessly across applications. So, they are may be at different centers; even for a one organizations and it can work seamlessly because finally, it should fit into the container in being shipped to the other things.

One can deploy Docker containers anywhere; on any physical and virtual machines and even on cloud. So, it is a Docker containers can be deployed anywhere; practically anywhere or any physical or virtual machines in VMs or physical servers and even on cloud. Since Docker containers are pretty lightweight, they are easily scalable; that is an

important factor. So, this container; this Docker container is light weight and as it is light weight then scalability is much easier.

If there is a heavy weight stuff, then there is resource requirement will be much higher and anything portability is becomes or scalability or scaling up or especially scaling up becomes extremely difficult; so, as this is light weight, this is much easier.

There are different variants or what we say different flavors of things or components. So, Docker for Mac, it allows one to run Docker container on Mac OS. Docker on Linux, it allows one to run Docker container on Linux OS. Docker for windows; Docker engine, it is used for building Docker images and creating Docker containers. Some sort of over the bare metal OS; you have the Docker engine, over this Docker containers are placed; some sort of analogy with the hypervisors of a particular virtualization system.

Docker hub; this is a registry the important thing registry which is used to host various Docker images. So, this Docker hubs you will find various Docker images; if you are a developer, you can basically submit in the Docker hub, which can be used by others to use it. Docker compose; this is used to define application using multiple Docker containers. So, it is a sort of composition service which is used to define applications using multiple Docker container. So, these are different components; there are various others or different flavors of these Docker or different components of this Docker.

(Refer Slide Time: 18:42)

## Traditional Virtualization

- Server is the physical server that is used to host multiple virtual machines.
- Host OS is the base machine such as Linux or Windows.
- Hypervisor is either VMWare or Windows Hyper V that is used to host virtual machines.
- One would then install multiple operating systems as virtual machines on top of the existing hypervisor as Guest OS.
- One would then host your applications on top of each Guest OS.

The diagram illustrates the traditional virtualization architecture. It consists of four horizontal layers stacked vertically. From bottom to top, the layers are: 1) Server, 2) Host OS, 3) Hypervisor, and 4) Guest OS. Within the Guest OS layer, there are three separate boxes, each labeled 'App' (Application). This visualizes how multiple virtual machines (each with its own Guest OS and applications) run simultaneously on a single physical server.

Ref: <https://www.tutorialspoint.com/docker/>

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

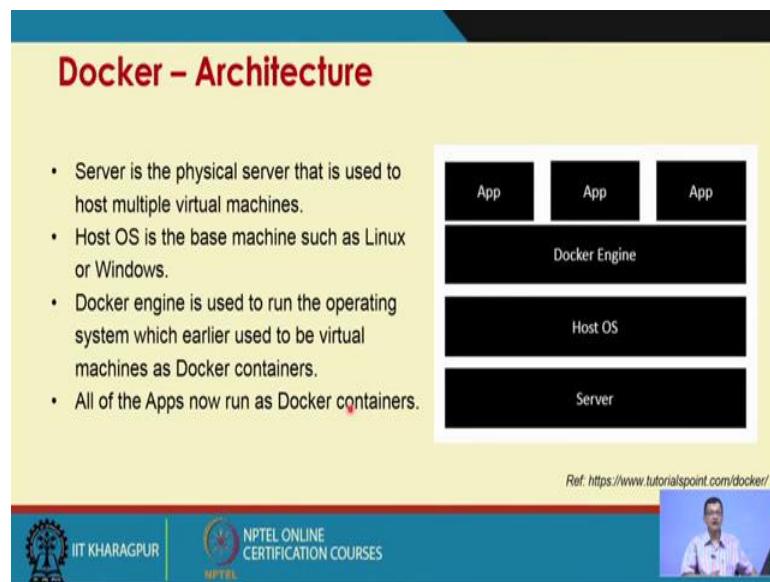


So, if we look at the traditional virtualization; so, we have that underlining server or bare metal or the backbone systems; physical systems. Over that there is a host ways which basically provide this service to this server, over that we have that hypervisor or VMM or Virtual Machine Monitor and this helps us in generating different VMs. And every VMs can have their own guest OS and over the guest OS; different applications runs.

So, this is the basic philosophy of virtualization already we have seen. So, the server is the physical server that is used to host multiple virtual machine. Host OS is the base machine such as Linux or windows, hypervisor as either some sort of hypervisor; it can be VMWare, Xen, KVM and type of things or commercially VMWare or windows hyper V that is used to host virtual machines.

One would then install multiple operating system as virtual machines on the top of the existing hypervisor as guest OS. So, what we do in case of a IAAS type of cloud and then host your application on the top of each guest OS. So, this is the way normally a traditional virtualization work.

(Refer Slide Time: 20:08)



A variant of this Docker architecture is that; it has the server OS; case two s and then we have the Docker engine. So, over the Docker engine we can run different apps; so, it is going to be light weight and of course, you may not do very large applications, but never the less if your application is not so demanding; you can basically deploy out here.

Server is a physical server that is used host multiple virtual machines. Host OS is the base machine such as Linux or windows.

Docker engine is used to run operating systems; which is earlier used to be virtual machines at Docker containers. So, it is running operating systems which is earlier used as a virtual machines as Docker containers. And then finally, all apps now run on the Docker containers; so, it has now Docker containers where this different apps run on this Docker container. So, what we see there is a; though the philosophically maybe same sort of aspects is there, but there is some difference with this traditional virtualization.

(Refer Slide Time: 21:30)

**Container?**

- Containers are an abstraction at the app layer that packages code and dependencies together.
- Multiple containers can run on the same machine and share the OS kernel with other containers, each running as isolated processes in user space.
- Containers take up less space than VMs (container images are typically tens of MBs in size), and start almost instantly.

Ref: <https://www.docker.com/>

 IIT KHARAGPUR |  NPTEL ONLINE CERTIFICATION COURSES

So, just to recap that containers are abstraction at the app layer that packages code and dependencies together. So, it is a; what is this container including Docker container? It is an abstraction at the app layer; that application layer that packages or bundle code and dependency together; so, it is together bundled. Multiple container can run on the same machine and share the OS kernel and other containers each running on isolated processes in the user space, so that is also possible.

Container takes up less place then virtual machines; container images typically tens of MBs in size and start almost instantaneously as that is exactly what we are discussing, it is a low weight and we can much less resource hungry than virtual machines and it can start instantaneous as it is a low weight.

(Refer Slide Time: 22:38)

**Container (contd...)**

- An **image** is a lightweight, stand-alone, executable package that includes everything needed to run a piece of software, including the code, a runtime, libraries, environment variables, and config files.
- A **container** is a runtime instance of an image—what the image becomes in memory when actually executed. It runs completely isolated from the host environment by default, only accessing host files and ports if configured to do so.
- Containers run apps natively on the host machine's kernel. They have better performance characteristics than virtual machines that only get virtual access to host resources through a hypervisor. Containers can get native access, each one running in a discrete process, taking no more memory than any other executable.

Ref: <https://www.docker.com/>

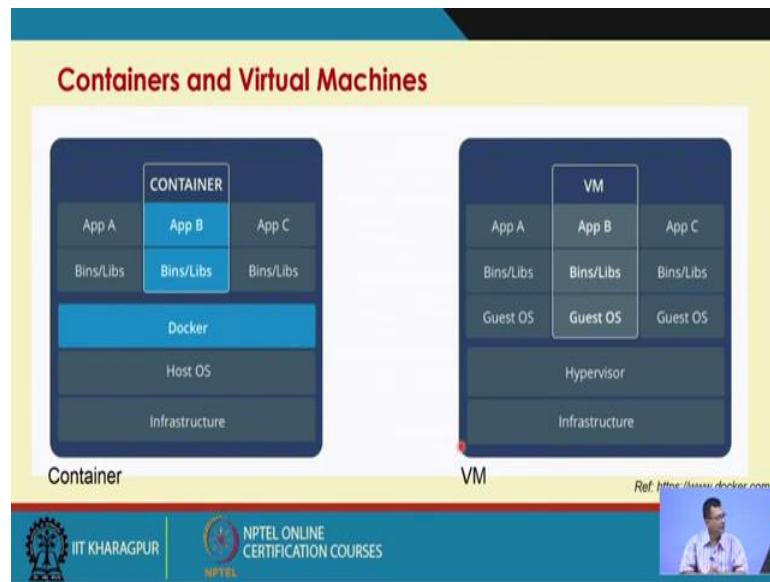
 IIT KHARAGPUR |  NPTEL ONLINE CERTIFICATION COURSES

So, there are other few concepts which comes back to back. One is that image so we are talking about images and type of things. So, an image in this context is a light weight, stand-alone, executable package that includes everything needed to run a piece of software, including code, runtime, libraries, environment variables, config files etcetera; so it is important.

So, it is a light weight standalone executable package that includes everything needed to run that particular software package or software including code, runtime, libraries, environment variables, configuration files everything. A container is a runtime instance of an image what the image becomes in memory when actually is executed. So, what is there? so, I have a bundle things once it start running; it is a instantiation of this image that is exactly the container, it runs completely isolated from host environment by default only accessing the host file and ports if considered to do so. So, it is independent of the host environment.

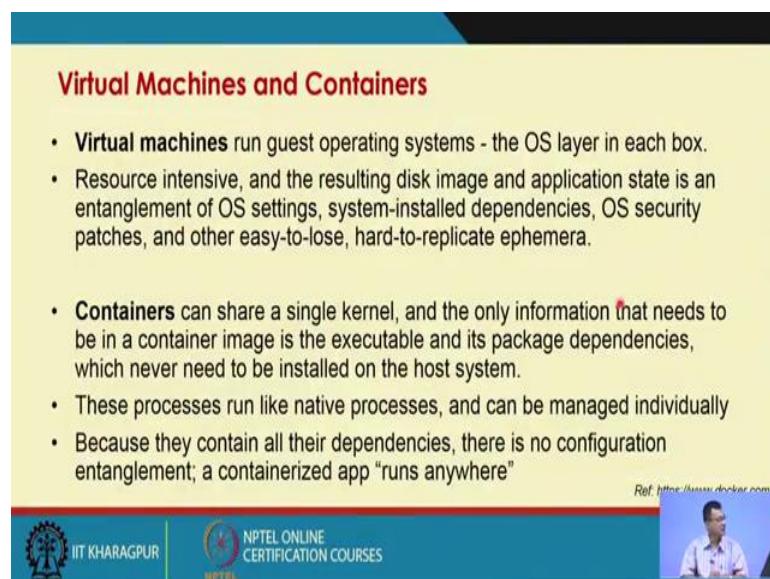
So, other things there container runs apps natively on the host machines kernel; they have better performance characteristics than virtual machine, that only get virtual access to the host resources through a hypervisor. So, it runs directly on the host machines kernel, containers can get native access each one running in a discrete process taking no more memory than any executable. So, it is not only light weight; it has performance wise also instantaneous.

(Refer Slide Time: 24:32)



So, if you look at containers and virtual machines; so, we repeat that things; so we have the underlining infrastructure of the server base. Host base Docker and then we have that container, it has the applications with other dependencies, whereas in case of a hypervisor or a virtual machine over the hypervisor, we have guest OS and rest of the things.

(Refer Slide Time: 25:05)



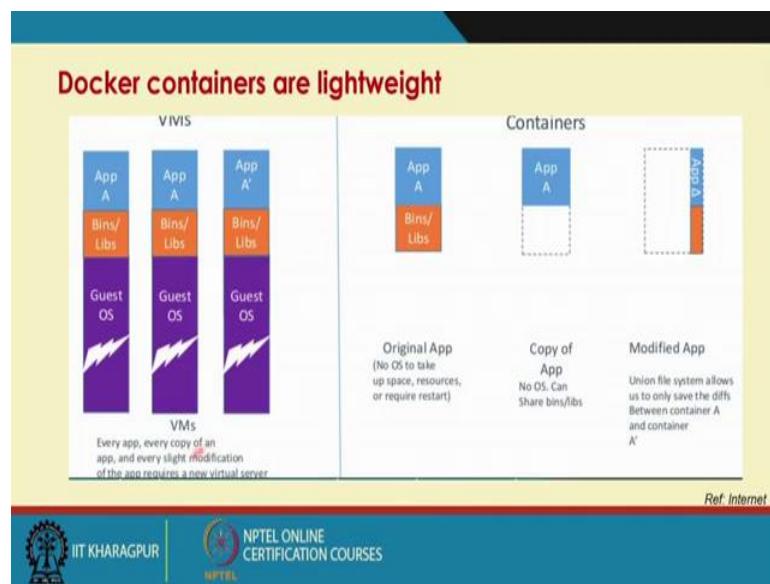
So, that is the difference; so, virtual machine run guest operating system; OS layer in each box. Resource intensive, so as it is requires more resources resulting disc image and

application state is entangled with the OS settings, system-installed dependencies, OS security patches and so and so forth. Whereas, container can share a single kernel and only information that needs to be in a container image is the executable and its package dependence.

So, what are the different other package dependencies which never need to be installed on the host OS. Because it is bundled together; these processes run like native processes and can be managed individual every container. Because they contain their dependency, there is no configuration entanglement and so containerized applications runs anywhere. So; that means, I do not have any so to say any binding with this host machine.

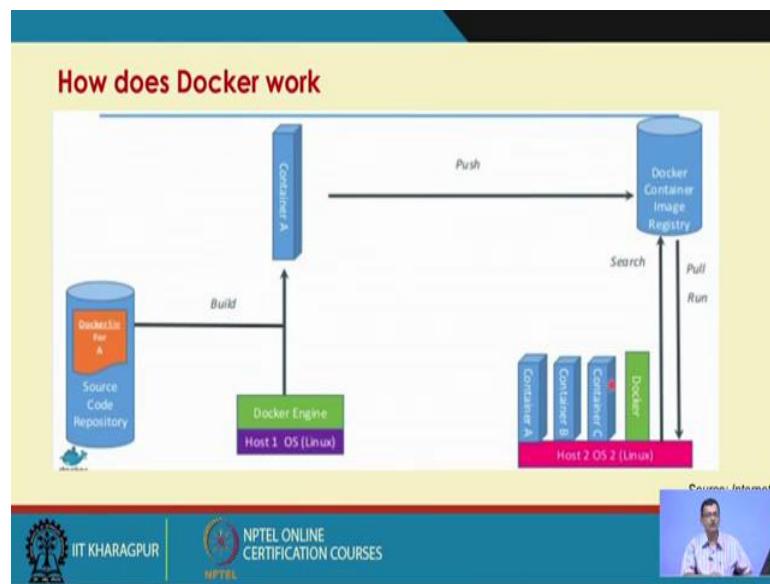
So that means, I isolate this my container based package with the host. So, it increases the portability and can be run any other system. So, same thing; so, if there is a running on a VM, then we have every app, every copy of an app, every slide modification requires a some new virtualized environment.

(Refer Slide Time: 26:35)



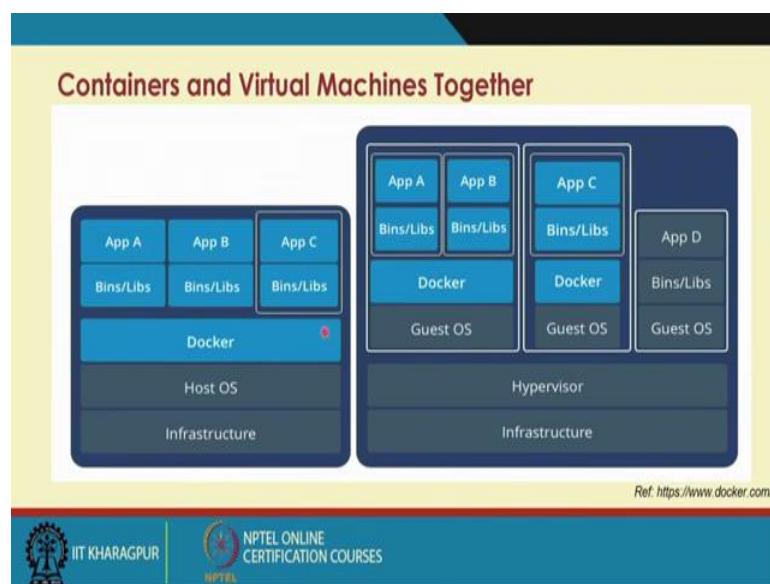
Whereas in this case slight modification can be done; allows us only to save the difference between container A and container A dash and that can be done at that container level.

(Refer Slide Time: 26:58)



So, overall working as we have the Docker, as we have seen this is the Docker engine is there. So, form the source repository built over this Docker engine; it is a container, a particular container; class container A, we push it to this container image registry; which can be searched by the users or the consumer. This container finally, is basically this Docker is a some sort of a client server mode operation. So, it search and then pull this things and run on this particular instantiation of the different container classes.

(Refer Slide Time: 27:42)



So, if we try to look at all together like how it can run on the cloud. So, we have infrastructure hypervisor over that guest OS, we can have a Docker services; in this case we have due to different container here only one and so and so forth. So, the cloud can host or can become a platform for running this container type of services. So, that is on some VMs; that is in a IaaS type of cloud, we can run this sort of services.

(Refer Slide Time: 28:25)

### Why is Docker needed for applications?

- Application level virtualization.
- A single host can run several spatial applications for utilization of resources.
- Build once, deploy anywhere, run anywhere.
- Better collaboration while development of applications.



Ref: <https://www.docker.com/>

 IIT KHARAGPUR |  NPTEL ONLINE CERTIFICATION COURSES



So, there are needs are enormous or needs, requirement, applicability is enormous like application level virtualization is possible. A single host can run several special application for utilization of resources; as we have seen when we discussed about special cloud there are several special applications and now we can have a single host which can run several special applications. Build once, deploy anywhere, run anywhere, type of things; philosophy. So, once I develop and build it and then I deploy anywhere, run anywhere. Better collaboration between developmental of applications, so I can have better collaborative applications into place.

(Refer Slide Time: 29:19)

## Terminology - Image

- Persisted snapshot that can be run
  - *images*: List all local images
  - *run*: Create a container from an image and execute a command in it
  - *tag*: Tag an image
  - *pull*: Download image from repository
  - *rmi*: Delete a local image
    - This will also remove intermediate images if no longer used

Ref: <https://www.docker.com/>



IIT KHARAGPUR



NPTEL ONLINE

CERTIFICATION COURSES

There are few terminologies some of them already you have seen. So, like images list of all local images, run, create a container from an image and execute a command in it. Tag an image, pull; download an image from a repository, rmi, delete a local image. So, this is also remove intermediate images if no longer is used so that resources can be released.

(Refer Slide Time: 29:48)

## Terminology - Container

- Runnable instance of an image
  - *ps*: List all running containers
  - *ps -a*: List all containers (incl. stopped)
  - *top*: Display processes of a container
  - *start*: Start a stopped container
  - *stop*: Stop a running container
  - *pause*: Pause all processes within a container
  - *rm*: Delete a container
  - *commit*: Create an image from a container

Ref: <https://www.docker.com/>



IIT KHARAGPUR



NPTEL ONLINE

CERTIFICATION COURSES

There are some terminology which are more associated with Docker container like ps; list all running containers, ps-a; list all containers including the stopped one, top; display process in the container, start, stop, pause, rm; delete a container, commit; create an

image on the container. What you may notice that many of them things are some sort of Linux commands; already we are used to it.

So, in this case is also for container we can use those commands those who are interested you can basically hook into Docker dot com and see that how the coding can be done and how a container can be built using a Docker engine. So, this is freely downloadable and you can work and you can run that things on your desktop or server or even on your android devices, make a particular application and run on different devices.

(Refer Slide Time: 30:59)

**Dockerfile**

- Create images automatically using a build script: «Dockerfile»
- Can be versioned in a version control system like Git or SVN, along with all dependencies
- Docker Hub can automatically build images based on dockerfiles on Github

Ref: <https://www.docker.com/>

 IIT KHARAGPUR |  NPTEL ONLINE CERTIFICATION COURSES

So, there are some few more things; one is the Docker file; create an image automatically using a build script called Docker file can be versioned in a version control system like Git or SVN, along with other dependencies. Docker hub can automatically build images based on Docker files on Github, so these things are possible.

(Refer Slide Time: 31:23)

## Docker Hub

- Public repository of Docker images
  - <https://hub.docker.com/>
- Automated: Has been automatically built from Dockerfile
  - Source for build is available on GitHub

Ref: <https://www.docker.com/>

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

And the Docker hub; where the public repository of Docker images are there; like hub dot Docker dot com, there is a good resourceful area for this Docker images. And automated has been automatically built on Docker file. Source for build is available on the Github.

(Refer Slide Time: 31:44)

## Docker – Usage

- Docker is the world's leading software container platform.
- Developers use Docker to eliminate "works on my machine" problems when collaborating on code with co-workers.
- Operators use Docker to run and manage apps side-by-side in isolated containers to get better compute density.
- Enterprises use Docker to build agile software delivery pipelines to ship new features faster, more securely and with confidence for both Linux, Windows Server, and Linux-on-mainframe apps.

Ref: <https://www.docker.com/>

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, finally if we look at that different view point or different uses of this Docker is a worldwide claim to be leading software container platform being used at various levels, in different environment and became very popular. So, developers use Docker to

eliminate work on my machine problem. So, that is dependency on the things when collaboration on code with coworkers are much needed.

Operators use Docker to run and manage apps side-by-side in isolated container to get better compute density; that is the view point or the usage of the operators. Enterprises use Docker to build agile software delivery pipelines to ship new features faster, more securely with confidence for both Linux, Windows Servers, Linux-on-mainframe apps and so and so forth.

So, there are different applicability of the Docker and this service becoming pretty popular and those who are interested in this particular Docker technology, you may go through that Docker dot com and there are several other open sources; where you can see that how this compilation can be done and can have your own Docker images and run out on different platform and see the things. So, as far as the cloud computing paradigm is concerned; it acts as a platform and has seen application portability across cloud, making cloud more useful, more acceptable at different environment.

Thank you.

**Cloud Computing**  
**Prof. Soumya Kanti Ghosh**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**

**Lecture – 37**

**Green Cloud**

Hello, we will continue our discussion on Cloud Computing, different aspects of cloud computing. So, what we are discussing during this phase of this lecture is more looking at the different other aspects of cloud computing which make this computing more efficient, what are the different supportive technologies around it and how the overall cloud computing may be made more practical.

So, one of the aspects of this cloud is its energy consumption. We have seen in case of when we discussed about resource management that resource is in the cloud though theoretically that is infinite volume of resources, but the resources to be managed appropriately. So, that the service can be provided to the maximize maximum or to maximize its benefit right. So, it is a type of situation it what we looked at that where it is a win-win situation for both consumer and the provider and how this resource can be managed. One of the aspect of a resource management what we have discussed is to reduce the energy consumption right. What we see in that this overall major the dark side so called dark side of this cloud computing is huge volume of energy consume or huge amount of energy consumption.

If you have experience in any type of going through or if you have red any type of data centers or if you go through this internet different resources you will find that typically if you look at a data center the cost of the computing equipments or computing infrastructure is somewhere match with the other environmental infrastructure like the space environment ac power and so and so forth. So, it is something like x is spent on these is more or less x is also spent on that side right. Rather in some cases is more costly because of the cost of the particular space and power and type of things.

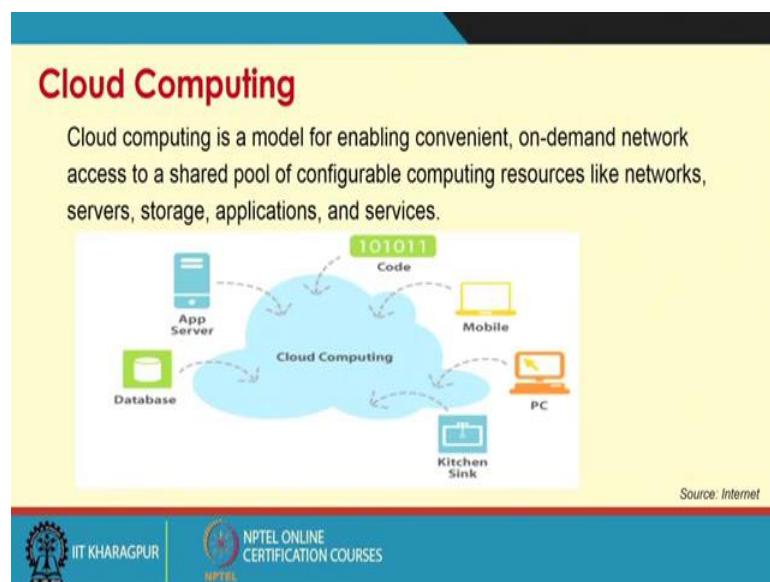
Secondly the amount of energy spent in maintaining this type of infrastructure is enormous right and what is been seen that unless this a overall consumption things consumption pattern can be reduced or the volume of consumption can be reduced in

some way or other it may not be feasible to scale up after extent right after a particular limit. So, though someone can have infinite or someone can have money to spend, but there may not be the supply is not available to the things.

So, in other and another aspect what we see that it is becoming sometimes it is a more hazardous or more negative effect on our environments specially the carbon footprint more you consume more energy being produced more may more is the carbon foot print. So, going to another unconventional energy sources may be other things etcetera. Nevertheless the overall these computing worlds need to look for some sort of a more quote unquote green computing. So, what we will discuss today briefly is that that different aspect of green cloud computing though it has a direct connection or direct relations with our resource management, but we will try to look at the different aspects of this green cloud computing and if we whenever we are setting up any infrastructure so this should keep we should keep in the mind not only the cost of the infrastructure, but also these consumption of energy how to reduce that consumption of energy and so and so forth type of things right.

So, that is what we discuss today is the green cloud.

(Refer Slide Time: 05:03)



If we quickly revisit our definition, so what it says that it is a cloud computing is a model for enabling convenient on demand network access to a shared pool of configurable resources like networks servers storage applications and services like this is the first line.

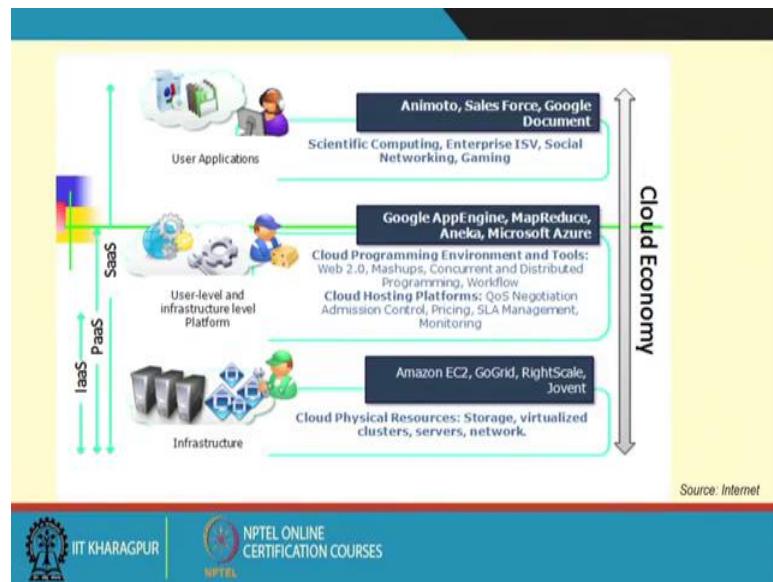
There are different characteristics like we have say that it can scale up or infinite scaling it has the ability of metered service broad network action and different characteristics are there omnipresent or ubiquitous access to these computing facilities and type of things.

Nevertheless these definitions what if we look at other in a little deeper way there is a lot of energy hungry resources are there right like if we may look at networks, servers storage and applications and services we run of the things they all take lot of energy if we can efficiently use those there may be there may be a chance that we can reduce the overall energy consumption otherwise the energy consumption may be considerably high. Like say if I say that if I have, if I my running VMs are say number of servers I am having 4 my number of running VMs are again 4.

So, every server can accommodate say running VM is 8 and every server can accommodate 4 VM. So, it may so happen that every server is distributed to VM per server it may look good that there is a load is distributed, but if you look at the energy point of view the energy consumption may be pretty high instead I could have packed them into two server of 4 4 each and so and so forth and other two server I can put on a off mode or a sleep mode or a passive mode and this could have saved energy in a bigger way.

So, these are the aspects which we need to try to look at and there is there is definitely a it is not like that that simple how we are trying to pose the problem how we are discussing, but there may be a lot of calculation of projects and type of things there are issue of SLAs, a key OS and type of things, but never the less taking all those in to consideration there is lot of opportunities may be there to go green right.

(Refer Slide Time: 07:36)



So, if you if we visit this means see the other slide. So, at the down we have a these infrastructure right where cloud physical resources storage virtualize clusters servers networks like Amazon EC2, go grid and different other solutions are there. At the middle we have a cloud programming environment or platform and cloud hosting platforms which uses this infrastructure and a there are Google app engine map reduce Microsoft Azure and Aneka type of things and at the top we have this SaaS or software as a service scientific computing etcetera.

Now these in turn every layer upper layer in turn uses the downwards layer and a more efficient is this computing at every more efficient is the implementation of each layer may help us in reducing the overall energy consumption right. So, it is not only that efficiency out here though we when we look about energy we mostly look at the IaaS type of things, but it is also that at a higher layers also contribute type some miss also contribute towards this proper energy management right. Like if I have an algorithm which takes unnecessary loops and takes more CPU time that may be energy inefficient then more efficient algorithm which is which where we can reduce the complexity and type of things right. So, it at much higher level I can do something which intern reduces my energy consumption or CPU usage time or network usage time and that intern reduces the overall energy consumption of the cloud infrastructure right.

(Refer Slide Time: 09:37)

## Green Cloud ?

- Green computing is the environmentally responsible and eco-friendly use of computers and their resources.
- In broader terms, it is also defined as the study of designing, manufacturing or engineering, using and disposing of computing devices in a way that reduces their environmental impact.
- Green Cloud computing is envisioned to achieve not only efficient processing and utilization of computing infrastructure, but also minimize energy consumption.

Source: Internet



So, looking at these if we try to look at a green cloud, so green computing is a environmentally responsible and eco-friendly user computers and their resources right. So, its first start is that it should be environmentally responsible means minimum carbon foot print and so on and eco-friendly use of computers in broader terms it is also defined as a study of designing manufacturing or engineering using and disposing of computer devices in such a way to reduces the environmental impact right. So, if we look at that how the designing manufacturing engineering all those things go on green cloud computing is envisioned as a to achieve not only efficient processing and utilizes now computing infrastructure, but also minimize the energy consumption right.

So, it is not only how not only it will be efficient in processing and computing and giving maintaining key OS and SLA, but also it gives minimize the energy consumption, so what when we look about talk about green cloud computing we look both side of the thing.

(Refer Slide Time: 10:50)

## Cloud Advantages

- **Reduce spending on technology infrastructure.** Maintain easy access to information with minimal upfront spending. Pay as you go based on demand.
- **Globalize your workforce on the cheap.** People worldwide can access the cloud, provided they have an Internet connection.
- **Streamline processes.** Get more work done in less time with less people.
- **Reduce capital costs.** There's no need to spend big money on hardware, software or licensing fees.
- **Improve accessibility.** You have access anytime, anywhere, making your life so much easier!
- **Minimize licensing new software.** Stretch and grow without the need to buy expensive software licenses or programs.
- **Improve flexibility.** You can change direction without serious financial issues at stake.

Source: Internet



IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Cloud advantages are well known if we again look quickly reduce spending on technology infrastructure. So, what we feel that that instead of infrastructure I shift to the cloud infrastructure globalized your work force in a cheap. So, that it is omnipresent stream line processes reduce capital cost improve accessibility minimize licensing of new software improve flexibility.

So, there is a host number of advantages are there which may interns try to reduce the computing at different in a different environment and in a and going like in terms of setting up your infrastructure at your local things and going to the things. So, in sense it may appear that we are in a sense we are reducing a energy consumption, but at the other end the consumption things increases. So, what we are trying to look at is not that that what we say about here we are more at look at the how to make those service provider ends how things can be made more efficient.

(Refer Slide Time: 12:08)

## Cloud – Challenge

- Gartner Report 2007: IT industry contributes 2% of world's total CO2 emissions
- U.S. EPA Report 2007: 1.5% of total U.S. power consumption used by data centers which has more than doubled since 2000 and costs \$4.5 billion

>> Need of Green Cloud Computing....

Source: Internet



IIT KHARAGPUR



NPTEL ONLINE  
CERTIFICATION COURSES

So, challenge specially in terms of energy consumption and carbon footprint like if you see the Gartner Report on 2007 it industry consume two percent of the world contributes to the 2 percent of the world's total carbon dioxide emission. So, it is a pretty high right it though its talks about it industry as a whole, but this major player or the cloud service provider are major contributor of the things right they are cooling they are overall energy consumption is pretty high.

Like big data centers installation people say that it is like a it is mini city like consuming power, so highly like. So, there is another report we say 1.5 percent of the us power consumption use by data center which has more than double since two thousand seven it is also around the around 2000 since 2000 and cause 4.5 billion dollar right it is also a report which is around 2007. So, in down the line 7 years it has more than doubled right.

So, as it is ever increasing phenomena that this shifting towards cloud or cloud industry is growing at a much higher pretty higher rate. So, this figure of energy consumption is likely to increase much more than getting reduced. So, it is more demand for cloud more energy consumption and so and so forth right. So, there is a need of green cloud computing definitely there is a need of green cloud computing.

(Refer Slide Time: 13:56)

## Importance of Energy

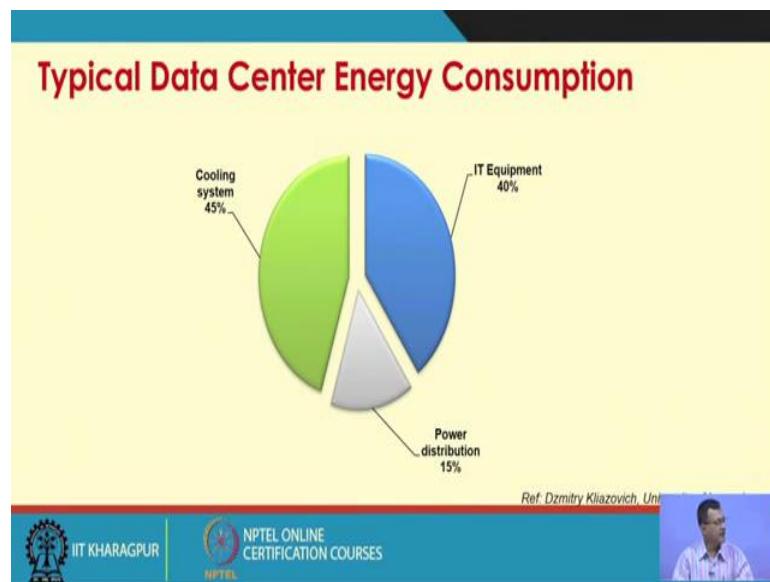
- Increased computing demand
  - Data centers are rapidly growing
  - Consume 10 to 100 times more energy per square foot than a typical office building
- Energy cost dynamics
  - Energy accounts for 10% of data center operational expenses (OPEX) and can rise to 50% in the next few years
  - Accompanying cooling system costs \$2-\$5 million per year

Ref. Dzmitry Kliazovich, University of Luxembourg

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

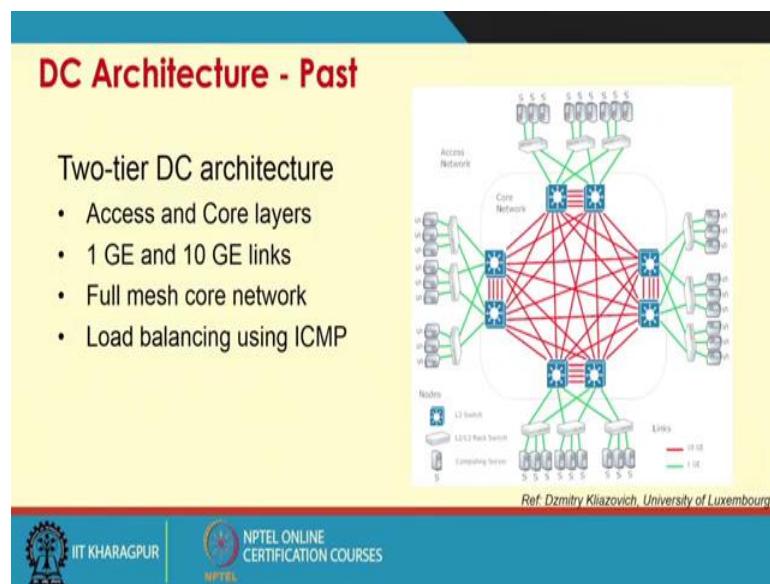
So, importance of energy increased computing demand data centers are rapidly growing consume 10 to 100 times more energy per square foot then typical of his building right this is a one rough cut thing energy is cost dynamics energy accounts for 10 percent of the data center operational cost right what other things might you say what we say OPEX and can rise to 50 percent of the next few years what they say that it is such a high rise going on it again one some report said that the cooling system cost around 2 to 5 million dollar per year this is also somewhere 2030 or so, that report which comes up.

(Refer Slide Time: 14:42)



So, again a rough cut if we see the energy consumption, so the equipment takes around 40 percent power distribution takes around 15 percent cooling system around 45 percent. So, if you say the if you look at the computing equipment is taking 40 percent over the total energy rest 60 percent is taking by this infrastructure is mesh thus end of a mesh cloud environment infrastructure itself right to set up this cloud that we need to set the environment that is sc power distribution appropriately power distribution etcetera that talking over the 60 percent of the energy the computing is around 40 percent. So, computing is still less.

(Refer Slide Time: 15:28)

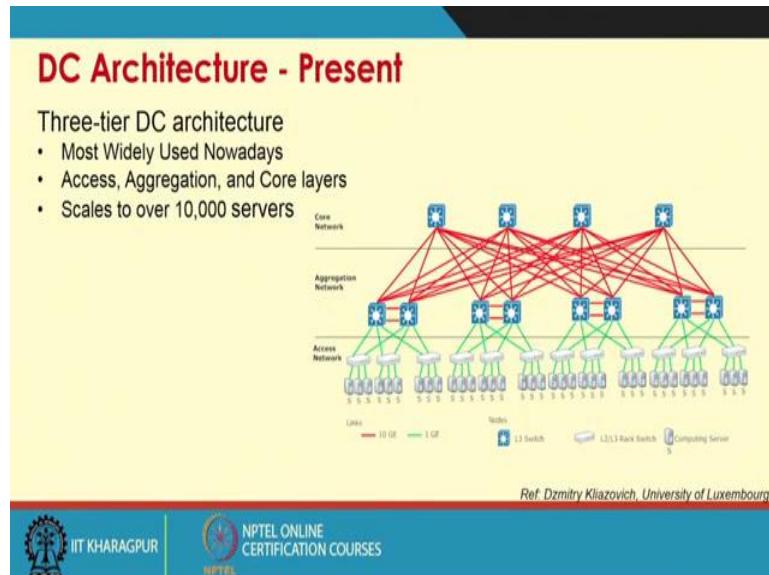


Now, if you look at the DC architecture. So, this has also evolving right it is also changing over time and more complex equipment are coming of course, more energy efficient equipment are also coming, so two-tier DC architecture access. So, initially it was access and cold layers. So, you have the centralized coal layer which is a highly mesh they are interconnected with a, vary concentrated mesh network and there are access layers from here the things are access these are the different access layers right this is the if you look at the DC architecture type of things computing architecture.

So, full mesh core network load balancing using ICMP protocol right, so this is typically there, so there are different type of switches then they are three switches they are two and they are three rack switches and computing servers. So, computing servers at the end mile or at the age and this layer three switches which routes traffic are at the core of the

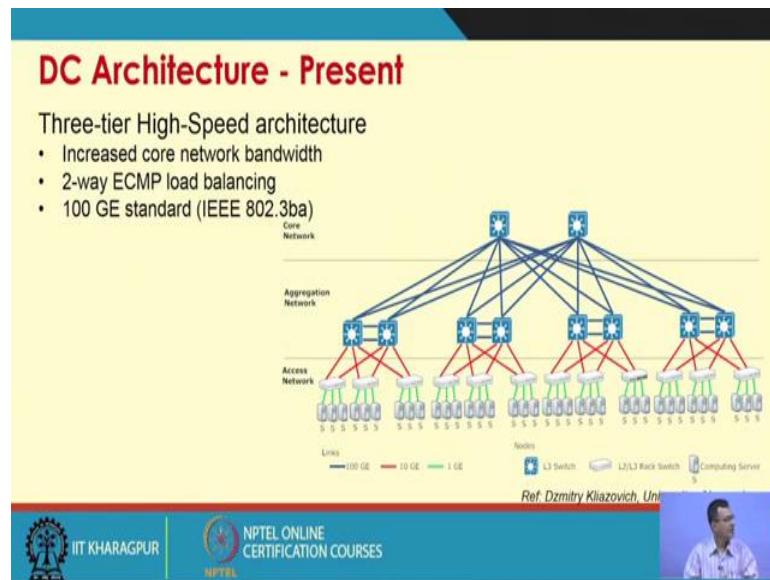
network. So, this is the typical architecture which were typically there in means old or past DCs though there are still some of things are still existing.

(Refer Slide Time: 16:57)



So, rather like all other large network this DC architecture present which see as a three tier DC architecture that is most widely nowadays access aggregation and core layers. So, there is a access layer, there is a aggregation layer. So, it is not only these layer three switches at the core, but also there are three switches at the aggregation layers and then we have that then this access layer. So, it is a hierarchical structure made to proper management and appropriate distribution which minimizes and of or better load balancing and type of things and so and so forth right. So, this is more for, it is more it is amicable for two scale up its scales up much better than our previous things like scales over ten thousand servers and so and so forth for a particular disease. So, it is this sort of architecture are there.

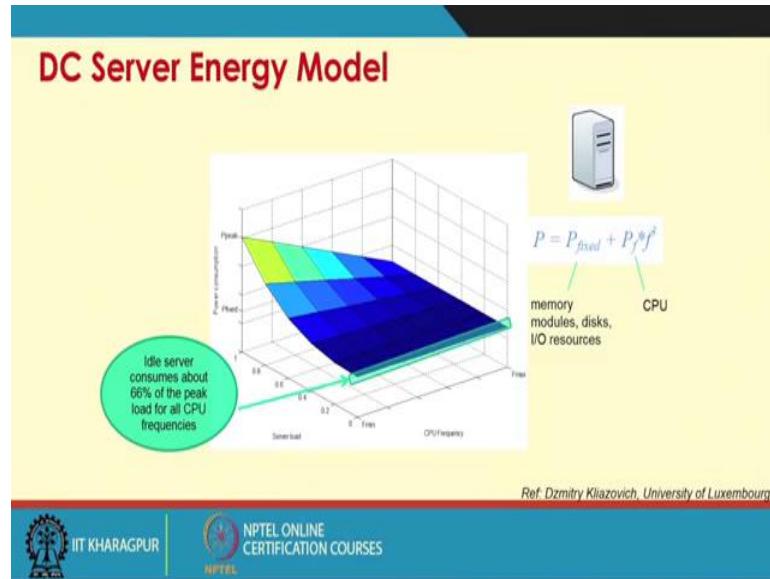
(Refer Slide Time: 18:03)



So, along with that it has high speed architecture three tier high speed architecture in increased core network bandwidth, 2-way ECMP load balancing. So, 100 giga bit ethernet standard connection over the thing. So, it is a much higher thing. So, what have thing previous thing was a standard connectivity with 1 to 10 giga bit connectivity whereas, here we have a jump of 100 giga bit connectivity. So, these links are very high speed links and then we have much lower links at the down the line right. So, it is better aggregations better scaling up and better management of that whole infrastructure is there. So, it is more better responsive type of architecture.

So, this present days architecture whenever we deploy all those things these are again we need to look at the energy consumption of the things. It is not only facilitates better computing or better accessibility of the computing things, but also it consume at times much more energy right. So, we need to have a tradeoff that whether there is a requirement of sub such things or that facilitating or providing services and increasing energy performance versus energy trade off should be three though we do not comprise on; do not want to comprise on performance, but making it efficient or energy efficient is one another goal.

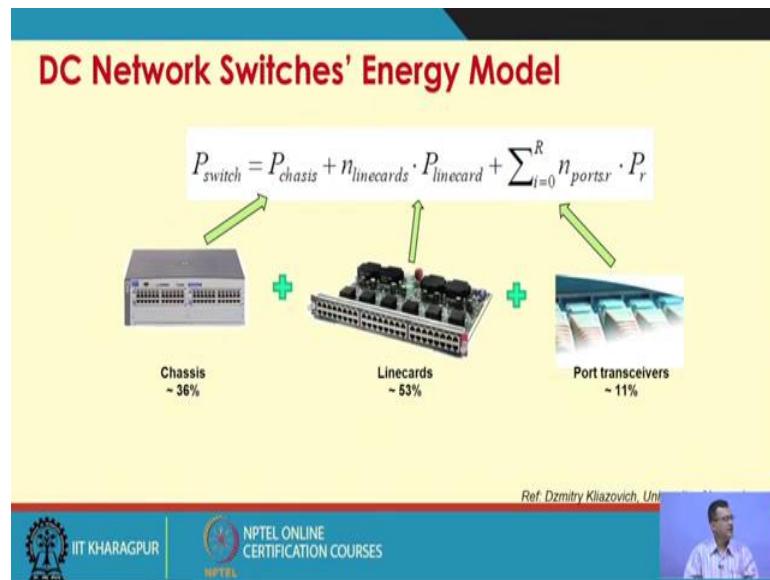
(Refer Slide Time: 19:55)



So, this is a typical energy model of DC server energy model what you see that even in the idle server consumes around 66 percent of the pig load of the CPU frequency. Even in this idle state it takes a considerable amount of energy so that it is always running energy. So, ideally if a particular service provider has no loads something it has to maintain the things. It is like if we consider that is in a if you in a particular shops say ice cream parlor or so, even there is no customer or not that particular customer it has to maintain as a particular level of cooling and type of things right. So, it is a some sort of energy level is there.

There are typical there are different model of energy models. In this case there is we have a fixed power model where the memory modules disk etcetera another CPU based on that frequency n number of fix CPU up etcetera we can have another other modules.

(Refer Slide Time: 21:10)



Similarly, if we look at the switches angry models so there are different category of things one is that what we see that the chassis, the chassis where if we have say a bunch of blade servers. So, it goes into a chassis right. So, typically say chassis contain the sixteen blades or of half height size so that chassis itself consume energy. So, if you look at that, the 36 around 36 percent of the a typical figure just to show you a rough cut figure that how things how the energy is important a chassis consume around 36 percent the line cards around 43 percent and these port trans receiver where the data being transmitted or transmission received is 11 percent right.

So, this model that p chassis plus number of line cards into p line cards and that number of aggregation of the number of ports and along with that the summation of things with that number of ports is that summation of the energy consumption will be the overall switching energy consumption. And that is also if we see that it can be considerable based on that if it is properly loaded and type of things like I can have traffic appropriately distributed I can have a better figure out here.

(Refer Slide Time: 22:46)

## Impact of Cloud DC on Environment

- Data centers are not only expensive to maintain, but also unfriendly to the environment.
- Carbon emission due to Data Centers worldwide is now more than both Argentina and the Netherlands emission.
- High energy costs and huge carbon footprints are incurred due to the massive amount of electricity needed to power and cool the numerous servers hosted in these data centers.



So, this all those things has a has a definite impact on the environment right or environment, this environment is not what we are talking about the cloud environment, but it is environment as a whole that overall environment like as we have talked about carbon footprint there may be effect of a heating there may be other different sort of a polluting effect right.

(Refer Slide Time: 23:11)

## Impact of Cloud DC on Environment

- Data centers are not only expensive to maintain, but also unfriendly to the environment.
- Carbon emission due to Data Centers worldwide is now more than both Argentina and the Netherlands emission.
- High energy costs and huge carbon footprints are incurred due to the massive amount of electricity needed to power and cool the numerous servers hosted in these data centers.

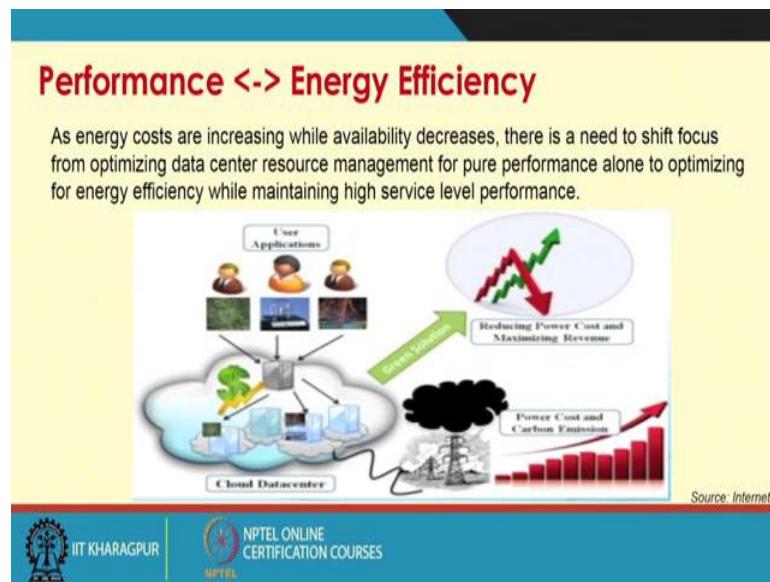
Source Internet



So, data centers are not only expensive to maintain, but also unfriendly to the environment it can be unfriendly carbon emission due to datacenter worldwide is now

more than more than both Argentina's and Netherland's emission. So, this is one figure it shows that the carbon emission is due to data centers worldwide is more than two counties overall emission high energy cost and huge carbon foot prints are incurred due to massive amount of electricity needed to power and cool numerous servers hosted in this data centers right. So, this is another major challenge of looking at it. So, it is a huge energy consumption.

(Refer Slide Time: 23:53)



So, you need to balance between these performance versus energy efficiency, as the energy cost increasing while availability decreases there is a need to see it focus from optimizing data centers resource management for pure performance along to optimizing energy efficiency while maintaining high service level performance. So, as the like when we talk about resource management we may be looking primarily on the performance alone right. So, you need to look at that performance vis-a-vis these energy management. So, how overall this can be achieved by performance versus energy efficiency can be achieved is there, is like reducing power cost and maximizing revenue may be the thing like power cost and carbon emissions are ever increasing. So, you need to look at in a more practical way.

So, if you look at in the modelling terms. So, in our this model this energy components would come into play right that how that need to be managed how need to be controlled should come into play.

(Refer Slide Time: 25:05)

## CSP Initiatives

- Cloud service providers need to adopt measures to ensure that their profit margin is not dramatically reduced due to high energy costs.
- Amazon.com's estimate the energy-related costs of its data centers amount to 42% of the total budget that include both direct power consumption and the cooling infrastructure amortized over a 15-year period.
- Google, Microsoft, and Yahoo are building large data centers in barren desert land surrounding the Columbia River, USA to exploit cheap hydroelectric power.

Source: Internet

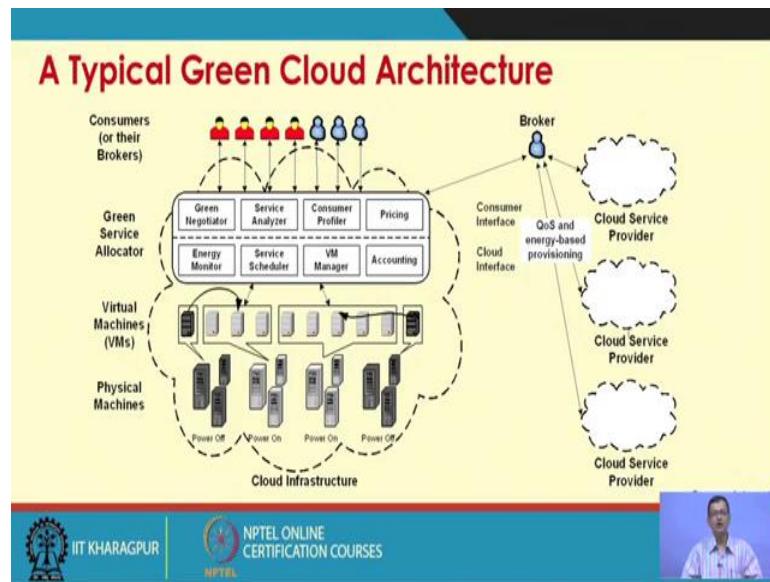


IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

There are several initiative needed from the cloud service providers end like cloud service provider need be adapt measures to ensure that their profit margin is not dramatically reduced due to high energy cost right. Amazons estimate one figure shows that the energy related cost to the data center amount to 42 percent of the total budget and include both direct power consumption and cooling infrastructure amortizes to over 15 year period.

Google, Microsoft, Yahoo are building large datacenter in barren desert land surrounded by Columbia river to exploit the cheap hydroelectric power etcetera. So, that is a tendency of make the datacenter more near to the power to the power generation unit. So, that your transmission of power transmission loss etcetera are reduced to a drastic element.

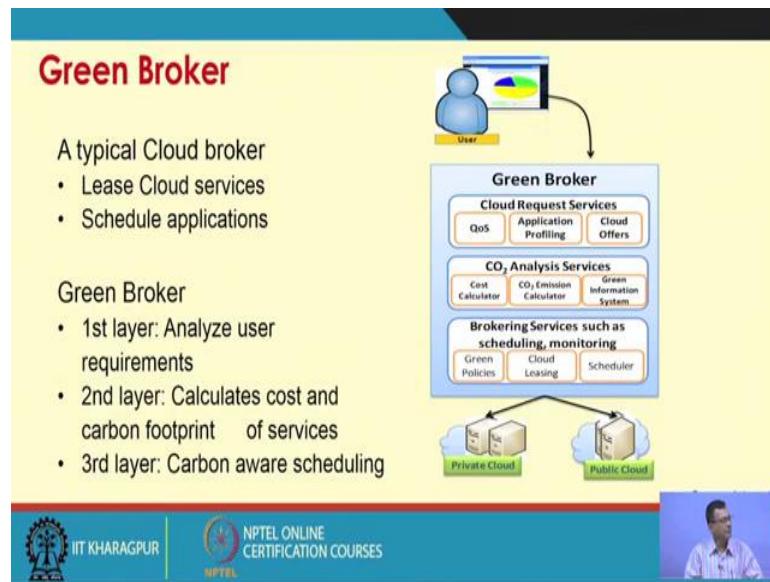
(Refer Slide Time: 26:04)



So, taking, so if this is typical green cloud architecture if we look at, so there are at the bottom end is physical machines there are several virtual machines and this green cloud allocator what we say now the green cloud brokering system, which brokers in favor of the consumers. So, this is a grid negotiator, service analyzer, consumer profiler, pricing, energy monitor, service scheduling, VM manager accounting this components where otherwise also they are in the architecture. What we see more is looking at that energy consumption energy consumption related parameter or matrix come into play right.

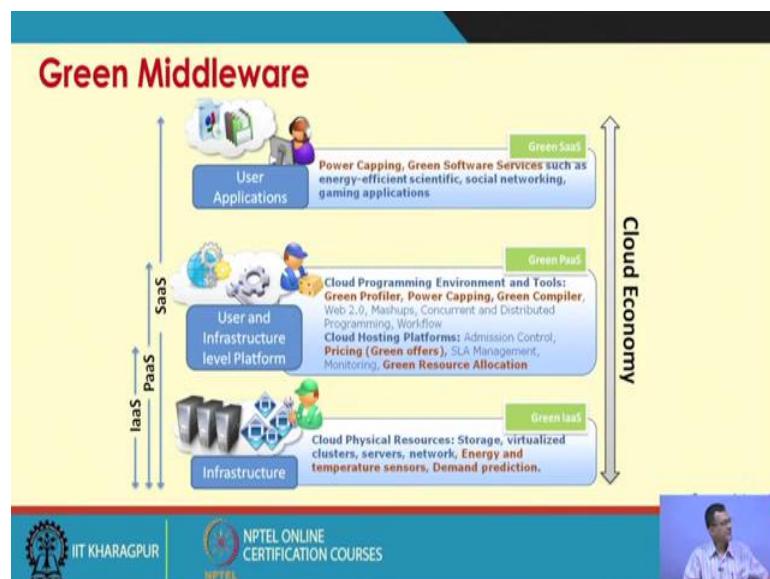
So, broker which look at QS and energy based provisioning of the different cloud service provider based on these they provision on these different cloud service provider. So, it is more energy aware provisioning of services.

(Refer Slide Time: 27:19)



So, since that green if we look at the green broker. So, it has, so it a typical ground broker lease cloud services schedule application that is the measure duty of the things. So, when you look at the green broker. So, analyzing user requirement calculate cost and carbon footprint of the services carbon aware scheduling. So, in now the scheduling is carbon aware scheduling right brokering services scheduling monitoring carbon dioxide analyzing services cloud request services. So, it does a carbon aware scheduling.

(Refer Slide Time: 27:54)



Similarly, we have a green middleware where what we say that green IaaS. So, what if you remember the initial figure, so we said storage virtualize service etcetera energy temperature sensor demand predictor is added. Here also we have at the PaaS label that green profiler power capping green compiler similarly green resource allocation system at the PaaS label, in the SaaS label power capping green software services and so and so forth. So, these are the different way we try to address this issues.

(Refer Slide Time: 28:36)

## Power Usage Effectiveness (PUE)

- \*  $PUE = \frac{\text{Overall Power}}{\text{Power Delivered}}$
- \*  $1 \leq PUE \leq \infty$
- \* “IT Load”
- \* IT Manager & Infrastructure Manager
- \* CUE
- \* Measurement, Modeling, Quantify
- \* Average PUE in US = 1.91

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

There is some effort of looking at the power usage effectiveness measure to find out that what is PUE and how to measure that how effectiveness is this power uses. So, there are different theories behind this. So, this there rough cart measures that how whether we can measure this power usage effectiveness of a typical infrastructure or a typical cloud service provider.

(Refer Slide Time: 29:05)

## Power Usage Effectiveness (PUE)

- \*  $PUE = \frac{\text{Overall Power}}{\text{Power Delivered}}$
- \*  $1 \leq PUE \leq \infty$
- \* “IT Load”
- \* IT Manager & Infrastructure Manager
- \* CUE
- \* Measurement, Modeling, Quantify
- \* Average PUE in US = 1.91



IIT KHARAGPUR



NPTEL ONLINE  
CERTIFICATION COURSES



So, to summarize clouds are essentially data centers hosting application services offered on subscription basis; however, they consume high energy to maintain their operations, so high operational cost plus additionally environmental impact which we try to ignore. So, that is one of the major aspects. Presented, so what we look at like look at that carbon aware green computing frame work to look at. So, there are several open issues lot of research to be carried out to maximize energy efficiency as a cloud centers, developing regions or to benefits the most benefits more that wares would be situation and so and so forth.

So, what we see overall that a overall this sort of computing aspects has a major concern not only from the service provider or consumer point of view it is a concerned worldwide from the environment point of view that the huge amount of energy being consumed which has a carbon footprint and a better and there is a need for better energy management to for so that this type, this sort of cloud computing environment in with lot of benefits are able to exploit it by the consumer. So, we need to head for some sort of a green cloud computing environment.

Thank you.

**Cloud Computing**  
**Prof. Soumya Kanti Ghosh**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**

**Lecture - 38**  
**Sensor Cloud Computing**

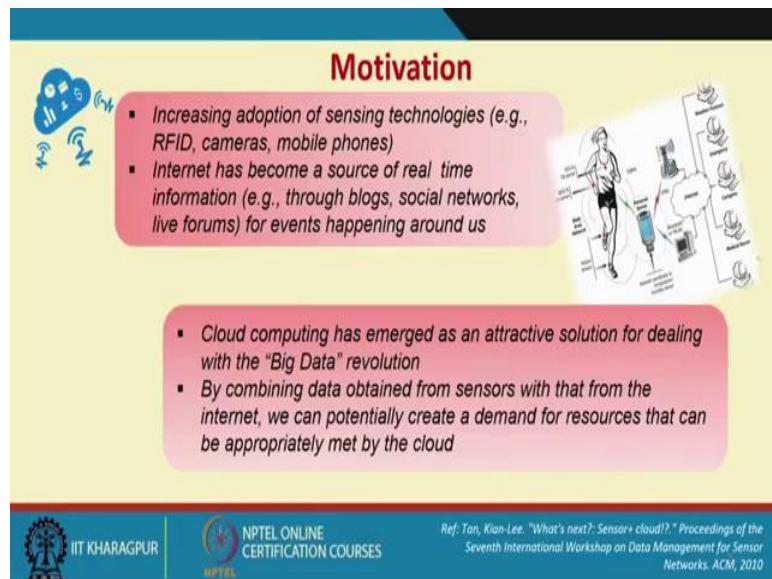
Hello. So, we will continue our discussion on cloud computing. Today, we will discuss on a topic which is very much relevant in today's context that is sensor cloud, right. So, what we have seen or what we are seeing or having these days, it is a world of sensors, right. It is you know enormous number enormous variety of sensors around us, right, it may start from temperature sensor or sensing pressure sensor to atmospheric different atmospheric parameters to if you can look at that camera as a sensors there are other different biological sensors which senses different aspect of our human physiology and other things, right. So, there are different aspects of sensors and it is a ever increasing phenomena, right.

So, what we have? We have a set of sensors which senses or accumulates information about environment about our health condition or about a particular device or so and so forth and based on the information. So, there is at the back end there is some intelligence or decision making systems which takes some call and activate or actuators are activated. So, we have a sensors and set of actuators in and in between, there should be a either physically or logically a decision support system or a intelligent systems which takes something, right, like if you look at our modern cars which are a bunch of sensors some somebody was saying other day that if the time is coming when where a particular car may have more sensors than mechanical parts right it may be a joke, but it says that the amount of proliferations of sensors for a monitoring management and activating different-different mechanisms and processes around us, right.

So, what we try to see that there is a amalgamation or there is a there is whether there is a need to integrate this sensors with cloud infrastructure right or having a cloud of sensors which has a property of emulating the properties of cloud, right, like what we have seen that the different properties of cloud which is making popular whether this type of ubiquitous sensors or a huge volume of huge number of sensors can emulate a sensor

cloud which is much more effective than individual sensors right or that we will try to look at.

(Refer Slide Time: 03:43)



The slide has a yellow header with the word "Motivation" in red. On the left, there are icons of a cloud with signal waves and a smartphone. Below the header are two pink callout boxes containing bulleted text. The top box discusses the increasing adoption of sensing technologies and the internet as a source of real-time information. The bottom box discusses how cloud computing has emerged as a solution for big data and how combining sensor data with internet data creates a demand for resources. To the right of the text boxes is a small illustration of a person running with various sensors attached to their body, connected to a network of devices like a laptop, a mobile phone, and a car.

**Motivation**

- Increasing adoption of sensing technologies (e.g., RFID, cameras, mobile phones)
- Internet has become a source of real time information (e.g., through blogs, social networks, live forums) for events happening around us

- Cloud computing has emerged as an attractive solution for dealing with the "Big Data" revolution
- By combining data obtained from sensors with that from the internet, we can potentially create a demand for resources that can be appropriately met by the cloud

Ref: Tan, Kian-Lee. "What's next? Sensor+cloud?." Proceedings of the Seventh International Workshop on Data Management for Sensor Networks. ACM, 2010

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Again as we see that it is more of an overview which we will open up the scope for you to read more somebody interested in research activity, etcetera to look into the things. So, what we see increasing adaptation of sensing technologies right that it is RFID camera mobile phones are nowadays sensors. So, locations as we have seen that different meteorological parameter are being sensed. So, internet has become a source of real time information on the hands internet has become a real time through blogs social network for events happening around us, right.

So, 2 technologies which are extremely pervasive and it is has become both of them eventually become a part of our day to day life, right, on the other hand, cloud computing has emerged an attractive solution for dealing with big data revolution which can able to manage and able to take a call on this type of huge or weak volume of data; so, by combining data obtained from sensors; we that from the internet, we can potentially create a demand for sources that can be appropriately made by cloud, right. So, what we are trying what we are trying to look at that whether we can have this amalgamation of this the technologies or the or the processes which are collecting data, there is a way of dispersing or dissipating the data that is a accessing the data. So, both sensors internet cloud coming together to make our need or to full fill the different aspects of our either

day to day or commercial or industrial requirements of the thing, right. So, so rather we have referred a one of the pioneer paper that what next it is a whether sensor plus cloud is the things which are coming in this particular decade.

(Refer Slide Time: 05:49)



So, already we know about wireless sensor network, I believe that you have more or less it is known just to put some points. So, it is a seamlessly couples physical environment with the digital world. So, it is what it is there; there is a wireless sensor networks which has different sensor nodes or modes sometimes we say which collects information and those are digitized and pushed through the scene to the central infrastructure; so, what we do that whatever the physical environment is there is being captured and coupled with the; our digital world. So, sensor nodes are small low power low cost and provide multiple functionalities, right.

So, typically there are small they are power hungry cost wise also not. So, costly and as different type of functionalities like sensing capability sometimes it has a processing capability memory communication bandwidth battery memory power and these are the different type of components of the things. So, sensing and processing at times or transmitting or store and hold like memory and then push it. So, communication bandwidth and it requires a management of the battery power in aggregate sensor nodes are substantially data acquisition processing capability usually when it use in a

aggregation mode, there are primarily using acquisition of the data and processing some sort of a processing and push it to the next node or to other node.

So, useful in many applications and we have we know already that environmental health care education defense manufacturing smart phones and anything where you require some sort of a sensing and monitoring of the things. So, what we see here; there is a battery of sensor node; these are been organized in a particular fashion, but that can be ubiquitously thrown and can reorganize among themselves there are some gateway with which it connects to the rest of the world right and this can be different type of things like it can be management of forest planes city transport and so on and so forth.

So, it finds application in different aspects like. So, it is either a the equipment itself as a having a sensor thing like if I say a car or a airplane. So, those are different sensing units doing it or they are sensing and sending data like a temperature center, moisture sensor or; hey sorry, humidity sensor, say velocity sensor they basically collect information send it out and with all different sensors collected information a call is taken by the management or the management or the control unit, right.

(Refer Slide Time: 08:51)

## Limitations of Sensor Networks

- Very challenging to scale sensor networks to large sizes
- Proprietary vendor-specific designs. Difficult for different sensor networks to be interconnected
- Sensor data cannot be easily shared by different groups of users.
- Insufficient computational and storage resources to handle large-scale applications.
- Used for fixed and specific applications that cannot be easily changed once deployed.
- Slow adoption of large-scale sensor network applications.

 IIT KHARAGPUR |  NPTEL ONLINE CERTIFICATION COURSES

So, there are typical limitation in sensor networks that we already we understand that is a very challenging to scale sensory network to large sizes suddenly scaling up or scaling down this sensor networks is a major challenge, right. So, you whatever you have deployed in the physical sensor is there suddenly instead of collecting a say I want to

increase I am collecting temperature or acquiring temperature and suddenly increase it to a all larger scale and to on to collect a large thing it is scaling up is different even changing the granularity is different, right.

So, proprietary vendor specific design in most of the cases difficult for different sensor networks to be interconnected. So, there is another problem interoperability sensor, they not data cannot be easily shared between different groups of users because the everything if this isolated, then it is extremely difficult to think share first of all there is a problem of the proprietary formats. Secondly, there is no there may not be there may not be mechanisms to how to share data. So, I have a bunch of sensors collecting and putting somewhere and another bunch of sensors and putting somewhere and those need to be talking to each other; right. So, that needs to be enforced, right.

In sufficient computational storage resources; in sensors to handle large scale applications, right, if I have a more complex application, it is difficult to do on this use for fixed and specific application that cannot be easily changed once deployed. So, usually the sensor requirement are for fixed and specific applications that cannot be easily changed or deployed slow adaptation of large scale sensor network applications. So, if we have large applications. So, the immediate deployment and adaptation becomes a challenge.

So, these are some of the things which are we see; so, to say with this type of sensor network including wireless sensors networks, right.

(Refer Slide Time: 10:47)

## Limitations of Cloud Computing!

- The immense power of the Cloud can only be fully exploited if it is seamlessly integrated into our physical lives.
- That means – providing the *real world's* information to the Cloud in *real time* and getting the Cloud to *act and serve us instantly*.
- That is – adding the sensing capability to the Cloud

5



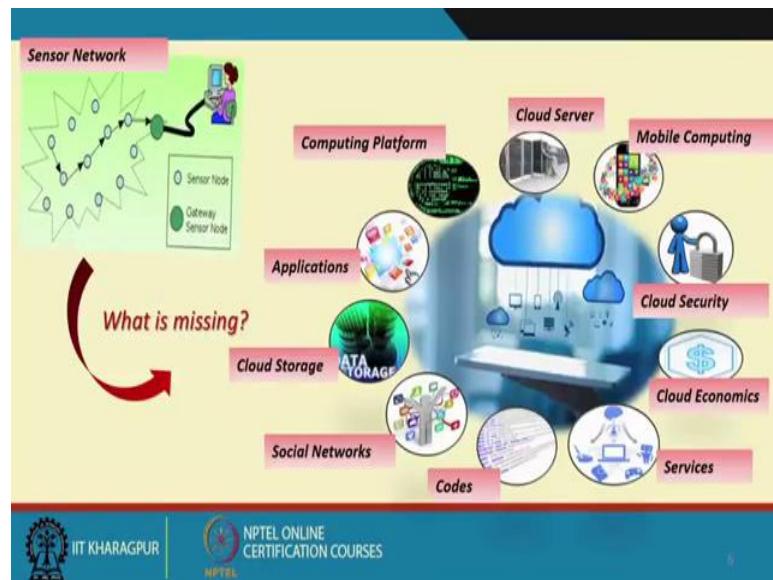
IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

And similarly, if we look at the there are in terms of the cloud other than all the limitations we have looked into there are other typical limitation like immense power of cloud can only be fully exploited or fully appreciated; so, to say if it is seamlessly integrated with our physical lines, right. So, it is it is not that when I demand I take things that it takes a some sort of a real world situation and process it on real world things, then only that actual power of the cloud is manifested or appreciated by the community at large.

So, what it means that providing a real world's information to the cloud in real time and getting cloud to act and serve us instantly right that is one thing or another thing adding the sensing capability to the cloud. So, cloud as such as the infrastructure wise just waiting and waiting for something to use it rather than it has a it has coupled with this sensing capability which sense do some processing and react to the things, right. So, having this sort of sensing capability will give a definitely extra age to these sort of a cloud computing.

So, to say it is if not; if we do not want to tell you the limitation of cloud, but to make a cloud really useful and effective we need to do this, right.

(Refer Slide Time: 12:14).



So, one side we have sensor networks typically we know that there is a bunch of sensors there is a way there is a sink node or gateway through which it is connected to the rest of the world and this there are different way of looking at the this sensing things like they can follow a particular path to the; say to the sink or they can form different clusters make the cluster hate and so on and so forth, there are technologies what you can see in the sensors network or on the other side if you look at the cloud infrastructure we have different sort of things like there are cloud server cloud computing platform several cloud level applications mobile computing cloud security aspects economics and so on and so forth, right and different models and different level of manifestation are there.

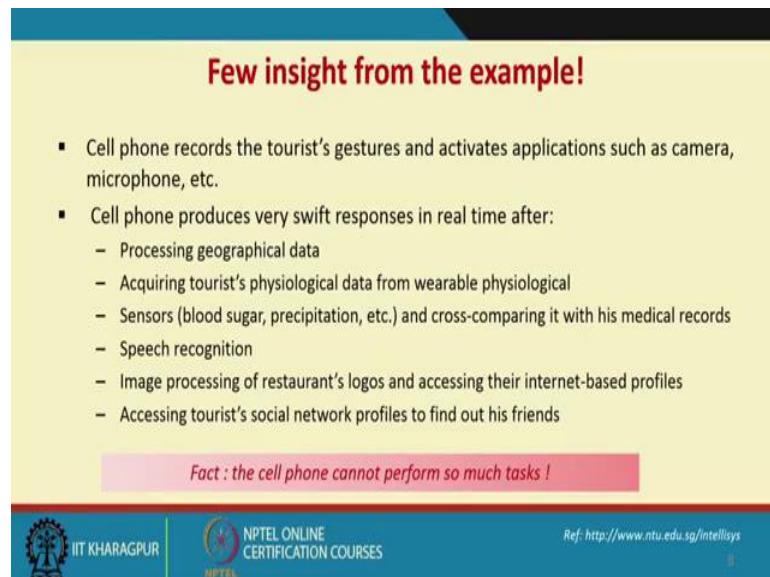
So, there is there is a missing link which we want to connect like making these sensor overall as a sensor cloud or connecting with the rest of the things which the cloud support which exactly the sensor cloud infrastructure or sensor cloud network try to address like.

(Refer Slide Time: 13:34)



So, like; if we look at a say a typical scenario like this sort of things where there are multiple parties multiple information are being exchanged these different type of resources are being different type of equipments specially mostly most cases our mobile phones are used to capture this informations, right.

(Refer Slide Time: 14:01)



And if we look at the cell phone is somewhat omnipresence and records the tourists gesture activities applications such as in camera microphone etcetera which could have been captured cell phone produces very swift responses to real time after the processing

geographical data to show that if it is interfaced with the geo cloud type of things acquiring tourist physiological data and wearable physiological sensors. So, physiological sensors like blood sugar etcetera and cross comparing with the medical records like. So, if somebody is need some medical help this automatically instantiated or before the some something serious happens there are speech recognition system to find out things image processing of the some restaurant logo or accessing internet based things accessing tourist social network profiles to find out his friends, etcetera.

So, what we see that even though is a simple scenario of visiting a particular place, but there are several type of sensors at different levels which are there some are the physically doing that some are may be doing some derived informations, right. So, it is a its according like whatever been posted in the social network finding the friend etcetera it is also trying to do some data information and try to couple with this physical information its location if at all; we are trying to do monitor its a other type of infrastructure those type of things are being used to that type of things.

So, the fact the cell phone cannot perform all the task which is needed right that we see that the cell phone cannot perform. So, we need require some of the something which is which is required for; which is something more, then this or coupling sensors with cloud philosophy.

(Refer Slide Time: 15:56)

### Need to integrate Sensors with Cloud!

- Acquisition of data feeds from numerous body area (blood sugar, heat, perspiration, etc) and wide area (water quality, weather monitoring, etc.) sensor networks in real time.
- Real-time processing of heterogeneous data sources in order to make critical decisions.
- Automatic formation of workflows and invocation of services on the cloud one after another to carry out complex tasks.
- Highly swift data processing using the immense processing power of the cloud to provide quick response to the user.

 IIT KHARAGPUR |  NPTEL ONLINE CERTIFICATION COURSES

So, there is a need to integrate sensors with cloud right the acquisition of data feeds from numerous body area networks right like a different type of body parameters and wide area like water quality monitoring that is a small or large scale things what we need to do that sensor networks in a real time type of things.

So, what we not need to do that it is not only real time informations, but also real time processing of the informations is there is a requirement which is coming up where sensors integrating as sensors with cloud may give some results; real time processing of heterogeneous data sources in order to make critical decision, right. So, we need to do some real time processing of heterogeneous data sources coming from different type of sensors like something coming from the body area network or sensors which are taking a more health related data accompanied by the other type of things like may be the overall environmental things along with some of the sensors which gives that like related to the heights and. So, and. So, forth and try to take a call integrating those reason.

These are the different heterogeneous sensor informations which need to integrate. So, automatic formation of work flows and invocation of services in the cloud one after another to carry out complex task tasks, highly swift data processing using immense processing power of the cloud to provide quick response. So, one side we have a capability of wide variety of sensors collecting information another side; what we have that computing power and the ability to interoperate between different type of heterogeneous data from the cloud we want to integrate together.

(Refer Slide Time: 17:54)

## What is Sensor Cloud Computing?

An infrastructure that allows truly pervasive computation using sensors as interface between physical and cyber worlds, the data-compute clusters as the cyber backbone and the internet as the communication medium

- It integrates large-scale sensor networks with sensing applications and cloud computing infrastructures.
- It collects and processes data from various sensor networks.
- Enables large-scale data sharing and collaborations among users and applications on the cloud.
- Delivers cloud services via sensor-rich devices.
- Allows cross-disciplinary applications that span organizational boundaries.

 IIT KHARAGPUR |  NPTEL ONLINE CERTIFICATION COURSES

10

So, if it comes to a scenario of what we say sensor cloud. So, what if you try to look at a the sensor cloud does an infrastructure that allows truly pervasive computing using sensors as interface between physical and cyber worlds the data computing clusters as a cyber backbone and internet are the communication media, right. So, what; let me repeat its a infrastructure that allowed pervasive computing using sensors as an interface to the physical and cyber world right and data computing clusters at the cyber backbone at the internet as the communication media. So, it is a amalgamation of this different technologies to realize a or to basically support a real time application or real time processing of informations, right.

So, it integrates large scale sensor network with sensing applications and cloud computing infrastructure it collects and process data from various sensor networks enables large scale data sharing and collaboration among users and applications of cloud delivers cloud services by sensor rich devices allows cross disciplinary application that is span organizational boundaries like it says lot of things, right. So, it says that what we are discussing that to make it the whole thing ubiquitous and omnipresent and to serve a variety of applications.

(Refer Slide Time: 19:28)

## Sensor Cloud?

- Enables users to easily collect, access, process, visualize, archive, share and search large amounts of sensor data from different applications.
- Supports complete sensor data life cycle from data collection to the backend decision support system.
- Vast amount of sensor data can be processed, analyzed, and stored using computational and storage resources of the cloud.
- Allows sharing of sensor resources by different users and applications under flexible usage scenarios.
- Enables sensor devices to handle specialized processing tasks.

 IIT KHARAGPUR |  NPTEL ONLINE CERTIFICATION COURSES

So, just to reinstate that enables user to easily collect access process visualize archive share and search large amount of sensor data from applications supports complete sensor data lifecycle from data collection to the decision support system as we are discussing vast amount of sensor data can be processed analyzed and stored using computational and the storage resource of the cloud, right. So, it is as a cloud as by virtue of it that having a huge infinite storage theoretically I can store this huge volume of data for further processing. So, many sensors are sending it is really a data challenge or a big data problem allows sharing of resources by different users and applications under flexible usage scenario enable sensor devices to handle specialized processing task.

(Refer Slide Time: 20:27)

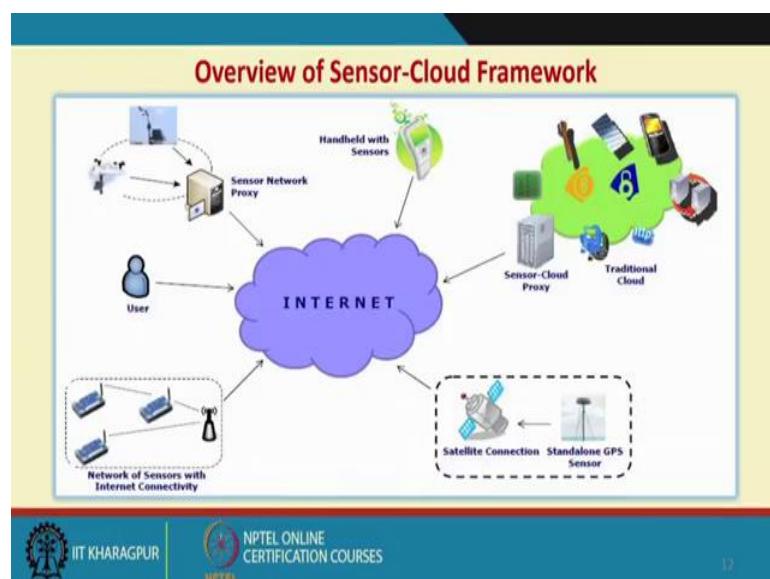
## Sensor Cloud?

- Enables users to easily collect, access, process, visualize, archive, share and search large amounts of sensor data from different applications.
- Supports complete sensor data life cycle from data collection to the back end. *Sensor cloud* enables different networks, spread in a huge geographical area, to connect together and be employed simultaneously by multiple users on demand.
- Varies huge geographical area, to connect together and be used simultaneously by multiple users on demand.
- Allows sharing of sensor resources by different users and applications under flexible usage scenarios.
- Enables sensor devices to handle specialized processing tasks.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

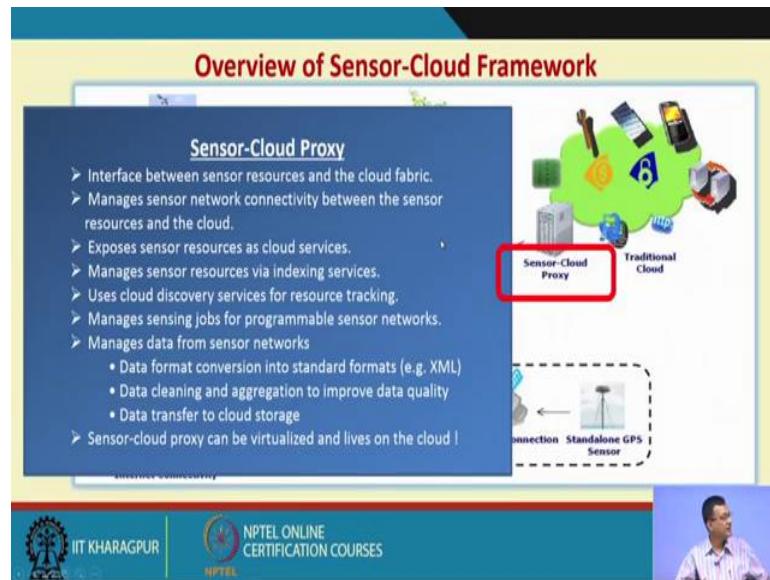
So, these are sensing processing actuating type of services. So, sensor cloud enables different network spread in huge geographical area to connect to together be employed simultaneously by multiple users on demand, right. So, if you can see that from the sensing technology we are in the; we are trying to incorporate or what we trying to do in the sensor cloud is putting that good properties or good use of this cloud in to this sensing overall sensor network infrastructure.

(Refer Slide Time: 20:58)



So, you can have a something overall overview of the things one side is sensor network where it; there is a sensor network a proxy which proxies it for the rest of the network. This handle devices or the handle sensors like mobile phones another type of things there is a traditional cloud we have a sensor cloud proxy which handles it and there are satellite connection standalone GPS sensors, there are network of sensors which capabilities different sort of things are being connected over the internetwork.

(Refer Slide Time: 21:33)



So, sensor cloud proxy it primarily interface between the sensor resources and the cloud fabric manages resource network connectivity between the sensor and resources which proxies on behalf of this; this say your sensors resources which can be heterogeneous which can be which homogenous type of things. So, it handles data format conversion like using technologies like XML interoperability data cleaning and aggregating data transfer to the cloud storage and so on and so forth.

(Refer Slide Time: 22:06)

### Overview of Sensor-Cloud Framework

The diagram illustrates the Sensor-Cloud Framework. At the top, there's a handheld device labeled "Handheld with Sensors". Below it, a central box is labeled "Sensor Network Proxy", which is highlighted with a red border. Arrows point from a "Sensor Network" icon to the proxy and from the proxy to a "Cloud" icon. A legend at the bottom left shows a blue square for "Network Internets" and a green circle for "Sensor Network".

**Sensor Network Proxy**

- For sensor resources that do not have direct connection to the cloud, this component provides the connection.
- Sensor network is still managed from the Sensor-Cloud Interface via Sensor Network Proxy.
- Proxy collects data from the sensor network continuously or as and when requested by the cloud services.
- Enhances the scalability of the Sensor Cloud.
- Provides various services for the underlying sensor resources, e.g. power management, security, availability, QoS.

**IIT KHARAGPUR** | **NPTEL ONLINE CERTIFICATION COURSES**



Sensor network proxy; for sensor resources that do not have different connection to the cloud this acts as a proxy. So, it is unlike, it is not; it cannot be expected that all sensor network will have that connection to the cloud and able to interface to the cloud. So, that is why we have a proxy where by which it is connected to the cloud, right, there are different other use cases like one is the traffic.

(Refer Slide Time: 22:26).

### Another Use case...

- Traffic flow sensors are widely deployed in large numbers in places/ cities.
- These sensors are mounted on traffic lights and provide real-time traffic flow data.
- Drivers can use this data to better plan their trips.
- In addition, if the traffic flow sensors are augmented with low-cost humidity and temperature sensors, they can provide a customized and local view of temperature and heat index data on demand.
- The national weather service, on the other hand, uses a single weather station to collect environmental data for a large area, which might not accurately represent an entire region.

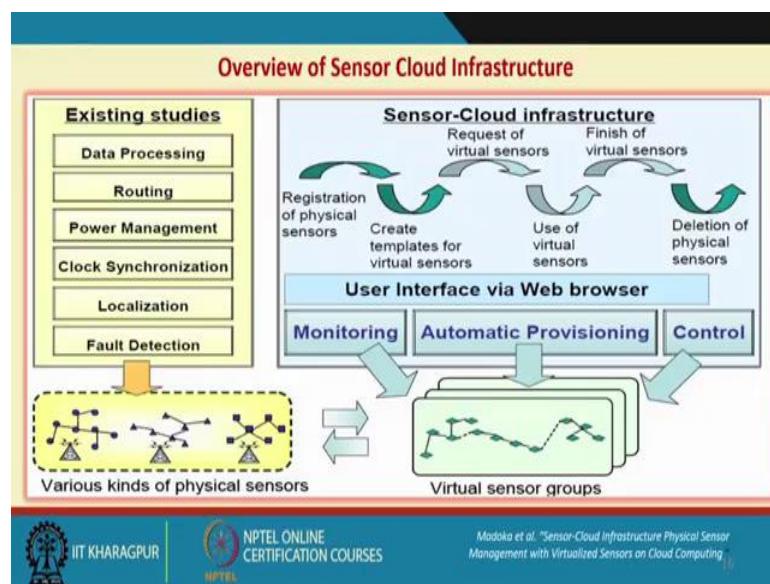
**IIT KHARAGPUR** | **NPTEL ONLINE CERTIFICATION COURSES**

We can say that traffic flow sensors that deploy large number in place city places like large cities right these sensors are mounted on traffic lights and provide real time traffic flow data, right.

So, these are on traffic lights systems on signaling system they can be mounted and real time things driver can use these data for his route planning right you see that there are this amalgamation of information and taking a call for somebody from the sensor things is there in addition if the traffic flow sensors are augmented with low cost humidity and temperature sensor that can also provide customized local view of the temperature and heat index on demand, right. So, it is also support the metrological department which other ways have fix number of sensors, right.

So, I can have a rough ideas, I have that precision sensor of the med department which keeps more precision and I have this local sensors which are on the lamp post and traffic signal and lights which keeps me a much more real time scenario of this temperature. So, the same type of infrastructure with little augmentation we can have different layer of informations or in other sense you can see that you have virtualized these things into different type of things, right, I have a one side that metrological departments looking for the med type of things traffic management department looking for that what is the overall traffic managing things I can have drivers on the road which uses for road planning and so on and so forth, right.

(Refer Slide Time: 24:08)



So, we can have one over view of the sensor cloud infrastructure like there are the standard way of looking at it data processing routing power management clock synchronization localized fault detection this is which are their basic physical sensors which has to be there and I basically what we try to do is a having a virtualized scenario or the physical layer over that as we have seen in the cloud a virtualized scenario which helps me in looking at different type of user related issues, right.

So, I have now a set virtual sensors or virtual sensor groups which can answer to the different category of applications like as we are discussing one may be med type of one may be traffic type of applications one may be action management type of things and something may be overall looking at that other phenomenon like how is that lightning or how is that overall environmental situations of that particular place.

So, we talk about virtual sensors what here is as we have said it is a emulation of the physical sensor that obtains its data from the underlying physical sensor as we have seen virtual machine, etcetera.

(Refer Slide Time: 25:23)

**Virtual Sensors?**

- A virtual sensor is an emulation of a physical sensor that obtains its data from underlying physical sensors.
- Virtual sensors provide a customized view to users using distribution and location transparency.
- In wireless sensors, the hardware is barely able to run multiple tasks at a time and difficult to run on multiple VMs, such as in traditional cloud computing.
- To overcome this problem, virtual sensors act as an image in the software of the corresponding physical sensors.
- The virtual sensors contain metadata about the physical sensors and the user currently holding that virtual sensor.

**IIT KHARAGPUR** | **NPTEL ONLINE CERTIFICATION COURSES** | **Modak et al. "Sensor-Cloud Infrastructure Physical Sensor Management with Virtualized Sensors on Cloud Computing"**

So, it is emulating the machines from the underlining the physical machines, right. So, virtual sensors provide customized view to the users using distribution and location transparencies right in a sensor wireless senor these hardware is barely able to run multiple task at time and difficult to multiple VMs and as in a traditional cloud. To overcome this problem virtual sensors acts as an image in the software of the

corresponding physical sensors right. So, that is as we have seen that in case of our cloudy cloud things also.

(Refer Slide Time: 26:17)

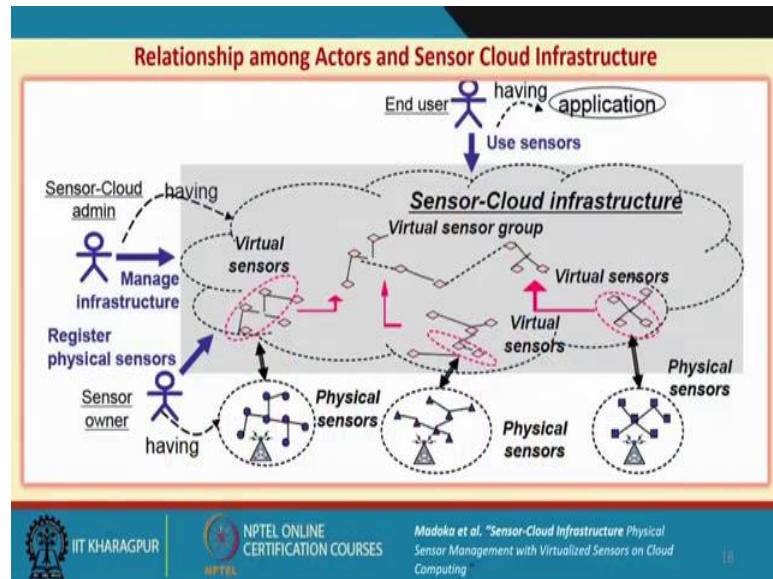
## Virtual Sensors?

- A virtual sensor is an emulation of a physical sensor that obtains its data from under
- Virtual sensor groups are used to group sensors based on their location to reduce the number of network requests required to access them
- In wireless sensor networks, virtual sensors are used to aggregate data over time and reduce the amount of data transmitted over the network. They are also used in distributed computing environments to manage resources and tasks
- To overcome the limitations of the current sensor technologies, virtual sensors are used to provide a more efficient and accurate way of monitoring and managing physical processes
- The virtual sensors contain metadata about the physical sensors and the users currently holding that virtual sensor.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | Madoka et al. "Sensor-Cloud Infrastructure Management with Virtualized Sensors on Cloud"

So, virtual sensors contain metadata about the physical sensors and users currently holding the physical sensors. So, we have a virtual sensors layer. So, there are physical sensors over that we emulate this virtual sensors and there are different virtual sensor group which uses that it may so happen one physical sensors may be contributing to the physical sensors. So, one to many-many to one many to many and this type of things scenario may come up.

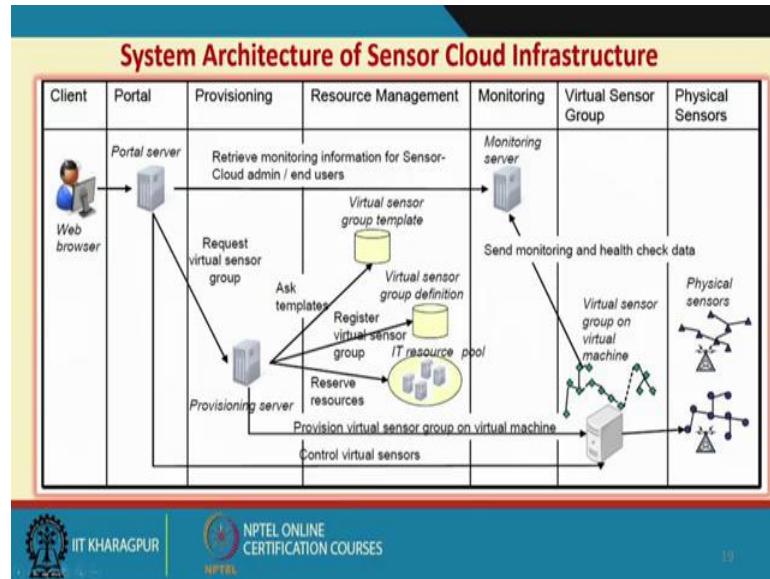
(Refer Slide Time: 26:28)



So, as we are discussing. So, there are sensor cloud admin. So, this is a sensor cloud infrastructure which has different physical sensors have virtual sensor groups which have underlining physical a virtual sensors which are coming from the physical sensors and then we have end users which are more-more looking for this more inclined to the applications or more bothered about the applications which talk to this uses this sensing network to look at it.

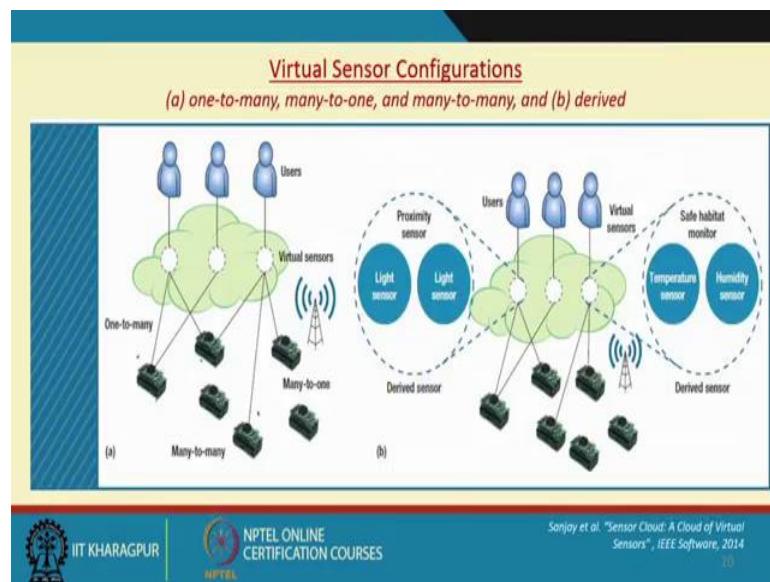
So, there is a management of this sensor cloud like registering physical sensors management mange infrastructure and that it may happen the sensors because this sensors are low powered manage by different type of in different situations. So, sensors may come up or can go down and type of things.

(Refer Slide Time: 27:38)



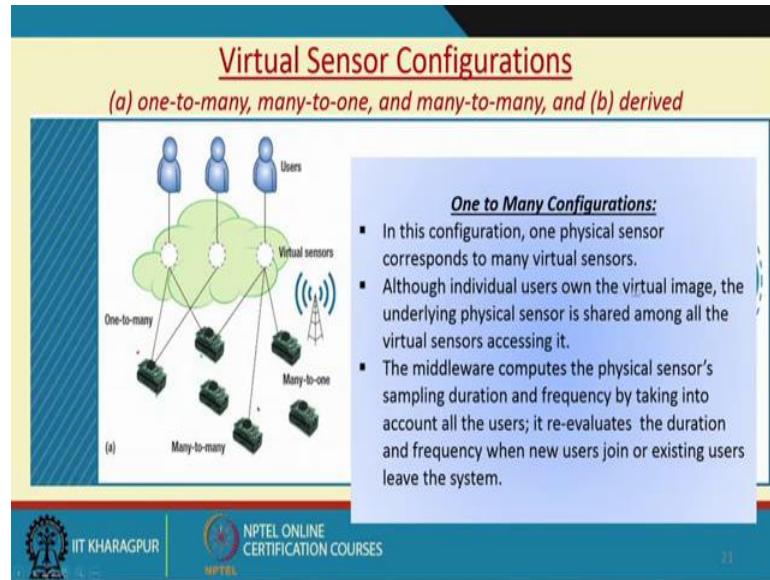
So, this sensor cloud need to be managed appropriately and the something; if we look at the client the portal it goes for a provisioning resource management monitoring virtual sensor group physical sensor groups and. So, and. So, forth this overall architecture if we look at in other way it is; so, client request to they to a particular portal servers which intern go for that process the query it needs to look at the different virtual sensors which can answer to the queries and which intern look for the physical sensors and so on and so forth.

(Refer Slide Time: 28:11)



So, as we have discussed there are virtual sensor configuration it can be one to many-many to one many to many and can be derived, right. So, there are several scenarios many to many one to many, then we have this derive type of scenarios.

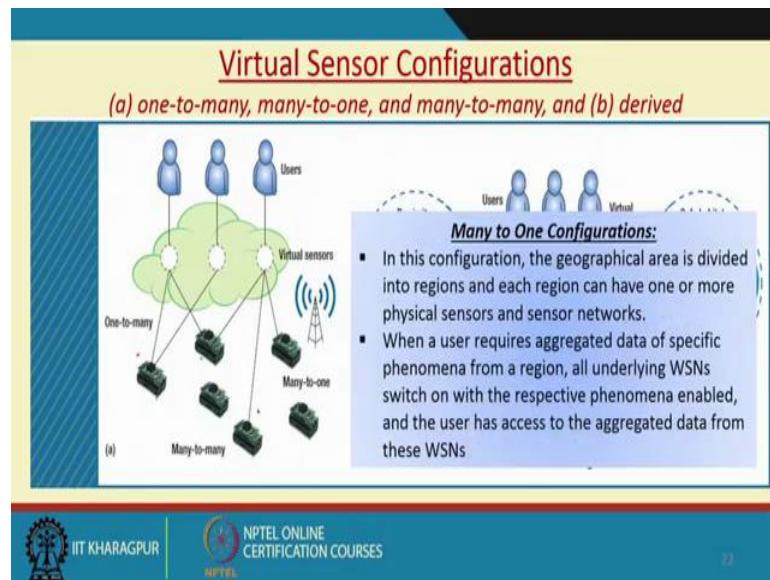
(Refer Slide Time: 28:36)



So, in case of a many to one to many; in this configuration one physical sensor corresponding to many virtual sensors although individual sensors on their virtual images the underlying physical sensor is shared among all the virtual sensor accessing it, right.

So, it is the middle ware which computes the physical sensor sampling duration and frequency by taking into account all the users it reevaluates and duration. So, it is basically the synchronization is ensuring by this underlining middleware and type of things.

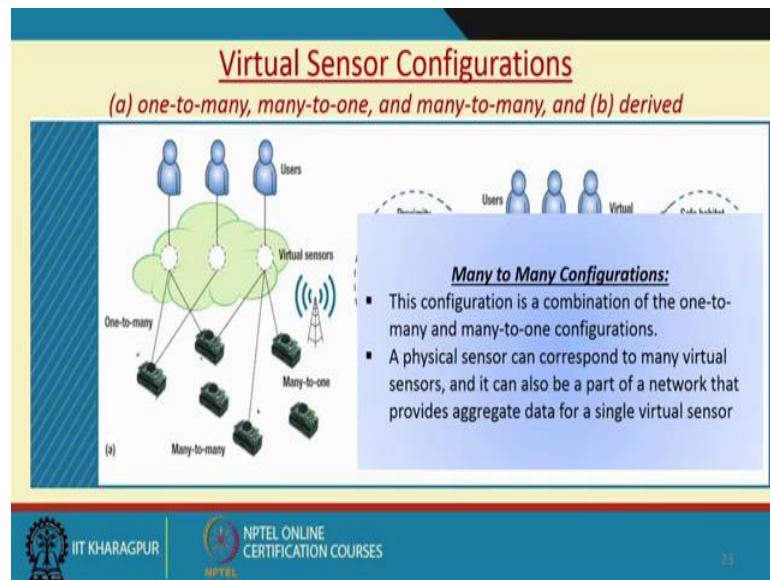
(Refer Slide Time: 29:14)



So, it is a; so, where we have one physical sensors corresponding to more than one virtual sensors; many to one in this configuration geographical area is divided into regions each region have one or more physical sensor and the sensory network when a user requires a aggregate data of specific phenomena from a region all underlining WSNs or value sensory networks which switch on with the respective phenomena enabled and the user can access the aggregated data.

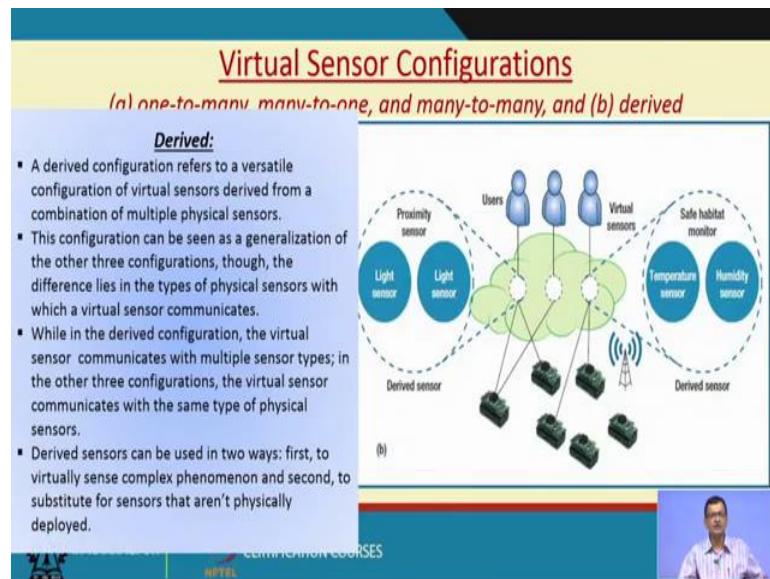
So, here what we have a large geographical space why which is different into regions and which has say for wireless sensor network and if there is a aggregated some processing is required then these are switched on and give a aggregated view on the things, right.

(Refer Slide Time: 29:54)



So, it is a scenario of many to one; we can have as a many to combination of one to and many to one type of scenarios.

(Refer Slide Time: 30:01)



There is a special scenario which is a derived. So, it is a derived configuration refers to a versatile configuration of virtual sensors derived from a combination of multiple physical sensors, right. So, there are virtual sensors which is a combination of the derive sensors like here you say there is a proximity sensor light sensors there are say habited monitor temperatures humidity sensor and these are integrated into to take a call, right.

So, this configuration can be seen as a generalization of other three configuration generalization; basically generalization of the other three configuration lies in the type of physical sensor with a physical sensor communicate; so, it emulates the sensors looking at different other physical sensors. So, it derived a sensing capability from the other physical sensors, right.

(Refer Slide Time: 30:57)

### Virtual Sensor Configurations

(a) one-to-many, many-to-one, and many-to-many, and (b) derived

- Many different kinds of physical sensors can help us answer complex queries. For example: "Are the overall environmental conditions safe in a wildlife habitat?"
- The virtual sensor can use readings of a number of environmental conditions from the physical sensors to compute a safety level value and answer the query.
- If we want to have a proximity sensor in a certain area where we don't have one mounted on a physical wireless node, the virtual sensor could use data from light sensors and interpolate the readings and the variance in the light intensity to use as a proximity sensor.

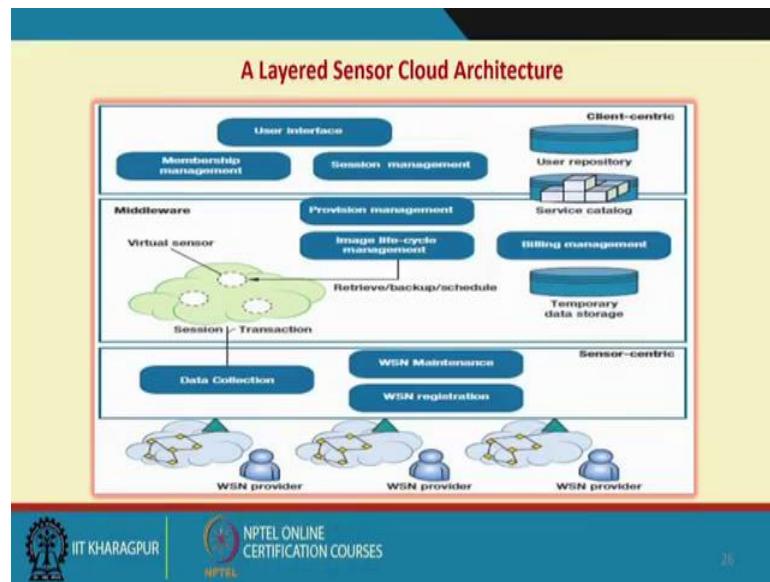


 IIT KHARAGPUR |  NPTEL ONLINE CERTIFICATION COURSES

So, these types of scenarios are being emulated. So, there are there can be many different kind of physical sensor that can help a complex question for example, are overall environmental condition safe in a wild life habitat. So, this is a very complex question, right. There is a there should be a definition of what is meant by the overall condition what is the meant by the safety of a wild habitat and this need to be known.

So, virtual sensor can use reading of number of environmental condition from physical sensor to compute safety level values and answer the query type of things, right.

(Refer Slide Time: 31:39)



So, there is quite pervasive applications which can use this type of scenarios and if we look at a layered sensor cloud architecture. So, we have different wireless sensor providers right they provide data information and type of things we have a sensor centric middleware which is a WSN maintenance WSN registration type of things and then we have a other middleware where we say as a virtual sensors emulating virtual sensor provision management image life cycle and so on and so forth and then at the top interfaces.

So, these are multiple layer one is that physical underlining layer over there one to have this WSN registration WSN maintenance data collection part over there that virtualization layer and then we have that physical user interface to connect to the things and they are we can have other type of things like user repository data repository session management membership management and so on and so forth. So, there are varieties of things which are can be there. So, if we can if we summarize sensor cloud infrastructure virtualizes sensors and provides management mechanism for virtualized sensors.

(Refer Slide Time: 32:38)

## Summary

- Sensor-Cloud infrastructure virtualizes sensors and provides the management mechanism for virtualized sensors
- Sensor-Cloud infrastructure enables end users to create virtual sensor groups dynamically by selecting the templates of virtual sensors or virtual sensor groups with IT resources.
- Sensor-Cloud infrastructure focuses on Sensor system management and Sensor data management
- Sensor clouds aim to take the burden of deploying and managing the network away from the user by acting as a mediator between the user and the sensor networks and providing sensing as a service

»



IIT KHARAGPUR



NPTEL ONLINE  
CERTIFICATION COURSES

Sensor cloud infrastructure enables end user to create virtual sensor group to dynamically by selecting the template of the virtual sensors or the sensor groups which are there. So, based on the user application, I can select a particular template to work on that sensor cloud infrastructure focuses on sensor system management and sensor data management. So, this is more inclined or the vertical towards sensor data management and sensor system management and aims at to take the burden deploying managing network away from the user acting as a mediator between the user and the sensor network. So, that is; it is a sort of connecting user applications with the deployed data or the sensing environment.

So, as we have started today's discussion that that is that is a there is a omnipresent presence of different kind of sensors starting from our mobile phones to different category of sensors which is collecting huge volume of data which could have been used for processing and supporting different user applications. So, that a virtualization mechanism is supported and evolving a sensor cloud which are which may help in addressing different real time applications or serving different real time applications with a variety of heterogeneous sensor collected or sensor acquired data system.

Thank you.

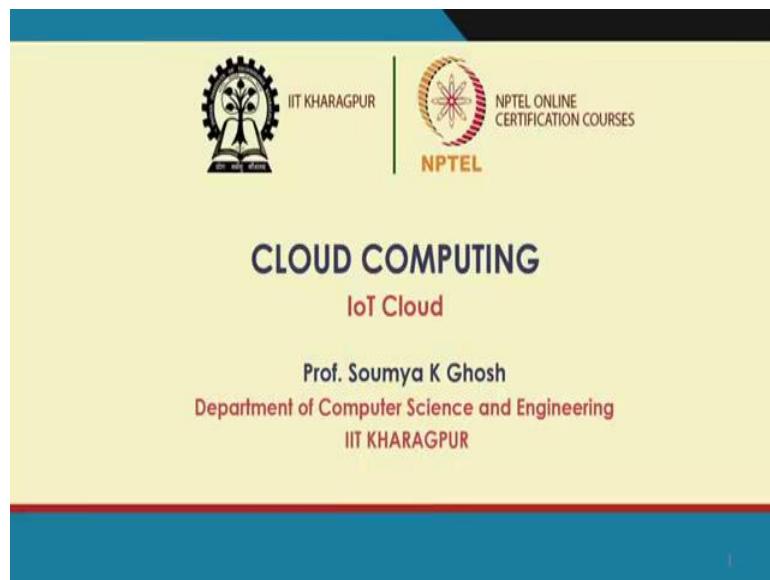
**Cloud Computing**  
**Prof. Soumya Kanti Ghosh**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**

**Lecture – 39**

**IoT Cloud**

Hello, so today we will discuss one of the another aspect of cloud computing rather another company and technology where could computing can be an enabling technology to have better services and better delivery of the services, so that is what we will discussed today about IoT cloud.

(Refer Slide Time: 00:45)



So, as we know that IoT is a buzzword that is Internet of things right so that means, anything rather anything and everything is now becoming Internet enable. So, that may in other sense that there are sensors less we see in a sensor cloud also, and there can be other type of a different variety of sensors which are enabling it to be connected to the Internet. So, anything and everything is connecting to the Internet, and there is a huge volume of data, different variety of services which are being possible with this sort of mechanisms. So, we would like to see that this IoT cloud that is amalgamation of this cloud technology or cloud philosophy with this IoT, how we it is likely to help the overall performance of this services of Internet of things or cloud services.

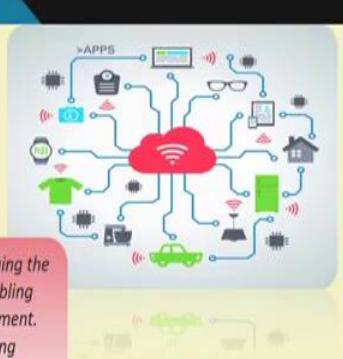
(Refer Slide Time: 01:58)

## Motivation

- Increasing adoption of sensing technologies (e.g., RFID, cameras, mobile phones)
- Sensor devices are becoming widely available

Wireless sensor technology play a pivotal role in bridging the gap between the physical and virtual worlds, and enabling things to respond to changes in their physical environment. Sensors collect data from their environment, generating information and raising awareness about context.

Example: Sensors in an electronic jacket can collect information about changes in external temperature and the parameters of the jacket can be adjusted accordingly



IT Kharagpur | NPTEL ONLINE CERTIFICATION COURSES

Truong, Hong-Linh, and Schahram Dustdar. "Principles for engineering IoT cloud systems." *IEEE Cloud Computing* 2.2 (2015): 68-76

So, if we look at the motivation is clear that increasing adaptation of sensing technology RFID, cameras, mobile phones and everything and anything and everything. So, sensor devices are becoming widely available across different type of mechanisms. So, when we talk about sensor cloud, we are more concerned about that the sensing device itself where what we are looking at these are sensing devices which are ubiquitously available. So, there may be very thin line between things, but it has a different way of handling, and more are some most of the cases are more application oriented type of things.

Like wireless sensor network or technology play vital role in bridging the gap between physical and virtual world, so it helps us to take the physical world to the digitized world and enabling things to respond to change to their physical environment. So, I say that the my ac will be air containing system will be changing based on the temperature of the things right; the temperature sensor activates the air conditioning process or it increases or decreases the temperature or the air conditioning controller system is being controlled or being activated by these sensors. So, this sort of activities, where the sensors in turn activating some other things, so those types of things are there.

So, what we see that sensor collect data from their environment generating information raising awareness about the context. Like we see here there is a variety of sensors right variety of so called objects or things, and they are having different sort of connectivity sometimes where or mostly wireless with these rest of the things. So, sensors in a sense

is like for example, sensor in a electronic jacket, a person wearing a electronic jacket may can collect information about changes in the external temperature and the parameters of the jacket will be sending like, it may the jacket may heat appropriate based on the fall of temperature and so on and so forth.

(Refer Slide Time: 04:28)

## Internet of Things!

- Extending the current Internet and providing connection, communication, and internetworking between devices and physical objects, or "Things," is a growing trend that is often referred to as the *Internet of Things*.
- The Internet of Things (IoT) is a scenario in which objects or people are provided with unique identifiers and the ability to transfer data over a network without requiring human-to-human or human-to-computer interaction.
  - *A thing, in the Internet of Things, can be a person with a heart monitor implant, a farm animal with a biochip transponder, an automobile that has built-in sensors to alert the driver when tire pressure is low -- or any other natural or man-made object that can be assigned an IP address and provided with the ability to transfer data over a network*



So, Internet of things like as we hear about it for in our every literature talking about whenever we talk about sensors and other things. Extending the current Internet and providing connection, communication and intern networking between the devices that is the physical objects or sometimes call things is a growing trend and often referred as IoT or Internet of things.

So, it is a scenario in which objects or people are provided with unique identifiers that means has been identified and the ability to transfer data over a network without requiring to human-human or human to computer interactions so that means, these devices are enable itself to communicate data to the Internet so called. So, thing in the Internet can be a person with a heart monitoring implant or a farm animal with a biochip transponder, or a automobile that is built in sensors to alert driver when the tire pressure low or it is some man functioning some of the any devices or it can be a natural or manmade object that can be assigned in a IP address and so on and so forth. So, that something that is uniquely identified IP address may be one of the mechanism to do that so that it can communicate with these unique things right. In case of a sensor normal

sensor deployment, I may not have a unique identifier. I am more concerned about the data, which is going to the things right. I am more concerned about here though the data is important, but I still I can I uniquely identify a thing or a object.

(Refer Slide Time: 06:20)

## Internet of Things!

- Extending the current Internet and providing connection, communication, and inter-networking between devices and physical objects, or "Things," is a growing trend that is often referred to as the *Internet of Things*.
- The I "The technologies and solutions that enable integration of real world data and services into the current information networking or human technologies are often described under the umbrella term of the Internet of Things (IoT)"  
with a microchip transponder, an automobile that has built-in sensors to alert the driver when tire pressure is low -- or any other natural or man-made object that can be assigned an IP address and provided with the ability to transfer data over a network

Source: 

 IIT KHARAGPUR |  NPTEL ONLINE CERTIFICATION COURSES

So, technologies and solutions that enable integration of real world data and services into current information networking technologies are often described under the umbrella term called Internet of things right at that is the core of the whole thing.

(Refer Slide Time: 06:39)

## More “*Things*” are being connected!

- Home/daily-life devices
- Business
- Public infrastructure
- Health-care and so on...



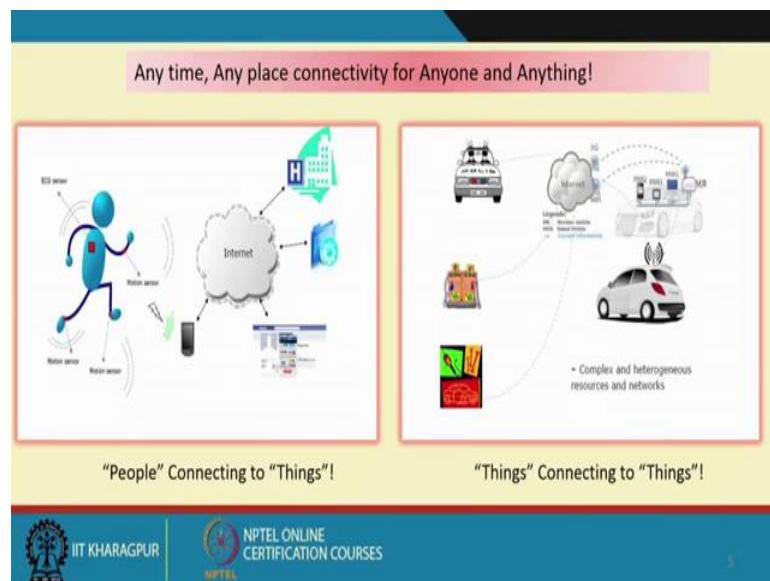
Source: 

 IIT KHARAGPUR |  NPTEL ONLINE CERTIFICATION COURSES

So, day-to-day so called more objects or more things are being connected right. So, we see starting from vehicle to some of the devices in a lab, where the device in a lab or even some filtering mechanisms or a lighting systems or lights right which are based on the things may be controlled by the things all are they all are communicating in case of smart grid also. And these access control and lot many things. So, home daily life devices are getting connected, businesses, public infrastructure, health care and etcetera, etcetera. So, it is a long list and everyday things are being enabled. So, we have the ability to control the devices actuates some other devices based on the sensed by some other some other applications and so on and so forth.

So, here it may not be always sensing. So, it is I can have it to actuate a thing like I can say have a controller circuit which is being actuated by something by external things like I switch on the lights, switch on the fan or activates something by using a application on the mobile. But the light or the electric switch is being is a thing or is the IoT device where it access a IoT device. So, it is it is not exactly sensing in the terms the environment, but in terms it is being activated by some through a Internet.

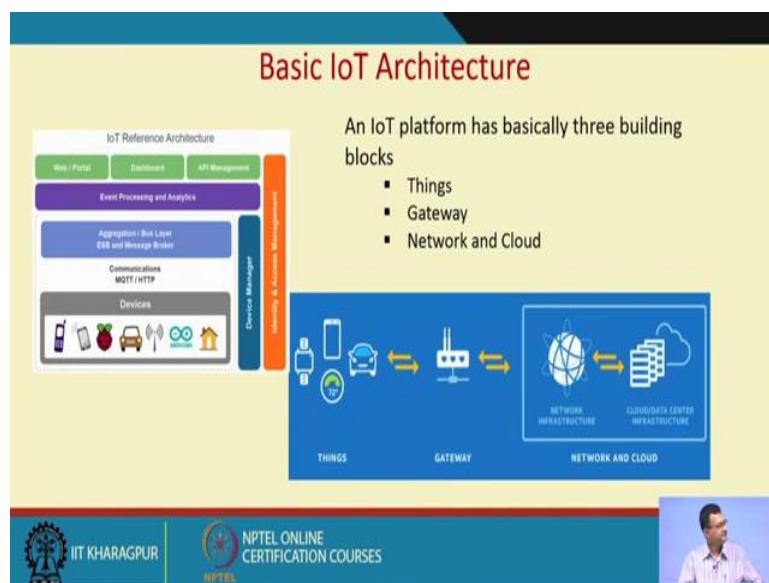
(Refer Slide Time: 08:29)



So, it is in other say I can say any time, any place connectivity for any one and anything. So, something like that right any time, so any place, location, space, time to anyone and anything so any objects and type of. So, let people connecting to things like it is connected may be to the hospitals to other type of a several mechanisms, and type of

things may be through a mobile device or it can be things connected to a things like a car connected to another car or petrol station, or mechanics, or workshops and type of things. So, a particular objects if it is malfunctioning or it triggers something to some other network and so on and so forth. Like if it is low on petrol not only it flashes the message to the driver, also at the same time it searches that what is the nearby petrol or gas station and who can do the things, similarly if requires say something to be done in the workshop and type of things.

(Refer Slide Time: 09:45)



So, if you look at the basic IoT architecture, so what it requires on the things, it is the objects or the needs a gateway to connect to the rest of the world that is network or cloud and type of things. So, it is a things gateway network and cloud. So, whether it is a car or mobile device, a smart watch or a temperature sensor or any type of things which can go through a gateway and connect it connect to the things. So, these gateway provides a connectivity with the rest of the thing rest of the Internet, it can be cloud, it can be other infrastructure and type of things.

So, if you look at so the devices at the bottom end. So, there is a communication path either through MQTT or HTTP some protocol aggregation and bus layer with ESB and message broker. So, aggregation bus layer. So, there may be enterprise service bus and sort of things and message broker. It goes to the event processing and analytics based on that whatever it sense is good for a event processing analytics. And there are web portal

dashboard API management at the top of the layer right. So, these are different structures. So, these are device manager, these are identity and access management at the across the whole vertical. So, this is this is broadly the generic or basic IoT architecture which more or less all devices follows, all devices confirm true that whichever or rather that I should say all things quote, unquote things confirm too.

(Refer Slide Time: 11:40)

## Several Aspects of IoT systems!

- **Scalability:** Scale for IoT system applies in terms of the numbers of sensors and actuators connected to the system, in terms of the networks which connect them together, in terms of the amount of data associated with the system and its speed of movement and also in terms of the amount of processing power required.
- **Big Data:** Many more advanced IoT systems depend on the analysis of vast quantities of data. There is a need, for example, to extract patterns from historical data that can be used to drive decisions about future actions. IoT systems are thus often classic examples of "Big Data" processing.
- **Role of Cloud computing:** IoT systems frequently involve the use of cloud computing platforms. Cloud computing platforms offer the potential to use large amounts of resources, both in terms of the storage of data and also in the ability to bring flexible and scalable processing resources to the analysis of data. IoT systems are likely to require the use of a variety of processing software – and the adaptability of cloud services is likely to be required in order to deal with new requirements, firmware or system updates and offer new capabilities over time.

 IIT KHARAGPUR |  NPTEL ONLINE CERTIFICATION COURSES

So, there are several aspects of IoT systems, one is scalability issue that scale for IoT system implies that in terms of number of sensors and so on and so forth, if it is increasing then how to handle. That is a big data issue that there are so many IoT devices and many of them are having sensing or data acquisition type of things, or event transmitting the data that is a huge volume of data which needs to be processed and there is a big data analysis and processing issue. And we find a role of cloud computing out here, so that not only in terms of the scaling in terms of IoT frequently involve use of cloud computing platform of our potential of large amount of resources, both in terms of storage, data and also ability and flexibility to scaling operations. So, there is a need for aspects of cloud type or cloud sort of mechanisms.

(Refer Slide Time: 12:40)

## Several Aspects of IoT systems (contd...)

- **Real time:** IoT systems often function in real time; data flows in continually about events in progress and there can be a need to produce timely responses to that stream of events.
- **Highly distributed:** IoT systems can span whole buildings, span whole cities, and even span the globe. Wide distribution can also apply to data – which can be stored at the edge of the network or stored centrally. Distribution can also apply to processing – some processing takes place centrally (in cloud services), but processing can take place at the edge of the network, either in the IoT gateways or even within (more capable types of) sensors and actuators. Today there are officially more mobile devices than people in the world. Mobile devices and networks are one of the best known IoT devices and networks.
- **Heterogeneous systems:** IoT systems are often built using a very heterogeneous set of. This applies to the sensors and actuators, but also applies to the types of networks involved and the variety of processing components. It is common for sensors to be low-power devices, and it is often the case that these devices use specialized local networks to communicate. To enable internet scale access to devices of this kind, an IoT gateway is used



IIT KHARAGPUR



NPTEL ONLINE  
CERTIFICATION COURSES

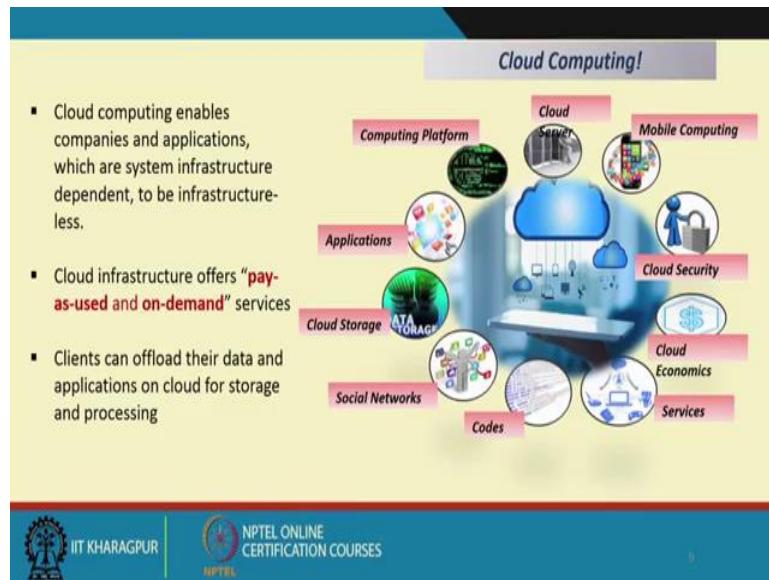


There are few more things many of the cases it is a real time phenomena. So, the processing etcetera requires some sort of a real time intervenes. So, IoT systems often function in real time, data flows continually about events progress and there can be a need to produce timely responses to the stream of the events, so that data being collected being processed to be need to be actuated at the at a real time. Highly distributed as we look at that whole lot of devices a variety of devices, it can span over a from a single room to buildings to even to a large much more larger geographical area like a campus or a city and type of things. So, it is a highly distributed and as it is distributed automatically heterogeneity come into play like there can be different objects, different type of devices and the type of data they transmit or consume or is they there can be a variety or heterogeneity on those type of things.

So, IoT systems are often build using very heterogeneous set of different devices and type of things. These applies to sensors to actuators, but also applies to type of network involved in the variety of progress like some may using WiFi some may ZigBee Bluetooth and type if things and it a whole lot of interoperability issues need to be addressed out there so that in order to address this sort of issue, this IoT gateways are required, so that it is addressed at a more localized fashion. So, whatever it communicates, it communicates with the gateway which is more localized fashion right say I say that there are several IoT devices in this room and I have a gateway which take care of this devices. So, rest of the world or outside the room they are not bothered about

this how the heterogeneity of the data formats, they are processing mechanisms, the communication paradigm which they do inside the things, the gateway takes care and the gateway represents or handles the interoperability or the heterogeneity issue of this IoT devices.

(Refer Slide Time: 15:19)



On the other hand, we have seen a lot of a cloud computing aspects. So, computing enables companies or enterprises applications which our system infrastructure independent to be infrastructure free that is one aspects. It offers pay as used and on demand services, which is pretty amicable for this type of IoT things. Clients can upload the data and application on cloud for storage and processing that is another these are the issues which are which provides. And as we have seen already that there is a number of applications or number of features of cloud right like a mobile computing to aspects of security to social networking and so on and so forth.

So, cloud computing enables services to be used without any understanding of the infrastructure. So, if it is the service provider takes SaaS type of things anything say at the SaaS level its do not bother about the what is the underlining infrastructure or the infrastructure or the platform and so on and so forth works using economy of scale. So, it has a economy what we say it works with the using he economy of scales, and it is extremely beneficial for several applications specially this sort of applications where these IoT type things are there. Data and services are stored remotely, but accessible

from anywhere, like as we are seeing that anywhere, anything anybody type of things, so this is again amicable for this IoT operation or IoT paradigm.

(Refer Slide Time: 17:11)

## IoT Cloud Systems?

- Recently, there is a wide adoption and deployment of Internet of Things (IoT) infrastructures and systems for various crucial applications, such as logistics, smart cities, and healthcare. This has led to high demands on data storage, processing, and management services in cloud-based data centers, engendering strong integration needs between IoT and cloud services.
- Cloud services are mature and provide excellent elastic computation and data management capabilities for IoT. In addition, as IoT systems become complex, cloud management techniques are increasingly employed to manage IoT components
- Thus, cloud services now act as computational and data processing platforms as well as management platforms for IoT. From a high-level view, IoT appears to be well-integrated with cloud data centers to establish a uniform infrastructure for IoT Cloud applications

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, why not then IoT cloud, so that is the things which naturally evolved; it is not like that we some we are trying to put them together, it is natural evolution to or merging of this two concept of the things. So, there is a wide adaptation and deployment of Internet of thing infrastructure what we see for various crucial application like logistics, smart city health care say even parking car parking and so forth. So, this led to high demand on data storage processing and management in the cloud based a data centers right. So, this IoT devices individually handling will be difficult and even then you have to have something some infrastructure which can handle those things why not the cloud.

Cloud services are mature and provide excellent elastic computation and data management scalabilities for IoT that is it has the potential to give data management scalability in addition and as IoT systems become complex, cloud management techniques are increasingly employed to manage IoT components. So, IoT systems not only the hardware wise, the systems in the processing wise are becoming complex. Cloud services now act as a computational data processing platform as well as management platform for IoTs right. So, the cloud platform can very well be a processing platform as well as data management platform for this type of IoT systems

may be in a localized manner may be more globalized manner or on a larger geographical spread or more complex type of situations.

(Refer Slide Time: 18:53)

## IoT Cloud Systems?

- Recently, there is a wide adoption and deployment of Internet of Things (IoT) infrastructures and systems for various crucial applications such as logistics, smart cities, and healthcare. This integration between IoT and cloud services allows coordination among IoT and cloud services. That is, a cloud service can request an IoT service, which includes several IoT elements, to reduce the amount of sensing data or the IoT service can request cloud services to provision more resources for future incoming data management platforms for IoT. From a high-level view, IoT appears to be well integrated with cloud data centers to establish a uniform infrastructure for IoT Cloud applications

Well as



IIT KHARAGPUR

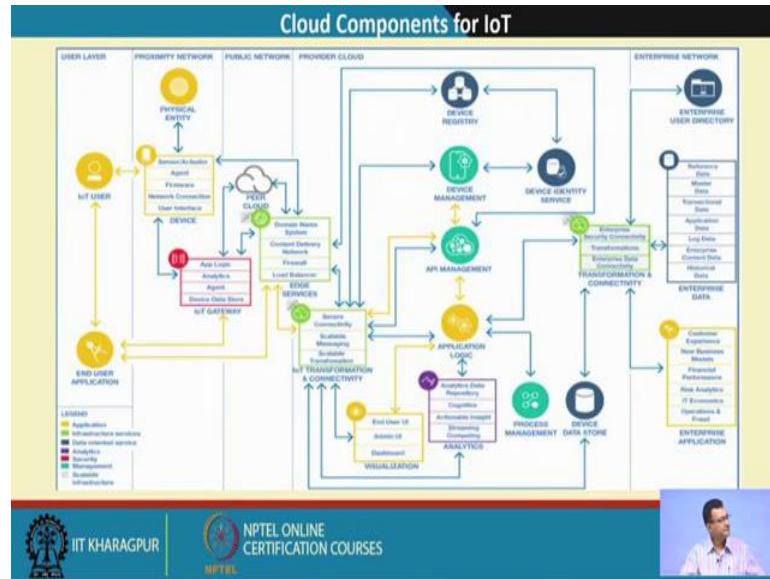


NPTEL ONLINE  
CERTIFICATION COURSES



So, what we see that integration between IoT and cloud services allowed coordinating coordination among IoT and cloud services in a seamless way. That is, a cloud service can request an IoT service which include several IoT elements to reduce amount of sensing data or the IoT service can request cloud services to provision more resources etcetera. So, it is what we say sort of a quote unquote helping each other to provide better services to enterprises to individuals or other type of services which uses this IoT services.

(Refer Slide Time: 19:37)

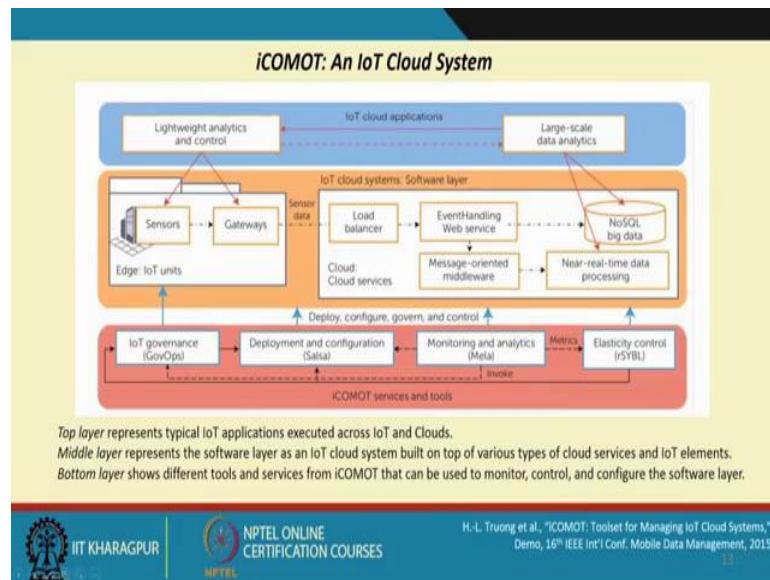


So, if we look at try to look at the cloud components for IoT, so there are you can see there are several verticals one is the user layer, where in end user applications or IoT users are there then we have a proximity network which are more near to these user layer. There are sense sensor actuators, agent, firmware, network connectivity, user interface, devices and so on and so forth. Here we have this IoT gateway, which connects this part with this public networks. So, the IoT devices or applications which are connected by the proximity network which can be again having heterogeneous type of things different IoT devices can do can connect to the things this IoT gateway connects to the rest of the world. So, peer cloud through a peer cloud to the other type of things.

So, what it goes as a application logic analytics agent device data store, these IoT gateway connect to this may be a DNS system or a content delivery CDN, firewall local load balancer and type of things. So, this in turn connects to this things which connects it to the provider cloud or a CSP - cloud service provider, which has all its cloud components already we know that can be a device registry, device management, API management which are there in that IoT cloud. So, application logic which helps in doing analysis analytics where the analytics data repository cognitive applications or cognitive mechanisms actionable inside streaming computing all this analytics related things can be there right and these interns can connect to a enterprise network. So, it has a user directory and so on and so forth.

So, if you see there are different layers. So, user layer proximity network through a gateway connected to the public network, these connect to a provider cloud and if required it goes to a enterprise network for accessing other type of services like I say that that I have a car, which detects some anomaly. It uses some gateway to connect to the cloud it with different type of not only from the or not only from different data from the cloud other environmental data these does some analytics then it may trigger to that particular car manufacturer or car related service provider or workshop. And goes to that enterprise and then gets a feedback and so on and so forth. So, make some, what we say corrective actions on the things like providing may be a service vehicle or sending alert to the driver and so on and so forth. So, there can be this sort of aspects.

(Refer Slide Time: 22:55)



So, there is iCOMOT. So, one it is from particular one publications an IoT cloud systems, what we see that different layers right at the top layer represent a typical IoT application executed across IoT and clouds. So, it is the top layer which is a typical light weight analytics and control, large scale data analytics. So, this is at the top layer. The middle layer represents the software layer as an IoT system build on the top of various type of cloud services in the IoT elements. So, this is the middle layer where the sensors, the gateways on the IoT client side, there is load balancer, event handler, message oriented middle ware and there are other cloud related databases like no SQL databases, near real time a real time data processing.

At the bottom layer shows the different tools and services from iCOMOT that particular framework that can be used to monitor control configure the software layer. So, this is more for deploy, configure, govern and control at different type of services. So, this is a typical one example of an IoT clouds system or IoT cloud framework.

(Refer Slide Time: 24:16)

Infrastructure, Protocols and Software Platforms for establishing an Internet of Things (IoT) Cloud system			
Types	IoT	Clouds	Purpose
Infrastructure machines	Industrial and common gateways (for example, Intel IoT Gateway) and operating system containers (such as Dockers)	Virtual machines and operating system containers	Enable (virtual) machines where software components will be executed
Connectivity protocols	Message Queue Telemetry Transport (MQTT), Constrained Application Protocol (CoAP), HTTP, control area network (CAN) bus	MQTT, Advanced Message Queuing Protocol (AMQP), HTTP, and so on	Enable connectivity among IoT elements and between the IoT part and cloud services
Platform software services	Lightweight data services (such as NiagaraAX/Obix), lightweight complex event processing (CEP) and data fusion, topology description and deployment service (such as TOSCA), and lightweight application containers (such as OSGI and Sedona)	Load balancers (such as HAProxy), message-oriented middleware (MOM) (such as ActiveMQ and Kafka), NoSQL, stream/batch processing (such as Hadoop and Spark), component repositories/marketplaces, and deployment services (such as TOSCA, HEAT, and Chef)	Enable core platform services for IoT and cloud tasks





Truong, Hong-Linh, and Schahram Dustdar  
“Engineering IoT cloud systems.” IEEE Cloud (2015): 68-76.

So, there are if we look at the infrastructure protocol and software platform for enabling an Internet of thing cloud systems. So, infrastructure wise in IoT industrial and common gateways, for example, Intel IoT gateway may be thing and operating system containers such a Dockers these are aspects. In case of a cloud virtual machines and operating system containers, so purpose is to enable virtual machines where software components will be executed.

So, connectivity protocol MQTT that is message q telemetry transport constant application was CoAP, HTTP different protocol which is there in the IoT for connectivity protocol in the case of cloud also we have MQTT, AMQP and HTTP and so, for the purpose is enable connectivity among IoT elements between IoT part and the cloud services. So, it this connectivity protocols allows that how to connectivity will be there. There are platform software services which enable core platform services for IoT and cloud task and we have different type of services they are light weight data services whereas, there are load balances such as ha proxy and so on and so forth. So, that means,

for if we look at the different types of infrastructure protocol and software platform which are there is a coupling between this IoT and the cloud to have a unified reason.

(Refer Slide Time: 26:00)

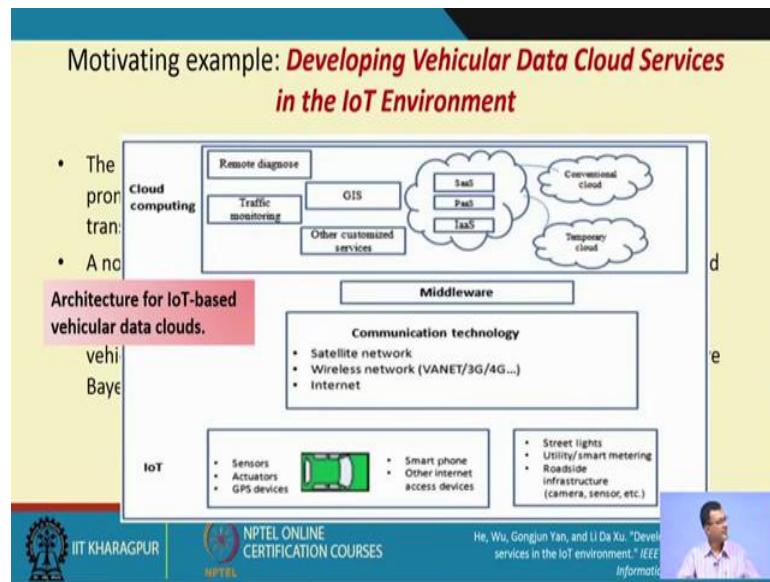
Motivating example: ***Developing Vehicular Data Cloud Services in the IoT Environment***

- The advances in cloud computing and internet of things (IoT) have provided a promising opportunity to resolve the challenges caused by the increasing transportation issues.
- A novel multilayered vehicular data cloud platform by using cloud computing and IoT technologies is presented. Two innovative vehicular data cloud services, an **intelligent parking cloud service** and a **vehicular data mining cloud service**, for vehicle warranty analysis in the IoT environment are also presented using a Naive Bayes model and a Logistic Regression model

 IIT KHARAGPUR |  NPTEL ONLINE CERTIFICATION COURSES |  He, Wu, Gongjun Yan, and Li Da Xu. "Developing vehicular data cloud services in the IoT environment." IEEE Transactions on Information Technology in Medicine and Biology, 2018.

So, there is a small example that developing regular data clouds services in IoT environment. So, advancement cloud computing and IoT have provided promising opportunity to resolve challenges caused by the increasing transportation issue. So, transportation worldwide is a major challenge that how to have safe driving with proper traffic management and so on and so forth. So, multi layered regular cloud platform by using cloud computing and IoT technology has been presented in this work. Two innovative vehicular cloud services like intelligent parking services and vehicular data mining cloud services for a vehicle warranty analysis etcetera are some of this motivating example.

(Refer Slide Time: 26:49)



So, if we look at that one side these IoT things are there, we it is architecture IoT based vehicular data cloud. So, different sensor actuator GPS devices smarts phone and other Internet services and we have things like street lights utilities, smart metering, road side infrastructure and so and so forth. So, these are different IoT devices. On the other hand, we have this cloud like remote diagnosis traffic, geographical information system, SaaS, PaaS, IaaS and other things. So, we have a middle layer like which allows this communication technology like satellite network if it is communicating wireless network, like vehicular ad hock network, 3G 4G services and Internet. So, this is allowing this is enabling this merging of information between this IoT devices and cloud for proper decision making or what we can say that meaning full decision making or some sort of a real time or real time decision making.

(Refer Slide Time: 27:58)

**Services for IoT-based Vehicular Data Clouds**

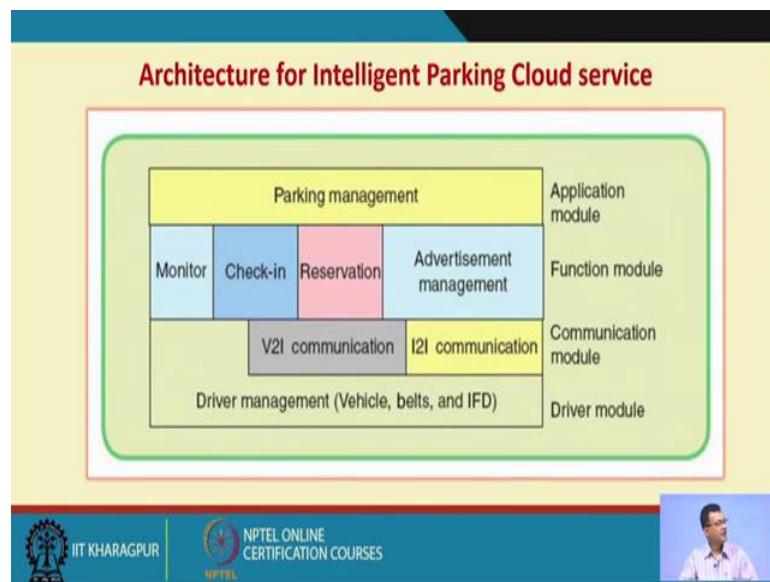
New services	Description
Network and Data Processing as a Service, i.e., Infrastructure As A Service (IAAS)	Vehicles provide their networking and data processing capabilities to other vehicles through the cloud
Storage as a Service (SAAS)	Some vehicles may need specific applications that require large amount of storage space. Thus, vehicles that have unused storage space can share their storage space as a cloud-based service
Platform as a Service (PAAS)	As a community, vehicular data clouds offer a variety of cooperative information services such as traffic information, hazardous location warning, lane change warning and parking availability

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES



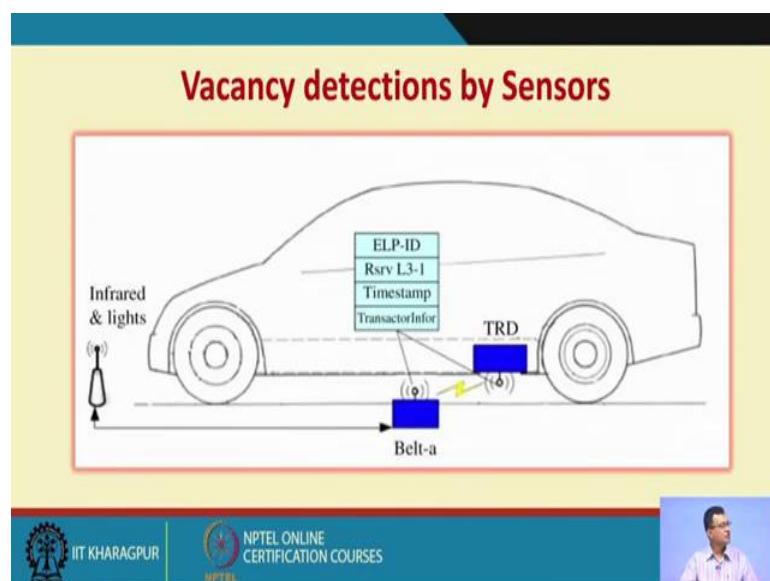
So, if we look at the services of IoT based vehicular data cloud, so there are network and data services as a network can data processing as a service. So, this is a new type of service like vehicle provides their networking and data processing capability to other vehicles through the cloud. There can be storage as a service that is some vehicle may need specific application that require large amount of storage space or a fairly good amount of storage space, which is not there in the may be the OBU or the onboard unit thus the vehicle have the unused storage space can share their storage like other vehicles. Platform as a service as a communicative as a community vehicular cloud offer a variety of cooperative information services such as traffic information hazard location warning lane change warning and so and so forth, what we say in the vehicular ad-hock network terms as a safety related services.

(Refer Slide Time: 29:05)



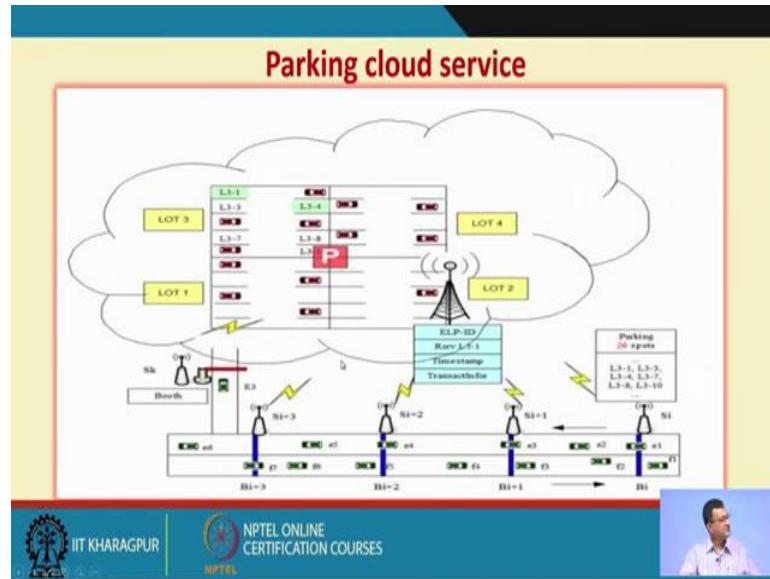
So, similarly another application is intelligent parking using IoT cloud service. Here also we have several things like starting from driver management by vehicle belts IFDs to vehicle to infrastructure, infrastructure to infrastructure, communication there is parking management, monitor checking, reservation advertisement, so there are application module, functional module, communication module and driver module. So, there are different type of modules which are there.

(Refer Slide Time: 29:45)



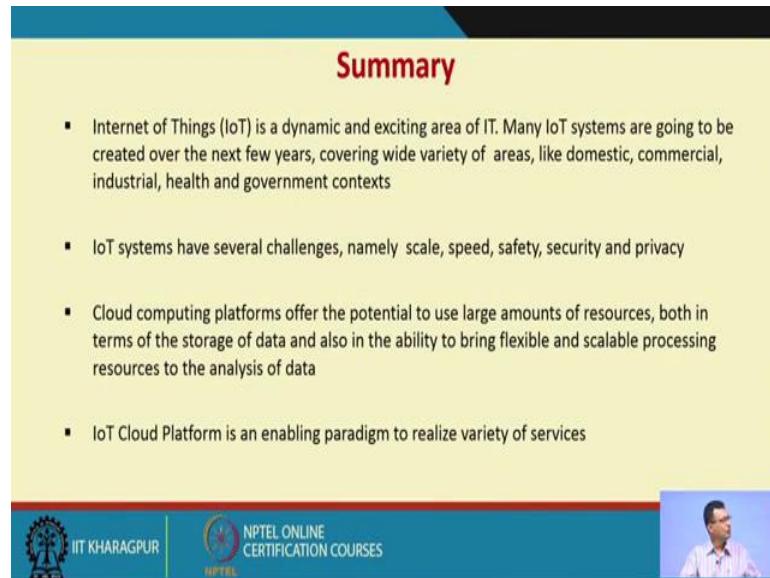
So, a particular car can detect by sensor using it different time stamping that which.

(Refer Slide Time: 29:54)



Lane or which lot there is a vacancy and type of things and before approach the pipe parking lot, enabling these IoT through gateway to the cloud can find out that where the vacancies are there. So, this can be one way of looking at that application which has a direct implication of our day-to-day life.

(Refer Slide Time: 30:23)



So, to summarize IoT is a dynamic and definitely exciting area. IoT systems are going to be created more and more IoT systems are going to come, and both domestic, commercial, industrial, health and different context and it is going to have a large

amount of processing storage requirements. And it faces thus it faces several challenges like scaling up or a speed of processing safety, security, privacy. On the other hand, cloud computing platforms offer potential to use large amount of resources both in terms of storage processing and have a flexible scalable processing infrastructure.

So, what we see these IoT cloud platform is going to be a enabling technology for our or several futuristic applications, it may be personalized application to some commercial applications to industrial applications, and several type of applications and processing need. So, that is that may be the next things which is coming into the things where it encompasses the cloud, IoT, sensors and different aspects of devices or which can communicate and produce data and communicate to the Internet. So, with these we will stop here today on this particular aspect of IoT cloud.

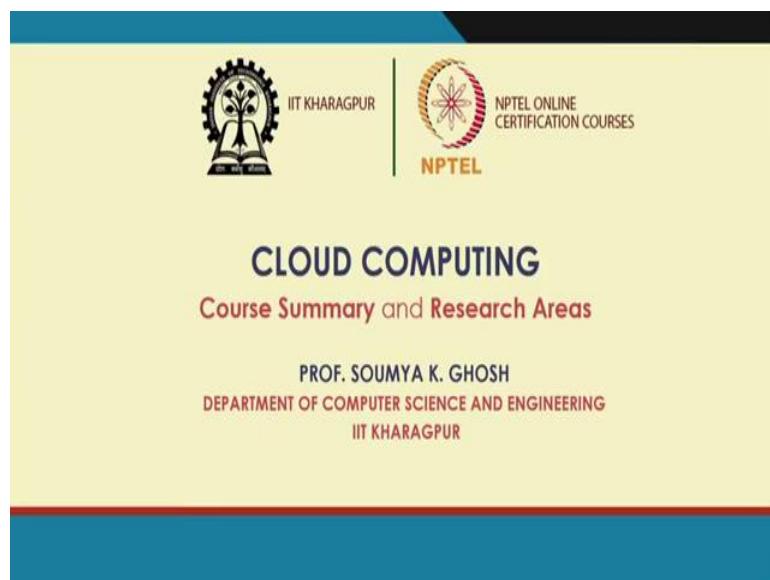
Thank you.

**Cloud Computing**  
**Prof. Soumya Kanti Ghosh**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**

**Lecture – 40**  
**Course Summary and Research Areas**

Hello. So, welcome to our final lecture on cloud computing course. So, today what we are trying to discuss is we will be discussing is primarily looking at; what we tried to cover in this course and also we will give you an quick overview or of that what are the possible openings or what are the possible research directions which you can explore or you can study more if you are interested in this particular field, right.

(Refer Slide Time: 00:57)



(Refer Slide Time: 01:05)

## Course Summary

- Introduction to Cloud Computing
  - Cloud Computing (NIST Model)
  - Properties, Characteristics & Disadvantages
- Cloud Computing Architecture
  - Cloud computing stack
  - Service Models (XaaS)
  - Deployment Models
- Service Management in Cloud Computing
  - Service Level Agreements(SLAs)
  - Cloud Economics
- Resource Management in Cloud

 IIT KHARAGPUR |  NPTEL ONLINE CERTIFICATION COURSES

So, in this particular course, we try to discuss on some of the topic or some of the areas of cloud computing as we all understand; there is a vast area; vast area of things, rather if you can see that every topic itself is a course by itself. So, what my major effort was to give you an overview and what are the possible challenges; what are the different characteristics of cloud or different properties of cloud and what are the different aspects; we tried to look into and which service that technology involved into things.

So, we tried we started with basic introduction with cloud computing with a basic NIST definition and which what we emphasize that it everything as a service. So, and then try to see that the basic property characteristics, rather advantageous and disadvantages of such type of computing as we discussed and try to reiterate that it is not suddenly a new technology which suddenly drop from the sky. So, there is a evolution process, we started, there are several effort towards this sort of activity, like we have gone through phrases of cluster computing, grid computing, distributed computing and finally, came up to this cloud computing.

The basic beauty or the basic characteristics of this sort of computing as we have seen is that it is what we call a form of a utility computing like as a utility services like electricity or water or anything, you can basically pay as you go model, right. It is a metered service, scalable service; you can scale to at infinite scaling and the other thing

is that it do not have to maintain the infrastructure at your end. So, you need somewhere as to hook into the thing and then start computing.

And also we have seen that if basic architecture and try to look at the basic cloud computing stack, there the popular service models like a though it is a anything as a service, we have seen primarily infrastructure as a service, platform as a service and software as a service, the primarily these 3 there are different other type of service anything which can be manifested it as a service mode is a is possible in this sort of framework right.

So, rather though it is not mentioned explicitly in the slide, what we have seen that the basically the foundation technologies like service oriented architecture web services and xml technology. So, we have gone through those things also, right, we have taken an overview of those things also which helped us into interoperate and make these service oriented architecture possible. So, this basic service oriented architecture or service oriented approach is the; what we can say foundation of or another pillar of this cloud computing architecture.

Then also we have seen some aspects of service management primarily with respect to service level agreements, right; SLAs which plays a important role because as a loosely couple heterogeneous services are talking to each other; these service level agreements plays a important role in realizing this sort of framework right, unless we have a strong service level agreement or then, otherwise it is very difficult to have a to maintain a quality of appropriate quality of services, right.

So, whenever someone is taking is leveraging cloud, then it is a organization or individual is expecting some sort of a reliability from the cloud, right like if I am taking if on behalf of our institute for running a lab; we are taking outsourcing our system to the cloud, then what I am expecting that there will be a faithful uptime, right, it is a of a particular level says I 99 percent; 99.9 percent and so and so forth which will be supported by the cloud.

If it is more missing critical operations like bank or similar financial or even some of the say defense or disasters management or those type of operations, then we what we require is more or health care systems, then why what we require is more high level availability and reliability on the things. So, how to bound that whether what I am paying

for or what I have agreed upon whether I get that services. So, for we require a strong SLAs and that every clouds service provider specially, these commercial provider have provide a SLA format though there are what we have seen there are differences between the formats and type of things, but it is plays a important role in our in having cloud services.

Similarly, we talked cloudonomics is cloud is always a good thing, right. So, what is the economy behind the cloud, when we should go for cloud, when we should not or stick to our own in house computing facility, whether it is infrastructure platform and so and so forth. So, though so, if there is a particular economy which comes into play and we have seen that we need to look at that those economic aspects whenever we are going to basically outsource some of my business or part of my business to cloud, right.

So, cloudonomics specifically what economy in cloud what we have looked into. There can be other aspects even cloud may be economically beneficially, but my other constraint like may be security constraint, my privacy constraint may not allow me to push all the data to the cloud right like IIT Kharagpur may decide that I can run the labs on cloud, but I may not put my student data, employee data on cloud because finally, it is on the third part domain. So, that type of thing also come into play. So, it is not only whether; it is economical in terms of money it is also to be that what should be the organization policy individual policy; how much I should push on the cloud and so and so forth.

So, never the less; cloud economics plays the important role to decide when to go for cloud and when to take it on in a house type of things. Then we have discussed I believe one or 2 lectures on cloud resource management in cloud, right. So, it is a very important issue, right resources resource management in the cloud is play a both from the providers and particularly also sometimes it is from the consumer end, but never the less for the resource management; the resources management in from the provider sense plays a vital role, right. Like I have though I say that I have infinite resources as my backend, right, but it is, but it is tangible, right, I cannot say that it is not there is a limitation of the things not only that resource also takes lots of other energy, right.

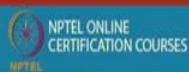
So, it comes with back to back that some sort of a consumption and so on and so forth like even I have; I can give a number of there is no constant the resources is to maintain

the resources; I need to have lot of energy at the things. So, appropriate resource management. So, that overall it is beneficial in terms of that say profitability of the service provider, it is not adverse to our environment or consuming more energy, this appropriate resource management plays an important role. Not only that if you see the literature or recent works in several that peer reviewed journals or top level conferences this resource management is always a plays a important part of that.

(Refer Slide Time: 10:16)

## Course Summary (contd.)

- Data Management in Cloud Computing
  - Data, Scalability & Cloud Services
  - Database & Data Stores in Cloud
  - GFS, HDFS, Map-Reduce paradigm
- Cloud Security
  - Identity & Access Management
  - Access Control
  - Trust, Reputation, Risk
  - Authentication in cloud computing
- Case Study on Open Source and Commercial Clouds
- Research trend - Fog Computing, Sensor Cloud, Container Technology, Green Cloud etc.



So, this is a very vital aspects in case of a cloud management or cloud computing. Then we have seen data management in cloud, it is a very tricky issue. Finally, whenever the data is in cloud; it is your data or our data is on third party, right. So, I do not know that what is happening to the data though; there are definitely security issues apart from the security issues, there are several issues like how whether what will happen that data if it is lost or something how what about the scalability of the data if the quickly I can grow on the data, what should be accessibility of the data is there in the cloud, it is not cloud storage is not lost however, I do not have appropriate bandwidth to access the data.

So, there are several issues in case of data management; we have seen some of the things like Google file system; HDFs, Hadoop file systems and what we have seen that that how the application policies works and how this huge data thing are maintained not only that it also while talking to this data management the one thing in variably come is that big data management how this big data can be managed in cloud, right. So, we have seen

that other technologies like a map reduce sort of technologies where we can paralyze this operations and do and can have a better efficiency or of processing a particular work in the cloud.

So, this is very again, they are very interesting area to work on or study more and type of things, then of course, one major concern in cloud is cloud security; we have also tried to looked into some aspect of cloud securities tried to show some of the recent train in the things; we have discussed these in 3-4 lectures. So, there are several aspects of the things like one is identity and access management, access control mechanism; whether it should be role base access control or use of risk base of access control and type of things. There are different aspects like trust, reputation, risks competence and type of things, right.

So, how to handle them; how to take a call considering all the things; how to further; for example, how do I calculate trust of a particular provider or even how the provider try to profile his own customer, right. So, there are major issues in their definitely authentication is another major aspect; how to authenticate, how the authentication protocol will run on the things and as we understand that security in say data security in cloud is also a major aspects like it may not be always possible to encrypt the data. Even if you encrypt the data, how this key management will work on the things, right your data is in third party aspects and then how these overall key management aspects will work in this type of there.

So, this is also a very strong field because as more organization, more individuals are going towards this sort of infrastructure pushing their data, their applications. services on the cloud. So, how this security is maintained or how this security data security can be guaranteed becomes a major resource. As we have seen in our country also now several major competitive exams are being conducted over clouding infrastructure.

Now, those securities like how this question paper will be encrypted; how the answers will be encrypted and securely transmitted and if it is stored in the third party whether there is a possibility of any leakage of any data. So, there that needs lot of study and what we have seen that is a major one of the research area also and one of the most what we say talked about concerned about our challenges in cloud.

Also we tried to show you some of the open source as commercial cloud, right, though it is not possible to cover everything, but we have tried with open stack and also we have

shown how in our IIT Kharagpur, we have implemented with help of students; a experimental cloud called Meghamala and how it is serving to a research community though in a small scale, but it is a operational thing though it is experimental. So, it is based on totally open source things on open stack, and there are several open commercial cloud like Amazon, Azure, Google cloud platform, IBM Bluemix and so on and so forth.

We have tried to show some example cases. So, that it will gives you a opportunity to quickly know about the things and work on this type of have a feel on the things, the basic idea is to have a feel of the things and finally, we last few lectures we discussed about some of the related technologies and some of the what we say companion technologies to cloud computing like one is fog computing. So, what we are seeing that the devices in recent days or last one decade are becoming more resourceful. So, that there is some sort of a it is not only acquisition of information and forwarding the information, we can basically have the opportunity to process the information.

So, if that is the situation why not do some operation at much lower level, and then we can provide the result to the things. This will not only help us in reducing the bandwidth requirement to connect to the cloud, but also some of the local decisions which could have been done at a local thing; we can basically achieve that this will help us in having more quote unquote some sort of a real time application to run more faith fully, right.

So, this some of the aspects; we have tried to see in then we have talked about sensor network sensor cloud, right that as we have sensors are omnipresent these days. So, why not if there is a formation of a sensor cloud. We talked about container technology like specifically on the Docker technology that how it can be used for in the context of cloud computing; how this cloud and this container technologies there and of course, this green cloud or like that it comes to that energy management, resource management of the cloud can be there.

So, some of the things what we have tried and what we have discussed may be discussed which may help you in finding your future research directions. Now what we tried to again I let me repeat what I tried to do in this course is to basically open up these different avenues, right. So, it is within some of this sort time period, it may not be possible to discuss everything in details, but definitely we see that lot of opportunities are there.

Now, let me with given these backgrounds. So, let just discuss about some of the research areas in cloud you may find in the internet; I do have found some of most of the things in the internet and like. So, just want to discuss some of these aspects of the things.

(Refer Slide Time: 19:01)

## Cloud Infrastructure and Services

- Cloud Computing Architectures
- Storage ad Data Architectures
- Distributed and Cloud Networking
- Infrastructure Technologies
- IaaS, PaaS, SaaS
- Storage-as-a-Service
- Network-as-a-Service
- Information-as-a-Service



5

So, one definitely if we look at the course and things, one definitely this infrastructure and services plays a important role like a cloud computing architecture storage data architecture distributed and cloud networking infrastructure technologies these part a important role and it is sort of a with the hardware and other related technologies upcoming technologies coming in to play.

So, it is never a saturation field, right it is always there is a scope of contributing into the things and making the infrastructure more intelligent and meaning full. Apart from there are services like there are basic services like IaaS, PaaS, SaaS which are also evolving every day or in a regular fashion; there are other type of services which are becoming very popular one is storage as a service again very tricky and critically issue critical service which everybody needs from the individual to organizations every federal agencies everybody needs it and there are other services like network as a service like I want to setup a network or based under that rather I can have different network; I want to have different network configuration at different part of the day or different time period based on my requirement or my clients requirement, I want to set up a things.

So, what I require instead of giving a physical network, I want to have a network as a service type of things; there can be other type of services like a information as a service. So, information being accessed as a service even people talked about concept like science as a service and type of things, right. So, there are there are lot of opportunity both research study and type of higher studies in the area of cloud infrastructure and services. So, this is one of the important core fields.

(Refer Slide Time: 21:09)

## Cloud Management, Operations and Monitoring

- Cloud Composition, Service Orchestration
- Cloud Federation, Bridging, and Bursting
- Cloud Migration
- Hybrid Cloud Integration
- **Green and Energy Management of Cloud Computing**
- Configuration and Capacity Management
- Cloud Workload Profiling and Deployment Control
- Cloud Metering, Monitoring, Auditing
- Service Management

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Of course the management operation and monitoring is another interesting area or important area. So, cloud composition service orchestration between services, cloud federation bridging busting cloud busting and those type of thing, cloud migration hybrid cloud integration like green and energy management of cloud computing, I kept it in bold because that is that stand out that this is not only a research field, it is also a major challenge and requirement to make this cloud computing a success, right.

Configuration and capacity management how to how do I configure reconfigure my infrastructure as a service provider and how to estimate; my capacity type of things cloud work load profiling and deployment control again another important aspects and if you see these are the things which are not exact these are not in isolation there are interconnected, right, then cloud metering monitoring and auditing of the services service management. So, there are different aspects of cloud management operation and monitoring which plays a important role.

(Refer Slide Time: 22:25).

## Cloud Security

- Data Privacy
- Access Control
- Identity Management
- Side Channel Attacks
- Security-as-a-Service

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

And as we already discussed that cloud security is an important need to be appropriately addressed. So, that people or organizations get confident to store their data in to this cloud there are not only technological issues, there are several legal issues, right if there is a data leakage; how this need to be handled in the law of things; if I need to be handed in the federal law then the physically the data should be stored may be has to be stored in the physical boundary within the physical boundary of a particular country or particular state, right.

So, there are several issues which need to be addressed. So, privacy data privacy access control issues identity management there are issues of side channel attacks which are becoming popular like the it is; it may not be the directly looking at the activities, but looking at the some other activities or basically understanding the basic operational things I can basically adjust some of the things; we have discussed during the security while we are discussing about the cloud security in our previous lecture that what we have seen that when we there is form a in a paper which we says that by looking by understanding the basic philosophy of allotment and type of things I am able to guess somewhat that say the IP address block and type of things in a particular of another client if I can; if somebody can do that then, there is a possibility of the security bridge, right.

So, these are different security aspects especially if it is mission critical things, then the organization or individual ones that security should be there. So, there is a whether; there

is a question that whether security as a service can evolve right whether we can provide security as a service in the in this context another.

(Refer Slide Time: 24:48)

## Performance, Scalability, Reliability

- Performance of cloud systems and Applications
- Cloud Availability and Reliability
- Micro-services based architecture



IIT KHARAGPUR



NPTEL ONLINE  
CERTIFICATION COURSES

Another interesting and of course, important aspect is that performance scalability and reliability of cloud computing, right.

So, performance and of cloud systems and applications is a another major issue like how to measure performance how to basically maintain that performance at particular level of performance and with proper resource management cloud availability and reliability is that like micro service based architecture. So, a service may that different micro service.

(Refer Slide Time: 25:46).

## Systems Software and Hardware

- Virtualization Technology
- Service Composition
- Cloud Provisioning Orchestration
- Hardware Architecture support for Cloud Computing

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES



So, whether my particular underlining architecture can support this type of micro services; so, I whether I can design a micro services based architecture for handling for better performance scalability and reliability; of course system software and hardware plays a important role like virtualization technologies whether the better virtualization technologies service compositions cloud provisioning orchestration and hardware architecture support for the cloud computing. So, instead of taking any hardware whether the hardware itself has is supported to this implement this cloud. So, this is also a important aspect; it need to for in order to work on that you need to know more about the internal working or internal more implementation details of architecture and development of architecture and type of things.

(Refer Slide Time: 26:36).

The slide has a yellow background with a blue header bar. The title 'Data Analytics in Cloud' is in red at the top left. Below it is a bulleted list of four items: 'Analytics Applications', 'Scientific Computing and Data Management', 'Big data management and analytics', and 'Storage, Data, and Analytics Clouds'. At the bottom, there are two logos: 'IIT KHARAGPUR' with its emblem and 'NPTEL ONLINE CERTIFICATION COURSES' with its logo.

Data analytics in cloud data analytics is the buzz; what these days we with huge volume of data and every aspects of the things, whether it is business banking metrological data or educational data where everywhere this analytics is there whether and as cloud; what we have seen its a it is a infrastructure which can hold huge volume of data may be and also application running on the data, right. So, yeah and not only that it tries to give ensure interoperability between the between the different data sources.

So, this data analytics in cloud is one of the; one of the very hot topic of research. So, there are different analytics application developing analytic as a scientific computing and data management So, it is not that data management and computing a separate, but whether I can my management goes in more hand to hand with the computing aspect big data management analytics storage data and analytics clouds. So, these are these are some of the things which are coming up in a big way and there is a lot of opportunity to work on this areas.

(Refer Slide Time: 27:53)

## Cloud Computing – Service Management

- Services Discovery and Recommendation
- Services Composition
- Services QoS Management
- Services Security and Privacy
- Semantic Services
- Service Oriented Software Engineering



11

And another aspect like cloud computing the service management in cloud is; they are like services discovery and recommendation services composition services quality of QoS management, right the QoS management of the services, right.

So, what we are trying to look at. So, security and privacy services security and privacy of these services. So, different services are there what about the security or privacies of these services there is a semantic services, right. So, that is whenever what like a different domain has different type of dynamics into the things, right different type of semantics into the; if I try to look at that whether predictions and type of things. So, there is underlying semantics into the things; how to incorporate that semantics and come up with semantic services so and service oriented software engineering aspects where how to put it on the cloud.

(Refer Slide Time: 29:00)

## Cloud and Other Technologies

- Fog Computing
- IoT Cloud
- Container Technology

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, these are the some of this other aspects which are which are becoming popular and finally, they are several technologies which are coming up and how cloud and those technologies come into play; we have already discussed these things like one is fog IoT cloud is another aspect sensor IoT clouds sensor cloud; things are there and this container technology as a we have looked into Docker technologies. So, how these technologies and cloud computing can go hand in hand to provide better services.

So, these are some of the aspects which we try to look into while as a future scope or what we say that recent trend in research. So, one major one good place to find the that what are the different what are the upcoming field or what are the present direction of the research is looking at top level channel general transaction and looking that what sort of special issues or sort of scope of things are there and also top level conferences say; what they are scope of the topics; they are looking for those type of things are there.

So, with these let me conclude this course hope you have enjoyed these particular course and there and you could find some interest in this particular aspect of cloud computing. So, what I believe or what I basically look forward to is that this will help you in further higher studies; in this aspects and also those who are interested in research in this field will find lot of opportunities are there.

So, there are some of the things which you can look into that there are several cloud simulators are available it is it may not be always possible to develop infrastructure or

getting infrastructure or higher cloud to experiment. So, there are simulators which we can; which you can work on to see that whether things are whether; whatever you experiment on the or whatever you are thinking can implement on the things. So, with these let me conclude these course and thank you very much to participate to for participating in this course.

Thank you.

**Cloud Computing**  
**Prof. Soumya Kanti Ghosh**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**

**Lecture - 41**

**Cloudâ€“Fog Computing - Overview**

Hello, so welcome to the course on Cloud Computing. Today we will be discussing on cloud fog computing paradigm, why it is important. So, already you have gone through the basics of and some of the aspects of cloud and fog computing. We will be in last in coming couple of lectures, we will be discussing little more in details that how this overall performance issue latency etcetera make sense while we have this type of paradigm, why this is important and why we have need this sort of a thing. So, more to say it is a cloud fog age sensors and other systems will come into play. Let us see that during the lectures that we will look into the things.

So, primarily we will give a overview and also we will look into a case study, primarily look into a health cloud fog framework. Some of the things that we will share that some of the work we did in our lab out here for realization of this cloud fog paradigm for a health domain. So, it may be in a lab setup, but I believe that that will give you a good understanding that how why this sort of a framework matters. And also we will look into some of the performance related issues, this particular paradigm have lot of promises to give.

So, these are some of the key words which will be there. So, to start with let us have a some quick recap. So, when we talk about cloud computing, it comes as a something as a service rather if we say anything as a service which comes into play. So, three prominent predominant services or three predominant model what we look into is one that software as a service arise, these are the different service model and or platform as a service and infrastructure as a service. So, if we look into the overall things, so infrastructure as a service gives me that basic infrastructure to do this computing thing.

When you have a platform as a service, so we have something or rather as a software as a service I do not care about the infrastructure and things like that, I directly want to have that realization of the software into the thing working of my tools and software's over the over the this computing platform. And in between we have this platform as a service which help me which provides me a platform may be for development, for execution some of the things to work on the things. Other than thing we have lot of other things like we are used to like sort of a storage as a service or something like database as

a service. So, what we are saying that somewhere ubiquitously the services are available and I can call that service and pay per use right as or pay as you go type of model, we pay as far as we use. And we know that some of the major characteristics are one is on demand self service that means, with minimal management interventions things should be there, resource pooling, broad network access, rapid elasticity and measured services right.

Though it is what we say in a it came from a distributed computing gradually distributed computing platform, but if you look at in some aspects that it is some sort of a some sort of a what we say as far as the cloud is concerned when we when we work with it. Say suppose I hire a cloud from say either Amazon or Google or my maybe we have a in IIT Kharagpur we have developed a cloud or a cloud infrastructure with our self with using open source that is Megamala whatever, what we whatever we look at as when I am working is this some sort of a centralized stuff. Not only that it is somewhere there. So, I have I need to send my data out there and get the result out of it if it is a something which is for computing right. However, in some cases this network delay may be much more right with all these different type of advantages what we are looking at instead of having owned infrastructure I can have a ubiquitous infrastructure which help me in increase mobility least botherization of my maintaining the infrastructure I just hire infrastructure like computing as a service.

So, with additionally there are issues of where we have this latencies and other things into play. In some scenarios specially mission critical scenarios and like that what we feel what what is seen that this may be sometimes more challenging of having this sort of a data traversal over network and delay etcetera. So, another thing which came up in after this cloud computing and what we studied is the fog. So, as the name suggests cloud is up in the sky and then fog is something little down or more near to the surface or more near to us right. So, it is a model in which data processing and applications are concentrated in devices at the network edge rather than the existing almost entirely on the cloud.

So, it is it is something where in the cloud. So, I we want to have some realization of this say in the processing term or analysis terms or the data somewhere in between right. So, in other sense what we see that instead of travelling direct to the cloud I if I can have in between this type of services it may make sense for better performance right. So, the term fog was originally coined by Cisco, Cisco systems as a new model to ease wireless data transfer to distributed devices in the in IoT paradigm right. So, as we have huge number of sensors and sensors actuators which are collecting data and being the data being analyzed and may be actuated for some other purposes.

So, what is seen that there is a tremendous data flow is in the network right with the increasing number of sensing devices. And so, Cisco with a huge infrastructural this network infrastructure what is what is seen that whether it can be used that unused or used resources in the network can be used in the things specially look out the nodes like routers etcetera which may be having you know good amount of computing resources, but for this transfer may not be there. So, I can I can do it in a at a much lower level. So, what happens? So, if we if we look at a scenario like as we have discussed say I am sitting in this room we have say 3, 4 such studios in this particular floor. So, every if every studio may be maintaining a temperature with some 10 to 20 sensors.

So, overall we may if there are 4 studios and there are 40 sensors which are sending data to the cloud, cloud analyzing those taking a call whether the temperature need to be increased etcetera and type of things right. So, while sending the data it is aggregating into a intermediate node or in an intermediate switch or a router where it being transmitted at the next things or what we say it is being aggregated at the gateway. So, what if see the this particular application if the temperature control is the application it could have been good that if a preliminary things suppose everything is no need of transferring this data of the things right. So, if I say then I need to maintain the temperature between 20 to 22 degree centigrade and it is between and plus minus 2 degree. So, I if instead 18 to 24 I mean if it is within if my local node of this 10 sensors of this of this room is thing that it is the within the range it will not transmit it will say that things are or it may transmit that it is within range type of things right.

So, what way this will help? It will help that I reduce this load into a much lesser thing. So, what we are trying to do? We are trying to do the computing at a much lower level because some resources are available right. It can be network things later on we can say it can be my mobile device, it can be some other things which are say excess resources are available whether I can use for computing at that purpose. So, anyway we need to transmit to the cloud because the cloud may be doing a global analytics where it sees all the 4 rooms say for this typical synthetic case right. Otherwise if I am doing on a local things I only see the data of this particular room.

So, as Cisco is having a huge infrastructure of the of the network. So, that may be one of their motivation of using that for computing purpose and it is much lower than the actual cloud which is on the little far away. So, it is what we termed as a Fock computing. So, vision of Fock computing is to enable applications on say millions or billions of connected device to run directly on the network edge. So, instead of pushing to the other end of the network whether I can execute at the edge of the network itself in thus reducing the overall traffic load of the rest of the network which saves network delays or of course, energy and quote unquote overall costing on the things right.

So, just this is a thing what we have we have already seen I am not going detail into the thing. So, there are some so called differences between this cloud and Fock computing paradigm like latency typically cloud in case of a high where Fock is lower than that or delay jitter say things like say geographically it is something some sort of a centralized cloud where it is particularly distributed and so and so forth. So, there are different pros cons both the things, but what we and if you see this typically this comparison table it may appear that why cloud why not only Fock type of things, but we do understand that the Fock is not a replacement of the cloud right. So, it is more of a something what we say a complementary or a supportive technology to the things which helps in better utilization of the infrastructure. So, cloud need to take a maybe a global call of what is going on overall means overall infrastructure wise and cloud can provision different type of service things which Fock may not be able to do that.

Fock has low resource mostly in comparison to the cloud and much nearer to the things and it does not have a global vicinity whereas, it may provide better security type of things if is if you are putting those type of resources on the things. It supports some sort of a mobility and there are the latency is low and so and so forth. So, it makes sense say if I have some sort of a infrastructure which goes hand in hand right what we say a paradigm where we have cloud Fock and rather we keep this age also where actually my sensors etcetera are capturing the data transmitting the Fock to the things. There are little what we say little thin line between that where this age ends and Fock some of the references you will find that interchangeably Fock and age thing age computing are being interchangeably used. Nevertheless what we want to see it like that.

So, I have a sensing devices right which can be IOTs and different category of things who are sensing data right and that it transmitting to the this Fock nodes right. So, in where we refer here the age is those type of sensors right. Nevertheless those sensing devices may still may have some sort of a intelligent or may have some resources by doing some processing there like typically reducing some trivial noises. So, that the more clean data being transmitted and so and so forth. So, what is the thing that bringing intelligence down the cloud close to the ground.

So, we have this cloud up in the so called up in the quote unquote sky and then bringing this intelligence down the down the towards the surface right. So, that some of the reason can be taken as a at a much lower level. The changes are there that as these are independent things. So, they do not have a global view. So, as such the cloud can still take a call of that the data or what we say this process data which is being transmitted.

But in doing this intermediate layer what we achieve that reduce the overall traffic

which are being pumped to the cloud right. We will later on see there is another concept or another technology or approach what we say due computing will come later on right. So, this like cellular base station network routers, Wi-Fi gateways will be capable of running applications right. So, if there are resources available. So, it can be cellular base station or our say Wi-Fi routers which are installed everywhere.

Even this network routers which are highly resourceful and maybe they have excess resources which can be utilized for application purposes right. I can have dedicated type of things which can do also as a intermediate node or the fog nodes right. So, end devices like sensors are able to perform basic data processing right. So, sensors will data processing in the sense data capturing and maybe some sort of a basic filtering like some sort of basic operation and transmit or transmit the raw data to the things right. So, process close to the device lowers the response time enable real time application.

So, once the response time is lowered which helps us in achieving real time applications right or realization of real time processing of the things. So, what we try to do in doing so that it is multilayer. There are definitely lot of challenges come into play right that how overall management of this data flow will be there right. How much we gain in doing so, alright and then if I have like if you are transmitting everything to the cloud. So, you are you know that how to handle things right you have that particular way to handle.

Now if you have intermediate lot of devices in between which are acting as your intermediate layer for computing like for computing. Then we have challenges like that how we can do that basically inter operate right. The data being processed by different say if my fog nodes are not homogeneous, if my fog nodes are like this that some fog nodes are say network routers or Wi-Fi routers and some fog nodes are maybe something say mobile device or something else. Then how do you handle this type of interoperability and type of things. There are issues of how what is the timing relationship will be there right some fog node sending some data at some rate some nodes etcetera.

So, these are challenges there are several challenges or in other sense we need to look into that overall what is the overall performance whether we gain or not for all cases whether this will be good or in which cases we need to deploy this type of scenarios right. So, there are things which need to be looked into. So, if we look as a cloud fog we keep the edge type of paradigm. So, what we see the same if you see the same thing in a different form. So, at the ground level we have different sensory devices right primarily they are sensing it if we look at the medical then it may be sensing different type of things like typical what we say body area networks right.

So, your things like say body pulse rate, heart rate, body temperature, pressure and it may have lot of other things which can be sensed right. So, those data are being transmitted to intermediate fog node. So, that may take some call like initial pre processing of the data that can be one way it is looking at the set and there can be other type of things also like if it finds that the basic analysis that a immediate alert is needed to be sent to the patient like you need to see a doctor or something like that that can also it can generate right. So, it is not a one way communication it is a both way communication between fog and the edge what we see out here also right. So, there are both way communications are there.

So, the fog nodes in turn accept this data do necessary processing at its end whatever is meant for and then this transmit this data to the cloud which is which may be a processed data right. So, if it is a health related data so, it may be sending a overall health metric along with the may be the actual data or it may be sampling the data in some rate and type of things right like if it is a metrological data then it is doing something else. So, based on the logic there whatever is there embedded in the fog it works on that those logic and work on it and that is being transmitted there. So, what we gain here some of the responses are being done out here itself right at the fog level. So, this overall this travel time from going from these sensors to this cloud and coming back and giving the result that is reduced.

So, I have a better response time right and it may helps in several real time applications right and the fog which is doing a partial computation or may be transmitting the not the whole data rather a subset of the data or a processed data to the cloud. So, in other sense it also reduces from this sector also the volume of data into the to the cloud. So, this also helps in overall achieving better response times, low energy cost and bringing down the overall costing of the thing right. And that it along with that this fog may be getting from some other data some other sources etcetera like if it is a data related to say agricultural related things the fog data may be getting some of the metrological data nearby. So, it may be getting other resources right or that may be data it may be in the cloud.

So, there are definitely advantage of having this layers in between which helps in what we say that partial computation at a much at the network age. The definitely the cost of all those things is that you need to have this fog devices right or your devices which are which can work as a fog node while during this overall transmission thing. So, if we look at that cloud issues or the limitations this one is the major thing is the latency. So, latency is pretty high at times. So, as we discussed in our previous lectures that it is not a it is not meant for high performance computing per say right.

So, high HPC is HPC. So, cloud may not be mistaken as a HPC type of things. It may provide different things like your high performance computing paradigm, but it is per say is not a HPC platform as such not looked as a HPC platform. So, large volume of data being generated right in as if we here in this case some of the data sets are restricted here otherwise huge volume of data is generated. There are which has a larger bandwidth requirement not designed for volume variety and velocity type of things like three V's of a big data type of concept right and generating which are being generated by the several IoT devices right which is which again has different interoperability issues etcetera. So, as we are just some few minutes back discussing that there may be interoperability challenges when we have different fog devices there may those challenges may be there when we are transmitting from the IoT rather I can have a fog device which fog devices which have agreed upon protocol of what need they need to transfer to the cloud.

In other sense it helps in it may help in achieving interoperability when the data set are coming from different fog sources right. So, if you look at the IoT or that is which are here in this case H devices. So, there one of the major challenge is a processing you do not have that much processing power at the edge. There are challenges of storage like usually these devices are low on storage like if you are using say what we say we are used to this pulse oximeter and etcetera they are not meant to store things right. So, they are low on storage right and also they have a power requirement in several cases they require a power requirement specially when you have something which is on the on mobility then you have a power requirement.

And fog much lesser latency permits uses in real time application less network congestion reduce cost execution at the cloud as we discussed more data location awareness like that is one part that if one fog and if you have distributed we have a one cloud we have a distributed the sensing of room temperature across IIT campus IIT Kharagpur campus. So, the fog device where it is located it knows that where the coordinate. So, this inherently I can have a spatio temporal data space time along with the other attribute data right. So, it is a better aware of the things otherwise you have to transmit always the GPS thing that is also ok, but that increases the load of the data right and increases the turnaround time better handling of colloquial data generated by the sender that is also a thing.

So, we will just look into a case study quickly. So, which help me help us in understanding that why studying these two things or overall paradigm make sense. So, here what we see that three sort of a layers one is the cloud at the top what we say level 0 that is no as such hard and fast that how many level etcetera we are with this is for the work this particular case study and discussion then we have a ISP in between and then

we have the area gateway. So, these are level 0, 1, 2 and then we have maybe some mobile devices which collects the data from different sensors. So, this is at the level 3 right and then we have the level 4 type of things. So, this fellow or these fellows may act as a fog network right which take a call.

So, in other sense the fog may not be a single device it may have different group of things or in other sense what you say I may have I may have a fog networks which helps me in achieving this. And the information flow both side one usually the up in the north direction the data is more going to the cloud and analyzed data is data low it is less, but nevertheless that information flow on the both side. Now this is what we did in a very synthetic environment in our lab and this there is a typical some of the configuration and this some of the configuration we basically used a simulation platform primarily that cloud SIEM and iFoXSim. Cloud SIEM is a from the University of Melbourne iFoXSim is a combinedly developed by a team of Kharagpur, IIT Kharagpur and University of Melbourne and now they are taking care of that tool. So, this is if we have typical infrastructure I think and if we have this different type of devices here and which have some latency.

So, these are some of the things what we try to see which with some reference material then we try to see that how overall performance things matter right. So, and if we look at this particular application that components and flow. So, this something EEG signal this client module it captures send to the client module data filtering module data processing module and even handling module different and even handling. So, there can be a confirmative module that whether this events has occurred or this alert need to be generated and it goes to the client module and then it is that goes to the display or what we say informing the user or the patient to the things right. If this is my overall flow so, what we have we have different modules to execute.

Now we have a module to execute and we have couple of layers where we can say this I can execute in cloud, this I can execute in fog layer and type of things try to see things. So, we did some experimentation again I with a caution that this is a very synthetic type of things it should not be generalize that this results will come everywhere, but we try to see that though things varies when we put layers in between. So, this is again that the placement obtained by different application module for fog and cloud architecture like client module in the mobile in case of fog based placement in the fog based module in the cloud based module in the mobile whereas, data filter module in the area gateway data processing module in the area gateway and error handling even handling area gateway and conformity module goes to the cloud. Because cloud has the global view that whether it confirm that it is a event has occurred or not like something a medical event and maybe if you work on a traffic type of thing something even the traffic event.

And this is in the simulation while we as I told that we use this iFoXim simulator.

So, these are the different simulation platform we used and this is the same thing as we see that these are cloud, ISP, area gateway, mobile devices and finally, we have these sensors. Now, if you look there at the performance is Neterwack uses. So, is very low in case of a fog architecture as only few like positive cases in the sense which cases need to be transmitted to the cloud is there right or candidate cases are there to cases the conformity module residing the is access right. So, the typically the fog load is low in case of cloud based architecture the uses is high as all module are on the cloud. So, it has go whether it is a whether it is say data filtering it will go to the cloud whether it is even handling go to the cloud and definitely conformity is going, but in this case it goes to the cloud.

Again I am just to keep you again just again I have to mention that we are we need to this is a very lab type of experimentation with some small number of nodes and etcetera. So, in different scenario things may be different, but nevertheless the overall story may be same. Now, cost of execution in the cloud because what we see that when you when you use your own infrastructure that is infrastructure why you are using because that is surplus right. So, that is practically no extra cost is involved into the things, but when we go to the cloud type of thing then we have a problem of this costing of the whole thing right. So, execution cost is much higher when we in this case in the cloud and even similarly if we look at the latency also.

So, in case of a fog it is more or less stable whereas, after some configuration we have used different configuration as I shown you. So, configuration we see that this latency increases in case of a cloud. Similarly in energy consumption also this different type of patterns come into play when we look at the different category of energy from the DC that means, data center that means, that actually in the cloud and then mobile energy and age and it varies in different aspects. So, also we tried with a prototype like try to see in the lab that how things works with using some of the medical devices like as you can see some hand band etcetera where the data being transmitted and we have used as very pi as fog devices which and AWS or in house open stack for the cloud, but the in house thing is the network delay is much less because it is within our IIT Kharagpur network. So, the delay you do not perceive, but if you have external kind like AWS and other things like Google cloud and type of things you can have things like incidentally Amazon has given IIT Kharagpur some free credits to the students.

So, we utilize those. So, use different data sets and customize to the formula for analysis like what we want to do in doing so is that we need to send a alert that when the in this case when the patient should go to a doctor or call for support etcetera. Now, when we

say that alert then we have to have some medical domain analysis into the things which should run some of the things in the fog devices itself provided that this fog device has the resource to maintain and those application and some of the things should run at the cloud end right. And we have tried different configuration with respect to resource allocation customizing the different physical devices and so and so forth. So, this what we in this thing what we tried to looked into in this our first lecture on this couple of series of not series we will have 3-4 lectures on this. We want to say that whether having this fog device or cloud fog paradigm or cloud fog age paradigm what we try to achieve is some sort of a better efficiency or better performance in time term in terms of delivering any applications right.

In this case the experiment in health application and we can see some of the performance metrics are giving better results with when we have this sort of a combined paradigm. There are few references and with this let us end our discussion today we will continue with this discussion in our next lecture. Thank you.

**Cloud Computing**  
**Prof. Soumya Kanti Ghosh**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**

**Lecture - 42**

**Resource Management - I**

Hello, so let us continue our discussion on this cloud computing course rather we were discussing about little bit more relevant topics or which are little bit above the general overview. So, what I believe that you have already seen the other lectures which gives a overview of the overall cloud computing thing. And today we will be mostly looking at that the challenges in resource management also the problem related to service placement in a cloud fog edge type of thing. This lectures as I mentioned in my previous talk that we will be concentrating more on cloud fog paradigm rather cloud fog edge paradigm and why it is important to realize a better resource management, better performance in terms of say timing or latency and type of things. So, mostly we will be looking at concepts like resource management issues and also we will look into the service placement problems rather this will be our base for going to other type of things like where we will look in subsequent lectures into things like migration issues and other type of things. So, these are keywords.

So, just to quick recap what we have seen that only cloud only type of scenario that means, you have this underlining IoT or sensor devices and above at the other end the cloud. So, like you have some of the things like health parameter like the type of things we discussed last lecture that you have body sensor networks or different sensors which has this different type of health parameters and that it uploads in a cloud it computes the whole thing and return the result to the either to the device of the mobile device or the display unit or any other display unit of the patient or the client right. So, what we have seen that processing IoT application directly to the cloud may not be always efficient solution because specially for time sensitive things right where it requires more time and not only those things it have a lot of resources right like lot of data being communicated over the over the this network. So, a promising alternative what we have seen that making a fogged computing thing paradigm.

So, these paradigm impose to process large amounts of generated data close to the data sources rather than the cloud. So, the what we have seen that it is by the definition it is bringing down. So, one of the consideration of this cloud based fog environment is the resource management definitely right like. So, it is not always a win-win situation like you go on increasing your IoT devices and then you expect that your age intelligence or the fog infrastructure will take care that may not be true. Not only that in number of cases what we will see what you what is been seen that some of the resource may be over constraint and

some may have still something spare.

So, appropriate managing this resources is also necessity. There is another point to be noted whenever we want to do any type of management thing which was may not be there in the cloud only scenario it also brings about some costing right. So, the a fog resource managers per say should work and that fellow will also it have is not only some resources it also it takes time to do the things right. So, nevertheless some sort of a resource management at this down the layers fog age type of layers is necessary that that is there right. And it is a sort of a resource allocation if so, this amount of resources are there how it will be allocated work load balancing right.

So, it should be somewhat equidistributed as per as possible right resource provisioning task scheduling QA's maintaining appropriate QA's level and type of things are coming into play. So, in other say sense we see that is a whole lot of things come into play. Nevertheless irrespective with irrespective means in spite of handling this type of scenario this type of paradigm where cloud with fog and age may be beneficial that we want to see. So, our today's talk is mostly over around how this resource management things will be what are the different parameters or what are the different means characteristics of this resource management problem. So, what we just to continue from the previous slide.

So, fog age to support for the cloud computing paradigm. So, the major issue is the latency issue right like may involve transfer of data for each single sensor to the cloud. So, whereas, this if we have this paradigm it will reduce that communication likely to reduce that means, as we have seen that I can aggregate the data in cases if it is possible and then transmit. Some of the things I can response at the fog or a age layer itself like some of the call it may not have to go to that level to things like as example we were discussed earlier like if I want to look at the room temperature and want to make a alarm call that the temperature is above the threshold level that may not has to take the go up to the cloud level and take a call that my even my sensor if we have a more computing things that means, at the age itself I can take if the temperature is more than something it sets up the alarm and say that it is there. Sometimes, it is aggregation is required like I all the sensors in this room should agreed upon things like it may happen that one of the sensor means may be misbehaving and set up a alarm and whereas, if there is a aggregated things.

So, that could have been done in a little much higher level not may be the age those fellows are sensing may be at the fog level again I reduce the load on the cloud right or sometimes we want to make a aggregation of this data and then send on the cloud. So, the overall traffic from this fog to the cloud reduces right. So, this is these are definitely latency issue and so, fog age computing made cloud in overcoming this. So, it is not a substitute rather it is a supportive or a cooperative or collaborative system came into play and there are lot of challenges also come into play right. Once you collaborate there are issues of interoperability, there are timing relationship and between this different processes and things like this comes into play right.

Rather what we will see in this subsequent lecture this event this fog layer may help you in attained in making this different type of devices interoperate right fog age layer right. Like I for example, if I am taking say even say come to our example of temperature sensor if we if even if we look at the temperature sensor that may be a variety of there may be more than one type of sensor that means, coming from different make and model. So, what will happen it may end up in sending simple or as basic as that they may be sending out this temperature things in a different format. For that matter suppose one is sending in centigrade and one is sending in Fahrenheit like. Now, this making in a uniform conversion could have been done in the either in that age itself or at the fog layer.

So, in other sense it allows me to interoperate in the things and if my if there are variety of sensors. So, it you may can imagine the complexity of the whole thing right. So, that this three technology can work together to grant what is improved latency, reliability in some cases like and faster responses right. So, though these things are interrelated things are apart from that there are that will help in at times in interoperability like to interoperate between the different things and later on we can we will see that it also may help in some sort of a what we say information security right. Now, like some of the things like fog etcetera may be in my control or in my premises.

So, while sending the data I can have I aggregate the data. So, individual data is not thing is not pin pointed or even I can skip of this say patient information the patient like patient identity and send that data up in the thing and while coming back I attach it I number it with the things right. So, starting from there are other crypto etcetera we can run we are not going to those complexity never this what we see that it will definitely help in having latency, reliability and faster response and may at times this trustworthiness and type of things. So, again we look at the same picture or same type of things. So, we have several age devices or IOT sensors and actuators like sensing and actuating based on the things and intermediate we can have a fog layer right.

So, there are different fog layers and if you can see this fog layers can have different type of resources like in this cases they are having different virtual machines right which takes care of the thing. So, it requires a hardware, a hypervisor and the virtual machines. Once you have you need to have a fog manager, server manager to manage this if there are more such things you need to have a some sort of a fog sort of a fog data center or a fog network which also may require a some sort of a management of the things right. There are other things what we will see that there are challenges of mobility right like as we our age devices or this IOT things can be mobile. So, this fog thing should take care of the things.

In other sense if it is under one paradigm of the one fog device right we are considering the fog devices to be stationary then once it cross and go to the another fog device area. So, that there should be a sinking of the thing that same sensor is sending data or if there is a data aggregation going on I may not do I may not want to lose this data type of things right. So,

those things are there is a underlining communication network definitely which push it to the cloud. So, there is a request goes on rather there should be a other way around as it is actuating some of the some it is request and this other arrow is the request response type of things. So, it sends some request and get a response and actuate on the different type of autonomous mobile sensors and type of things right.

So, what we see that fog environment model typically is composed of three layers a cloud client layer age, a fog layer and a cloud layer right. Fog layer is to accomplish requirement of the resource resources of the client and if there is no limited no or limited availability of the resources in the fog layer then the request to be passed to the cloud layer right. So, there may be two scenario where this age is pushing to the fog is pushing to the cloud one is that it do not have the capability of computing that type of things right that particular analysis or that computing model or it is getting constrained on the resources. Once it is getting constrained on the resources especially in case of a fog it can delegate to the other what is collaborative fog devices right or it can push it to the cloud. Another thing is that it do not have the way to compute this right like if I want to say if we looking at the say all temperature sensors in all over this building all rooms and then want to take a call that or in this whole our institute and then want to take a call that which are the things which are maintaining and which are not maintaining that individual fog devices and aggregatedly how many percentage of the this air conditioning systems are working faithfully or in a proper way.

So, that can be so, this global thing can be done at a cloud level right. So, the major functionality so, what we see if we have this type of fog age layer. So, major functionality is that the there should be a fog server manager which employs all available processors to the client right and there are definitely VMs operate on the fog data server process them then delivers to the result to the fog manager and fog servers content fog server manager virtual machine to manage this using a server virtualization technique etcetera. So, in other sense the type of virtualization techniques and computing things that comes down below that the fog layer from the cloud infrastructure. So, that may be a overall managing this fog layer right.

So, this type of things may make lot of sense in application specific domains right it will change over to a like if I if we look at say something like vehicular ad hoc network type of things right where we have underlining this cards and vehicles which are having their own sensing units and not only that they are having that on board units or OBU's and it communicates with either another car right. So, if I look at these are the ages right with sensors at the down. So, it can either communicate with the things what we say V to V communication vehicle to vehicle or it communicates with the road side infrastructure. So, V to I infrastructure right this road side infrastructure is connected to a back end cloud data centers. So, for all type of things suppose a revocation of one license etcetera or something is there certificate is there then it is basically taken a call at a much higher level and done like that something is misbehaving or something is happened or means something is some

group of car is misbehaving like slowing down not expected.

So, it may have to be taken a call that whether there is a in whether it is a there is a call in there is a accident in front and type of things which may have to be correlated with other information which is known to the cloud right. So, that type of things are there. So, what is happening now this infrastructure may maintain different resources where different virtual machines etcetera are there and it has a separate things which acts as a fog layer for all those things and which push it to the cloud. Cloud also can definitely have other means resource management things right.

So, it can be different. So, this is one scenario if we look at the scenario like healthcare systems then we have say if we have a body area network. So, it captures the things transmit to a nearest hop or a it may be a local server or even a mobile device which acts as a fog and in turn it push it to the back end health cloud and type of things right it can be that way also. So, in case of a say vehicular network it can be too much mobile right it is make and break are there where in case of a other scenarios like either health or like temperature measurement or temperature humidity and etcetera measurement across the rooms those are more or less static sensors they are not moving around now and then. So, the trend as we can see is to centralize some of the computing resources available in large computing server things that is why we are having this fog as we have seen dedicated micro data centers and type internet nodes such as routers gateways switches are augmented for the computing facilities and as another set of definition I kept it a little in italics that because you may be by a time bombard with so many definition, but nevertheless the philosophy is still remains same bringing down the computing from the cloud to down the to the as much as to the client end or the where the sensor end so that the computing is more facilitated right. So, if we look the overall resource management paradigm.

So, this is the thing which we referring that reference is given below a nice very nice survey paper which takes care of this deals with lot of things will be primarily looking at that stuff. So, if we look at that resource management in fog and edge computing rather fog cloud fog and edge computing. So, we like to divide it into three major components right one is architecture wise how they are infrastructure wise and algorithm wise which is dealt in this particular reference paper you can look at the whole document very nicely written survey paper and pretty recent one end of 2019. So, and if we look at everything like architecture. So, architectures also have different components like segment like data flow where as we are discussing that aggregation of the data or sharing and offloading of the data right.

So, I like aggregating the data if it is make sense sharing information for other computing purpose or things and offloading to other devices or offloading to cloud and type of things. Control another aspect is the control right which looks as the whether the control is a centralized control even if we look at the fog layer itself whether it is a centralized control or distributed or it can be hierarchical control right I can have different layers of

hierarchical control. So, this now we are having this fog edge things in a much bigger perspective. When we look at the infrastructure so, infrastructure whether we are looking at hardware infrastructure or system software or middleware right. So, different aspects are being studied or lot of research going on that how overall things works right.

So, on the other side given that we have a infrastructure and architecture in place what are the different type of algorithms which will come into play like things like discovering the resources or benchmarking the resources or how things will work load balancing issues or load balances algorithms and there are challenges of placement right given a given a job where do I place right the job. Even if you can see that in even those who have gone through distributed network, distributed computing type of paradigm they are also what we have seen that this placement of the things is are challenges rather all these components are somewhat present in those type of scenario. So, if we look at the resource management approaches one is the architecture as we are discussing use for resource management in cloud fog edge computing based on data flow or classified data flow control and tenancy right, but where it will be there where it will be where it will be residing right so, tenancy. The again infrastructure composed of as we have seen hardware and software to manage the thing network storage resources for application and this like overall resources how it can be managed right and the other side we have the algorithms. So, underlining algorithms to facilitate cloud fog edge computing paradigm.

So, these are the things which are which are very much needed for this type of scenarios right. So, again coming back to the figure if we want to have resource management so, these are the major component and where it need to be we need to looked into. Now, we will let us see one component other. So, architecture if we look at the architecture as we have seen it is data flow right control and tenancy. So, it is based on the direction of the movement of workloads right when we talk about the data flow in the computing overall computing ecosystems right.

So, workload can be transferred from the user to the devices to the edge nodes or alternative the cloud server to the edge nodes and type of things. So, how this data is flowing based on your application requirement also dictates how the resource need to be managed right. Like if I say my data is from the user end to the edge and fog and they are it likely to be computed and come back then I have to make the resource management that way. If I want that my the if we the requirement is the data has to travel to the cloud and something has to be done at the cloud and come back then I have to take care of that so called intermediate this computing network resource management also into play vis-a-vis how this data being transferred etcetera right.

The other part is the control. So, based on how the resources are controlled in the computing overall computing for a single controller or a centralized algorithm may be used for managing a number of edge nodes or it may be a distributed things. It all depends that how much your overall paradigm is right. And the thirdly what we see when you look at the

architecture issues based on the support provided for the hosting multiple entities in the ecosystem either a single application or multiple activation could be hosted on the edge. So that means, if we look at this architecture so one is that one looks into the one aspects looks into the overall data flow mechanisms other is the control mechanisms like how overall resources can be controlled whether it is in a centralized way or in a distributed way.

Suppose I have a one or a set of fog devices more or less within the things then I can do in a centralized way or if I have say a fog network itself then I may have a distributed control into the things right.

There may be issues that based on the applications like say vehicular network type of scenario where I may want a distributed computing because there is a that the vehicle is moving over a large geographical space and things like that and need to be looked into. Whereas, in a other type of scenario where temperature control etcetera which are more or less static things we I may look for a distribution. So, it also depends on that and then tenancy means which algorithms which will be running and where they will be residing and where it will be running is a important thing right. A single or multiple application hosted on edge nodes or which things will reside on the thing. So, that is also important that where what will be the tenancy.

If we look at the infrastructure as we have seen the other aspect of the infrastructure. So, there is a things called hardware system software and middleware right. So, when we look at the hardware is the things like if we look at fog edge type of paradigm. So, it exploits the small form factor devices such as network gateway, Wi-Fi router, set top boxes, small home servers, edge ISP servers and so on and so forth. And if we say vehicular things like car and vehicles even drone as a computing servers for resource efficiency things like that right.

So, that means, if I having this other resources in place like if we look at Wi-Fi router or in a vehicular communication car or in a drone capturing images or other information so drone. So, if we can add on computational module or computational resources it has a additional computational or unused computational resources I can utilize we can utilize that for this fog edge computing right. Whereas, typically cloud have that sufficient resources to work on right. And the other aspect is the system software runs directly on the fog edge hardware resources such as CPU, memory and other devices right. So, once the hardware is there based on that my system software runs right.

So, it has there should be a underlining system software which manage the things. It manages resources distribute them with the fog edge like operating system virtualization software and type of things right. So, those are the things these are system software which are not per say actually the application specific software right. So, these are more running the running the devices and type of things. And then we have middleware which runs on the operating system and provides complementary services that are not supported by the system software right.

Like middleware coordinates the distributed computing nodes and performs deployment of VAM and containers to each etcetera. So, these are the middleware which are not typically supported the jobs which are not typically type supported by the system software, but it supports this type of things as we are seeing that deployment of VAM or container in the thing. So, those are those are the things which are middleware right which cut across maybe more than one resources right. So, if you look at the infrastructure so, these are typical component of infrastructure. And finally, in this paradigm if you look at that algorithm so, discovery benchmarking load balancing and placement.

So, these are the four major components when you look at. So, what we say infrastructure is in place the architecture based on the architecture and now we have the algorithms to run over and above the things right. So, discovery is identifying the easy source so, workloads from the cloud or from the user devices can be deployed on them right. So, this algorithm I should have underlying algorithm that which resource is free and which are the edge devices where it can run. So, benchmarking is capturing the performance like it may be CPU, it may be storage device, networking and type of things of a computing this.

So, it is very much needed because this actually dictates that how overall this overall performance over the of the system will be there. And load balancing as edge data centers are deployed across the network edge issue of distributing task using say efficient load balancing algorithm has gained gaining significant important like. So, it may deploy different optimization technique, load balancing techniques and different type of searching mechanisms and type of things. So, it is important to have the load balancing otherwise sometimes it may be skewed into the things right it if it is going to the setting the same type of devices.

So, this the algorithm need to take care correct. And finally, the challenges of address placement, the placement of services and applications right. Address is the issue of placing incoming computing task on suitable resources. So, how I can place this resources considering available resources in fog edge layer and environment if there is a environmental change right. So, it can be dynamic layer condition dynamic condition aware techniques or iterative techniques.

So, I can we can have different type of things. So, what we see that this four components is the algorithmic things. So, now, again if we come back to this big picture right where this resource management with respect to fog edge rather cloud fog edge type of paradigm is there. So, one aspect is looking at that what sort of architecture we are having which has data flow control and tenancy type of issues to be handled. Other part is that the infrastructure hardware system, software, middleware how things are there and the and then given this scenario we need to have this appropriate algorithms to run on those things so that the applications are there. So, our basic bottom line is that this our different sensors which are being there I want to take a call based on this data right whether it can be done at the edge intelligence or at the fog layer or at the cloud based on the things.

If I want to use this more efficiently this fog edge type of paradigm along with the cloud I need to have a proper resource management right. We will continue our discussion on this topic or more precisely some of the areas in my in our subsequent lectures to see that how overall this cloud fog edge paradigm helps us in achieving better performance in the overall what we say this ecosystem. Thank you.

**Cloud Computing**  
**Prof. Soumya Kanti Ghosh**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**

## Lecture 43

### Resource Management - II

Hello, so we will let us continue our discussion on Resource Management in the Cloud Fog Age Paradigm rather we will be continuing our discussion what we are talking about in our previous lecture in this lecture also. So, our basic consideration here is that what are the different parameters which influences this overall mechanism of this cloud fog age paradigm or more specifically when we are now looking at the resource management more we are looking at that on the point of fog and age because these are more resource constraint unlike the cloud. So, if we recap quickly our basic overall consideration in this situation is that in having this multilayered thing like cloud at the top then fog and age we want to have a better efficiency of the overall system right. Let it be healthcare system, let it be traffic management, let it be vehicular networks or even disastrous management any type of things where we require a different type of data acquisition, gathering, inference drawing, delivery to the end user or actuating some other devices how this device how this combine things will work faithfully right. Other thing definitely motivated by resource reach this what we say end devices sometimes smart devices or IOTs or intermediate devices where the surplus resources like CPU like computing or memory etcetera available. So, that we can leverage on this.

So, it is continuous. So, the same type of extension of those things we will be little looking at the service and offloading hardware software consideration type of say algorithmic consideration and so and so forth right and also this keywords also remain same. So, if we look at the so a few recap of the previous slide so that we are in sync right. So, if you look at the resource placement problem.

So, if you if we see that it is from this end that edge devices or this different sensors and IOT devices and the other end is the cloud intermediate we have fork devices and we require some of the things here in this case what is referred as the fork orchestrations engine or orchestra which try to see that how efficiently it can be there. So, it can be the load should not be very much skewed or equidistributable with the things and overall efficiency in terms of say in terms of in time or in terms of energy utilization and things like that is optimized right. So, a variety of devices at this end and cloud usually we have data centers or if there can be clusters or different data centers out there and intermediate we have a variety of

devices right primarily what we see these are mainly this can be this intermediate network devices which have resources to do that or there can be also dedicated fork devices and different data caching nodes and type like this. So, service placement problem in more specifically as we are telling that fog and age. So, what we are what we are trying to see last lecture if you remember.

So, if I have a cloud only situation things are like that whatever you are doing it is going to the cloud and getting executed right. Now you what we are seeing that given a overall application we want to look at it that it can be distributed over fog layer and the age layer right. So, in order to do that which service to be placed where is a big consideration right and it also varies from application to application right. So, some of the things which may be need to be calculated at the cloud end itself some could have been calculated at the age and down the line right. So, the fog computing as we see it is highly virtualized right that offers computational resources storage control between end user and the cloud server.

So, it is a intermediate layer as we discussed in last lectures. So, it basically try to give a compatibility or interoperate between the two layers at one set cloud and other is end user or where the edges are there and IOT devices. So, it is a so, it need to be it is a very so called virtualized platform. So, and what we see that our so called conceptually centralized cloud need to now work in a this sort of a distributed edge devices and some of the analysis required to be calculated in a centralized fashion where some can some of the call or some of the functionalities can be done at the much lower level. So, in order to do that so when we go for say distribution of resources.

So, we have several considerations like so some of them where which are pretty prominent are is location awareness right where it is there and location awareness of the things like specially if you say vehicular network the vehicle needs to communicate to the nearest RSU and type of things right and the vehicle is on the move then things goes in a thing. So, a hospital ambulance on the move needs to communicate the say patient related information health information to the hospital concern hot features or health care centers. So, has to take this data to the intermediate say for devices. So, which we which are more in the vicinity or it need to be aware of the location right in the both side right if I know that mobility pattern of the ambulance then I can basically know that where it is likely to be there and where the data transmission data uploading requirement will be there right. And of course, we what we try to see is in order to do that there is low latency right.

So, that is one of the our broad goal that low latency and so that our overall means latency is optimized and one aspect is that one other aspect is that better bandwidth replacement in doing. So, I should be again optimally using the bandwidth right. System should be scalable right that is one requirement that rather that is the paradigm what we started with the cloud computing overall this discussion at that is one of the major aspects of whole thing is a scalable. Now, in order to doing this if you if we compromise the scalability of the things then it may not make sense for the user at large. And with today's world where

mobility is the order of the day.

So, it should be mobility aware or it should have proper support for the mobility right. So, let it be whether it is a vehicular network or somebody some person moving from one place to another. So, whether it is a some other type of situation. So, mobility should be is a inherent thing. So, that means, in other sense the whole this paradigm should be aware of the things that there is a that there can be a mobility of the underlining devices.

There can be a situation where the your fog devices may be may not be very static, but nevertheless we still see that the cloud should be somewhat centralized and static. Now, again if we refer as I mentioned in my last lecture we are referring to this work which is very recent work on a survey work which accumulates different research things. I also encourage you to refer this paper which is available on the internet right very nicely written. So, if we look at the service placement problem. So, if we try to segregate it is a one is definitely the looking at the problem statement, then the what is the basic taxonomy of the service placement, what are the different optimization strategies and evaluation environment.

So, these are the four major group what they refer to also what we see that this can be a thing and if we look at the problem paradigm. So, it is it can be a looking at the infrastructural model, application model and deployment pattern right how the overall problem is there and service placement taxonomy. So, how the control plane has to be designed, what are the placement characteristics that overall system dynamicity and of course, this mobility support right. And when you look at the optimization strategies. So, what are my optimization objective any optimization problem I need to look at the what are the optimization objective and how what are the matrix for that and accordingly how I can formulate the problem and what are the this resolution strategies.

And similarly evaluation environment analytical tool whether I have a support of simulators of looking at it and experimental test beds whether things are like that is there. So, deployment thing. So, one example where which is referred in this particular paper let us it is a it will be good to look at it to have a more clarity say that is something a one video capture and analytics and recognition system is there. So, what it goes on it captures the vision right then like and then goes for a course feature extraction from the thing right what are the course features. So, it may be here we they refer to a Google glass.

So, a b may be the things could have been done out here in this particular device itself then it goes for a face recognition right. So, I want to go for a face recognition taking image from the go through the Google glass and if we go for the face recognition then it requires a more resources right the Google glass may not be able to do that. So, we fall back to something smart Wi-Fi gateway right it can be at the fog level or if I consider these are the devices which are the IOTs and sensors and then we can consider that as a at the age also devices

right it as I told you that there are different people look at it different way nevertheless it is not in the device itself it is out of the device right and if it is so, rather it is done here in a laptop which is connected to another device. So, it is it can be considered here as a age computing platform which supports this through a network device right. So, this smart a Wi-Fi gateway is the gateway where the traffic is moved to this device for the things.

I could have pushed it to the cloud for recognitions right, but it can be done in a much lower level right even this device had it been a more resourceful, but anyway face recognition is not that straightforward job that we can run on some devices like Wi-Fi gateway. So, there are other things like object recognition. So, what let us look at this left side. So, I require vision capture we require a course feature extraction like what are the things I coarsely I want to feature extraction it has two three things one is that face recognition that if there are faces objects like this there is a object recognition if there are objects and there are OCR or character recognition if there are written things. So, these three domains are if you those who have worked on this these are separately a large system building domain right they are they itself makes separate segment of development things right.

So, these three are there once this is done we need a learning based activity inference right. So, based on this I want to do some sort of a learning based activity inference and then with those and this inference we want to render and we can have an effective display of the things. So, what it is happening it is capturing the thing recognizing rendering the thing and putting appropriate display on the thing. So that the it I get a feel of it or we get a display out of it. Suppose this is my activities right so, our thing is not this application is not that added it if it has this type of components.

So, how we can map to our edge fog cloud type of things. So, as we see that in the device itself these things can be there A B C A B I vision course correction and your this inference can be there in this particular device itself right. And then we have the C here in the in this local loop I think there is a typo here it should be H right finally, it renders and effect the display out here itself right. So, it should be H it was typo as given as a I right nevertheless. So, these components can be worked out here I can distribute this C component in a local laptop which is connected to a smart gateway or local what we say computing server and this is a if we look at this rendering process which is which can be done out here right.

And whereas, at the back end cloud what we require this learning based activity inference has to be done at the back end cloud. Two reason one is the may be the computing need may be much higher right which cannot be done things like activities like this. In some of the things there it may require some of the other data from other sources right. It may not be true for this in some cases I require the data from other sources may be some models may be some other external data which may help me in the in that particular process which has to be in the cloud. Other than this object recognition it requires a much little more resources that can be run on a some sort of a LAN router.

There are there can be cloudlet which were rather people used to say that the cloudlets are more popular when this fog and age things were not much into the things. So, the large cloud base with the cloudlets which does that. So, nevertheless we can have OCR type of things in the cloudlets. I can also have OCR type of things in some other type of computing devices which can do this that what it is doing as a face recognition can do the OCR recognition also right. And doing all those things finally, we have a rendering which this smart in this case is smart Wi-Fi gateway does it.

And finally, this display effect which this I should have been H is done out here right. So, what it is what we are trying to showcase, what we are trying to showcase, what we are trying to look at it that a particular job of this capturing to having a means display effect of the thing has several components. There is a very example scenario it can be in. So, these components could be divided into the one way is that I could have pushed everything to the cloud and that cloud could have computed things and bring it back. In this case we distribute it in the different layers right some at the age, some at the fog, some pushing at the cloud.

What we try to achieve? We try to achieve that better efficiency less latency, better scalability of the whole system. And it may it also it may support the mobility in a better way like if this person with the Google gas moves from one place to another. So, another this Wi-Fi router can do it and work on the things etcetera right. So, that can be there. So, try to achieve nevertheless dividing this placing this different application different your activity into different layers is one of the major thing.

So, if we see that application placement problem defines the mapping of pattern by which application components links are mapped into the infrastructure graph right. So, I have a infrastructure graph that computing devices, physical edges and type of things how it is mapped that is the thing we want to do. So, application placement is finding the available resources in the network that is I need to find out that where are the resources are there not only the resources how much free they are and type of things. And there are there can be like application requirements which should not say suppose a I require for face recognition, I require a memory space of say minimum 500 MB or so right. And the resource where you want to run it that stuff it is not having that free space.

So, that it will not satisfy the things. So, what we require in some sort of a overall manager or some sort of a orchestration engine which try to handle those type of things right. It can be a centralized thing or maybe a thing or it can be a distributed and collaborative stuff, but it requires right. So, service providers have to take into account this constraint like limit of this space providing an optimum or real optimal placement and type of things. So, what we see that there are several constraints one is that resource level constraint like infrastructure is limited by the finite capability in terms of CPU, RAM, storage, bandwidth etcetera.

So, resource level constraint, network level constraints right such as latency, bandwidth and so and so forth which are at the network level. And there can be a application level constraint like locality requirement, restrict certain services, execute in the specific location etcetera right. Delay sensitive some application can be specific for a deadline and to be done in a things right. Some of the things that I some of the computing things I do not want that need to be at a should be done much within my control area. Like I can say that if it is from the if something is there to be distributed, but this applications you cannot leave say IIT Kharagpur network or computer science departmental network whatever has to be computed has to be computed here or I require some of the things which is more time sensitive or delay sensitive some applications.

So, I need to do something which is appropriately so. So, there are several constraints as if you remember in other placement problems also this type of constraints are there. So, as if you refer this particular paper if you see that the service placement taxonomy we have this control plane design that how overall this control to be managed. We have placement characteristics right system dynamic and mobility support as we are talking about in the initial things. So, if we look at the optimization strategies already we discussed some of the things.

So, our objective is this latency resource utilization cost and energy consumption right. So, these are the typical optimization strategies not only here in different type of when we do have distributed environment we do have this sort of optimization strategy do we do have this type of thing. So, based on that your heuristics or algos need to be designed like how things will be there like in the how I can optimize this service overall application placement. So, that we can have in the cloud fog edge environment a overall a optimal service into the things. So, one when we talk about all those things what we are trying to do we are primarily offloading some of the things right I could have done here I am offloading somewhere else right like everything could have been done in the cloud right.

So, we are offloading something on the fog some of the thing on the edge computing in order to attain those efficiency right. So, it is a offloading as we know it is a technique in which a server and application and associated data are moved into the edge of the network right. So, I would go on offloading the things rather. So, primarily it augments computing requirements of individual or a collection of user devices bring the services in cloud that process request from the device closer to the things. So, if we look at it has a two component one is from user device to the edge right.

So, augments computing in the user device by making use of edge nodes right usually a this user to edge is a single hop. So, one hop away of the things or in other things the device where it connects to the next things or operating from the cloud to the edge device. So, the a workload is moved from the cloud to the edge right. So, it is maybe server offloading there can be caching mechanisms where which helps me in helps us in doing this offloading things. In other things so, offloading is a well a popular or a I should not say use the word

popular, but it is a well studied or well research area where I how I can offload and how much I can offload things are like that.

Two things are important when I am be offloading the whole thing or I may be offloading the part of the things. So, once I offload I need to again the result and output I need to ensemble also right I need to aggregate those things. So, those it those things. So, this is the big picture of the things. So, what we see that from edge to device to edge and the cloud to edge and application partitioning and caching mechanisms are the popular things when we go from the device to the edge that is end device or either devices to the edge.

And then you both of them has different way of different characteristics and different methods to approach the thing to handle the things. From the cloud to edge we have the server offloading. So, it can be replication of the server to some of the edge things or it can be a partitioning of the things like the partitioning into in of the this server into different edge devices or some of some part of the things. We have the also caching mechanisms where this content popularity based or multilayer based caching mechanisms which are there. So, another aspects of this overall scenario is the control like how the control is managed whether it is a centralized control or distributed control right.

So, two aspects are there one as we are discussing that overall this management can be in a centralized way the overall control control can be centralized way or it can be a distributed way of managing the things. Both has different characteristics like things like solver based approaches or graphs matching based approaches are there when we go for centralized whereas, in the distributed block change graph game theory sorry game theoretic or genetic algorithm based different approaches are there. So, we are not going to the nitty-gritty of the things our first thing is that there should be there need to be a control mechanism otherwise we cannot achieve this sort of a I could not manage this overall this type of service placement and this managing the thing on the cloud fog edge type of things and it can be centralized or distributed control mechanisms. So, two aspects always come in hand in hand is one is the hardware part of it right or hardware infrastructure another is the software or the system software part of it right apart from the application software etcetera which are running. So, the fog edge computing forms the computing environment that uses low power devices right.

So, usually which are there low power devices like it can be mobile devices, it can be routers, something gateways, some home systems and type of things right. So, in other sense what we are trying to leverage upon that something is already there having excess resources why cannot we use it. So, low power devices a low resource devices right. So, combination of this so called low or small form factor devices connected with the network enables the cloud computing environment that can be leveraged by a rich set of application processing the internal IoT and CPS type of data right. As the huge volume of data being generated this IoT devices and different cyber physical systems.

So, this along with the cloud with this battery of devices with appropriate control mechanisms, service placement strategies etcetera we want to have a overall thing. For that I require my our hardware infrastructure should be compatible. One is that compute hardware computing devices another is network devices right single board computers or community products those type of things can be there that where the hardware is the computing. Otherwise my network devices like we seen in the example couple of slide back that gateways routers, Wi-Fi access points, edge racks all those things will be there. Other part is the system software for this fog edge or I should say cloud fog edge paradigm manage the computing network storage devices resources etcetera.

One is the hardware part and I we require a system software which manage this whole scope right. So, system software needs to support multi tenancy and isolation right that is that is one of the requirement. If you remember in our earlier lectures also we discussed about things when we look about the cloud computing infrastructure what we look at right, specially when we are looking at the IES type of deployment. So, there are two broad category one is the system virtualization. So, system level virtualization what we have seen earlier also or network virtualization right and if you look at the overall picture.

So, system virtualization where we have VMs virtual machine containers will we discuss something on containers will be taking up containers little bit more. Virtual machines and container migration how things will this is a major another big aspects of looking at the means of to the research community that how this migrate in appropriate way. And for network virtualization SDN or software defined network or network function virtualization. So, this is also another big area of research and big area of means virtualized virtualization world that is network function virtualization and SDN as you know that it has it is proliferating in a big way. These days all modern switches are mostly SDN enabled where the control plane and data plane are being segregated.

So, for overall better management of the network management and also we have overlay networks. Another part if we need to mention it will be in appropriate note that is the middleware. So, it provides complementary to the services to the overall system software. So, system software is the handling the this devices and allowing virtualization at system level and network level. Whereas, middleware in fogged computing provides performance monitoring coordination, orchestration, communication facilities, protocols etcetera.

So, that is very important right though the system software plays the say big brother role the this middleware actually supports it to have a faithful or reliable scalable operations right. So, there are different components like volunteer edge computing, hierarchical fogged computing, mobile fogged computing, cloud orchestration management and so and so forth. Another aspects as we look at the hardware infrastructure and other things other aspects is the algorithm. So, algorithm used to facilitate fogged computing. So, it has four major components so to say one is discovery, identifying the edge resources within the network that can be used for the distributed computing.

Benchmarking another important aspects the capturing performance of resources for design making to maximize performance of means what we are trying to do placement deployments etcetera right. So, there should be benchmarking of the resources which is of the performance of the resources which are at your hand. Load balancing so that it is not skewed distributed workload across resources based on different criteria like priority fairness and etcetera. And finally, the placement right what should be the algorithm for placement, identifying resource appropriate for deploying a workload, understanding that what is the requirement of the workload and piece of piece that how your things. For this what we require a different set of algorithms and also they have again major components like if we look at the discovery what is the programming infrastructure, hand checking protocol, message passing and similarly for benchmarking, load balancing, placement and things like that.

So, this is the given that hardware, given that system software, given the middleware in place and all these things how I implement actually how this overall things what should my algorithm so that part of the things right. So, that is another important aspect. So, what we try to see that overall in this overall resource management paradigm what are the major players and what are the major considerations right. They are in the sense what are the major considerations when we look for this resource management in a cloud fog age paradigm. More as we are looking at that as the resource constraint wise is fog and age is more resource constraint.

So, that is more is looked into rather than it need to appropriately synchronized with the cloud based on the resource availability and things like that. So, that we give something in a efficient way right. So, with this we conclude our discussion today. We will continue our things in the next class. There are few next session, there are few references these are very nice references I encourage you to look at those papers. Thank you.

**Cloud Computing**  
**Prof. Soumya Kanti Ghosh**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**

**Lecture 44**  
**Cloud Federation**

Hello. So, welcome to this course on Cloud Computing. Today we will be discussing on Cloud Federation right. So, we will try to see that why Federation is important and a overview of what are the basic approaches or architectures of cloud federation right. So, we will have a overview of cloud federation, the characteristics and federation architectures. These are the typical keywords.

So, when we talk about cloud federation, so what we are what does it mean? If we even by if we go by just intuitively it means that there are there are one or more cloud providers who are trying to cooperate among themselves to give a enhanced service right. So, it may be better this means serving the better user requirements, it may be the less delay or latency in giving the service fault tolerant and different. We will see that what are the why we require this type of federation architectures right. So, when we talk about federated cloud also referred as cloud federation is a deployment and management of multiple external and internal cloud services to match a business need right.

So, I have a broad business need and based on that it can be a means a set of cloud service provider or set of cloud can work together and do the things right. Like say if we take example that IIT Kharagpur as I mentioned earlier we have a in house open source cloud platform which is we refer as Megumala which is a internal cloud of IIT Kharagpur which is not accessible from the external world outside IIT Kharagpur network, but say some job is running on that say we are conducting a exam of say all first year students which is around 2000 students right. So, if it is not sufficient to handle by the cloud we can basically take a external from external service provider and though there are it is not that straight forward. So, that the applications which is running has to run seamlessly there are timing constant etcetera nevertheless when my capacity is full or my capacity in a constraint then we can take external or other cloud providers to do that or in some cases some of the providers can federate among themselves. So, that they can have a overall load balancing of the things better service provisioning and so and so forth right.

So, a federation by definition and by meaning also is a union of several smaller parts that perform a common action right. So, this is the way we look at a means meaning wise right. Similarly, here also we see one or more cloud service provider or one or more clouds try to provide a some service which is a as such in a transparent to that external world and it is gives a integrated form of the things right. So, different cloud are come together to give a

common service platform right or common to have a typical business rule to be executed in a seamless manner. Now, so collaboration between cloud service providers why we do that to achieve first of all may be capacity utilization right like somebody is constrained on the capacity somebody is having surplus capacity.

So, that it is I may if I am having a surplus capacity I may say that I am ready to lend it right definitely at some price. And also sometimes required to have enhanced interoperability sometimes the federation is self is a challenge sometimes I know that this portion can be done by better by this cloud CSP 1 other one is the CSP 2 and then we can at the back end talk to each other and this can be we can interoperate. So, also cataloging of services. So, these are the type of services provided one this CSP 1, 2, 3 and this providers can say that their catalog and say other catalog can have a as a what we say a global catalog of this federation and it shows that these are the services we are providing. And inside about providers and SLS that is one another requirement when that how much we allow to during the federation.

So, we can have some sort of a sharing the whether it is loosely capable or tightly capable based on that how much individual cloud expose their means infrastructure or they are back or back end configurations and the SLAs what they maintain with others those things make a there may be based on the architectures we are following, but nevertheless there is some sort of a sharing is required when we federate. So, motivation as I was mentioning. So, first of all in order to make this federation different CSPs join hand and so, to maximize resource utilization that is one of the primary goal. So, if somebody is having surplus resources and somebody is resource constant. So, that can be utilized or we can overall we can have a large resource pool to cater a variety of customers.

Minimize power consumption sometimes it may so, happen that overall requirement is distributed and we can achieve some efficiency in terms of power consumption. Load balancing definitely that if something is more loaded than the other. So, it can share that loading and global utility. So, utility wise also it is enhanced and having different service by the service providers and have a more say encompassing or more services can be provided and of overall it increases the CSPs footprints right. So, like if you will go to the other previous slide say what we see that CSP 1, CSP 2, CSP 3 is there.

Now if they are federating the CSP 1 maybe now footprint instead of his own footprint now it is share the footprints of all the three. So, it is a that maybe a chance to expand a individual footprints right. So, which makes sense in case of their marketability and profitability of individual CSPs. So, characteristics as we are looking at so, one is to overcome the current limitation of the cloud such as service interruption where it is loaded all there are faults etcetera lack of interoperability degradation of the services those things can be they can be handle if there is a defined or well established federation between more service provider. So, in case of overloading or if there are some sort of a fault or interruptions so, that the other members of the federation can take care of those situations.

Many inter cloud organization have been proposed like how this that organizing or how this working of the inter cloud will work cloud federation is one such inter cloud organizations right. So, in other sense it is a inter cloud organization in voluntary characteristics like who wants to join the things maybe voluntary it is usually not regulated by a things that you need to federate with this two other cloud provider type of things it is more of a voluntary individual things it should be maximum geographical separation. So, it is a large geographical span or catering to a larger region or the means of the service area. Well defined marketing system and regulated federated agreement right. So, there should be a well defined market system and once we collaborate there should be a what we say some agreement or some basis for collaborations right what would you say some regulations of or agreement over this federation right.

Another important means things what just means that maximum geographical separation some what why one of the thing what we try to give in a cloud environment is what we say is high degree of reliability right. So, that it is even so, ideally so, ideally this federation partner should be in different maybe power zone different network zone different seismic zone and type of things right. So, spread such a way that it is not like that a some say environmental or some disaster may affect all the federating partner right like. So, if so, others will be working as usual so, that that federation can effectively can support the their client base right. And it is an environment where multiple service providers come together and share their resources right.

So, it is a what we say a ecosystem so, called or a environment where the multiple survive service providers can share their resources to maximize what we say utilization of the resources satisfaction of the user community and so on so forth. So, when we talk about architecture or federation architecture so, associated with several portability and interoperability issues right. So, it is not pretty straight forward when once say 2 CSP, 3 CSP or n number CSP talk to each other or agreed upon federating and giving a single interface or sorry single say what we say view of that globe of their federated structure for some purpose definitely for some goal and type of things. So, what is important at the back end though the issues like portability between service provider, interoperability issues with respect to services, data and different challenges on the interoperability issues will need to be addressed right. So, if we look at typical federation architecture so, they those are one is cloud busting, brokering, aggregation and multitask right.

So, this is these are the typical federating federation architecture which are being practiced or which are prominent or important architectures. So, these architectures can be classified according to the level of coupling or interoperation among the cloud instances involved right, ranging from loosely coupled that means, no or little interoperability among the cloud instances. I should not say no interoperability there is there has to be interoperability otherwise you cannot make the federation working, but what is there that what it requires what this loosely coupled what it is basically do not have much information about the other

providers say infrastructure and things like that right. So, it is less what we say less transparent to each other right. So, it is not all is all is feature sets etcetera exposed to the other providers right.

So, that is loosely coupled or we can have a tightly coupled where the interoperability is very tightly managed among the cloud instances even they say that the type of architecture and other systems things what they are maintaining at their end. So, it is very tightly coupled. So, there are both need and advantage disadvantage whether we go for loosely coupled tightly coupled things based and it is more based on the applications what you are looking for right. Some of the things may be loosely coupled some may be tightly coupled and if it is my federation at different level say at the as a higher level or SAS level then something is there or if it is in the infrastructure level then something else is there. So, there are different challenges and issues which need to be addressed.

Now to look at if we have loosely coupled federation architecture. So, limited inter operation between CSPs and cloud instances. So, there is not much inter operation between the CSPs and other cloud instances right in case of a loosely coupled. Like for example, a private cloud complementing its infrastructure with resources for an external commercial cloud for some requirement. Like we are telling that we are running a exam and we found that the basic cloud infrastructure what we are having it not able to serve in a in a faithful manner.

So, what we try to do that we purchase more cloud resources or subscribe to cloud resources which can be other public cloud or other private cloud, but the challenge here is how wow is so a integrated fashion. So, as we do this type of federation these are all loosely coupled right. So, they are not exposing everything to you and you are also not exposing your feature set to the external world, but on a on the other hand you need to run the what we say quote unquote business process in this case may be running that exam in a particular stipulated time duration and that that makes things pretty challenging. So, a CSP has little or no control over the remote resources right. So, if it is a loosely coupled say IIT Kharagpur, Meghalwala cloud with some other cloud.

So, CSP do not have much control to the resources right. Remote resources like these and about the VM placements are not known. Monitoring information is limited for example, only CPU, memory, disk consumption is VMs. So, these are limited right. And usually no support for advanced feature such as cross side networks or VM migration and these are the things which modern day cloud supports those things are if you have a federation things that is difficult to support all those things right.

First of all they may be having their underlining different sort of configuration at their end and type of things right. So, this VM migration what we will see in subsequent lectures will be may be a challenge out here. The other is the partially coupled federation different CSPs or the what we say partnering clouds establish a contract or framework agreement stating

the terms and condition under which one partner cloud can use the resource other. So, it is a partially coupled that means, not so loose as loosely coupled not so strong as a tightly coupled, but in between them partially coupled where they establish a contract or a framework agreement stating that the terms and conditions and under which one partner will share resources with the other and so and so forth right. So, this is also good and this is also help in achieving better efficiency than may be pure loosely coupled thing.

So, this contract can enable a certain level of control over the remote resources right. So, this contract as we have a contract previously loosely coupled we do not have any so to say any formal contract actually there should be some agreement definitely otherwise how you federate, but that is not that strong, but here it allows to some control over the remote systems like allowing definition or say affinity rules to force two or more remote VMs to place in the same this etcetera. Like you want to have that this for this particular working this three VMs should be on the same physical cluster. So, those type of things are may be allowed right it depends on installation to installation, but this is things like that. So, may agree to interchange means more details monitoring information for example, providing information of the host where the VM located energy consumption etcetera right.

So, it may help in exchanging or interchange or more monitoring information right and may enable some advanced networking feature among the partner crowd. Also it may enable some advanced networking features partner crowd like creation of virtual network across boundaries and type of things right. So, what we see here that more flexibilities or more hand holding between the two cloud providers are given with respect to our loosely coupled thing. Now we other thing what we are having tightly coupled right other totally other side tightly coupled in this case the cloud are normally governed by the same cloud administrator right. So, once you have to a tightly coupled in other sense you require more information about the other party in other if we see the other sense it is basically the type of stringent requirement if it is there for strong coupling then we have some sort of a some single administrative authority at times right or if it is also different administration it reports to a single authority.

So, that anything you do here the other party or other parties know immediately or a priority. So, that they can be ready with their configurations and things like that right. So, a cloud instance can have advanced control over remote resources like for example, allowing decision about exact placement of area of remote VM and type of things right. So, that type of these and that which cluster where this VM will be placed and so and so forth will be more stronger and can access all monitoring information available about the remote. So, in loosely coupled we do not have any practically anything in partially coupled we have some control, but here availability of the information about remote resources is much high.

May allow other advanced feature including creation of sort cross side network, cross side migration of VMs right implementation of high availability techniques among the remote cloud instances creation of virtual storage systems across side boundaries right. So, when

we have a tightly coupled we have other advanced features of cloud can be looked into like creation of cross side network, cross side migrations of VMs if there is a loading that I can migrate to the VM without much negotiation right it is already agreed upon right implementation of high availability techniques when we require among remote cloud instances right and creation of virtual storage. So, as storage is one of the important factor of or important consideration in any implementation. So, this virtual storage systems across side boundaries all those things are maybe they are in the tightly coupled federations. Now, if we see the a broad picture of the thing.

So, we see there are four category one is hybrid or bursting architecture another is broker architecture right another is aggregated another is multilayer architecture right. There are some commonalities between this architecture or between any two such architecture, but nevertheless their overall focus is different or based on different type of business need the what type of architecture you will be looking at may be important. So, we will quickly try to see that what they try to mean right. So, bursting or hybrid architecture cloud bursting or hybrid architecture combines the existing on premise infrastructure usually a private cloud with the remote resources from one or more public cloud to provide extra capacity to satisfy peak demand periods right. So, what we see that especially when the requirement is pretty high or what we say when it is reaching to the peak sometimes what happened we can predict some of the peaks some of the peaks we do not predict, but if we as we have seen in earlier lecture if we design of our my cloud based on my peak requirement then many of the time it may be underutilized right.

In other sense if I have a federation in place and if required I can have a hybrid or burst out to the other things and get my work done definitely there will be some requirement of SLA quality of service and of course, there may be some financial constraint also like charging when you use the other things, but if it is done seamlessly if you can see that from the customer point of view it is a you know it is a something running and going on run even if there is a constraint in the provider rate. So, as the as the local cloud OS has no advance control over the virtual resources deployed in the external cloud beyond the basic operation and providers allow right. So, has no advance control over the virtual resources. So, of a local or one service provider do not have much advance control over the virtual resources the architecture is typically loosely coupled. So, more existing most existing cloud manager support this hybrid architecture right.

So, it is loosely coupled and it is well taken by this different CSPs. The other one is a brokering which is more sounds to be more intuitively sounds to be ok or it is more natural. So, a broker that serves various users has access to several public cloud infrastructure. A simple broker should be able to deploy virtual resources in the cloud as selected by the user. So, in this case we have a broker or which acts as a managing the resources of one or more cloud service provider.

So, broker is the common most common federation scenario. So, that makes more easier to

implement, easier to maintain and easier to keep the what we say quote unquote privacy of this individual cloud providers right. So, that is there and say as we are looking at that advance broker offering service management capabilities could make scheduling descent based on the optimization criteria like cost, performance, energy, consumption to automatically deploy virtual user service in the most suitable cloud. So, what he is telling that the cloud as the broker is there. So, request coming there it can take a more global descent based on the what type of resources are there, what is the requirement and type of things like that right.

It may even distribute the service component across multiple clouds. This architecture is also loosely coupled not that strongly coupled at times though people refer it is something say means not tight or loose it is in between that is partially coupled type of things. Since the public cloud typically do not allow advanced control over the deployed virtual resources right. So, the public cloud may not allow. So, it is difficult to go there, but definitely broker acts as a thing and if we have multiple cloud infrastructure the broker will help in some sort of a load balancing, finding out appropriate service requirement and responses to those things right.

So, this is typically the broker architecture. So, we have a federation broker they are these are the external user it can be human user or system it does not matter and there are different CSPs and this broker interact with them to give the overall increase the efficiency. The other one is aggregated architecture involves two or more partner clouds that interoperate to the aggregate or rather interoperate to aggregate the resources and provide users with a large virtual thing. So, what we do as the name suggest. So, it is aggregation of the resources right and gives a aggregated view to the external world right. So, this architecture is usually partially coupled since partners could be provide with some kind of advance over the remote resources right.

So, it is partially coupled unless you have some information about the things a priori or even on the fly aggregation all those things are not feasible usually the aggregation decent etcetera based on the job type of job etcetera are somewhat pre decided or done not on the fly it is possible that to do on the fly, but there are different constraints. The partner clouds usually have a higher coupling level when the when they belong to the same cooperation and when they owned by different companies agree to cooperate and aggregate the resources right. So, this if you see this partner cloud to have a. So, there are constraint like they want to collaborate or it may be owned by different say it can be different CSPs all together, but agree to cooperate and aggregate their resources. So, that means, they need to share some information about the resources which is permissible based on their own business rule or corporate rule and then it can aggregate the resources.

The other thing is the multitask finally, we have that multitask involves two or more cloud sites each running their own cloud OS and usually belonging to the same corporation that are managed by a third cloud OS instances following a hierarchical arrangement. So, in the

in case of multitask instead of having all flat things we have a hierarchical structure right. So, some sort of a root or top cloud OS instance has a full control over other resources in different cloud size. So, it is a tightly coupled scenario and it is it exposes the resources available of the different cloud sites or the cloud providers domain and if they are located in a as if they are located in a single cloud. So, in hierarchical it is the control is in a hierarchical fashion and but they are more tightly coupled unless that is otherwise it is difficult to maintain those thing for a for executing a common goal.

So, this architecture is beneficial for corporation definitely with geographically different cloud infrastructure because it provides uniform accesses right. Like if I say that I have a organizations which have say in all major cities of India and some of the some second level cities etcetera. So, making them hierarchical it make me sense that to manage the whole thing and there may be different requirement and different regions. So, it can be in like a hierarchy it can be maintained. So, it may be useful for implementing advanced management features such as high availability, load balancing, fault tolerance right.

So, once you have a this type of multi-tier hierarchical architecture. So, I can have we can have different advanced features which can be worked on right. So, in this today's discussion what we try to look at is this cloud federation that when one or more cloud service providers come together and say quote unquote join hands say join hands to give a to deliver a common business process then we require federation. We have seen that there are different sort of federation architectures which are being practiced by the things right hybrid or bursting, broker, aggregated and multi-tier architectures which helps us in working with these things right. So, with this let us end our discussion today we will continue with we will in the subsequent lecture or lectures one or two lectures we will see we will see another important aspects of this cloud computing which is VM migration right.

So, that type of things we will see. So, there are few references and which you may go through. So, let us end our discussion today. Thank you.

**Cloud Computing**  
**Prof. Soumya Kanti Ghosh**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**

**Lecture 45**

**VM Migration - Basics Migration strategies**

Hello. So, let us continue our discussion on different aspects of this cloud computing paradigm right. Today we will have we will start a discussion on VM migration. So, virtual machine as all of you know that it is a one of the major service what a cloud provider provides right. It is primarily as a in the level of IAS or infrastructure as a service. So, what you will do? You can have your own virtual machine configured based on ideally configured based on your requirement when you request for the things right.

In reality what will happen is that there are different flavors of the things are available like different configurations available like if you have used any open source like OpenStack or any commercial like from Amazon or any other sources. So, you will see that the VMs there are different flavor available like. So, different category of virtual machines are there. So, you can basically subscribe that machine for a period of time and then work on that machine and then release when you do not require that machine that VM right.

The challenge is sometimes. So, what is happening? There are several servers where the virtual machine are initiated where the users are using that consumers or the users of the VMs are using this right and then there may be a situation like and suppose there are there is a server where or a particular server there are say N virtual machine and may be every server has a capacity of say M number of machines somewhere N1 somewhere N2 N3 VMs are running right and what is happening like once it is coming it is getting requisition in one of the VMs then sometimes once that work is over there is release that gap is created and the overall cloud that cloud broker so called or the controller units is going to allocate the things. And in a in this type of scenario one thing may happen there may be imbalance in the thing some of the servers are loaded heavily some are less loaded and so you may need ideally need to migrate some of the VMs working in this say server 1 to say server 6 because the server 1 is more loaded right or server 3 to server 4 and server 14 anything right. So, there I require a virtual mechanism. Secondly, some of the things may require a shutdown that or maintenance go may need a maintenance right.

So, once there is a need of maintenance then also you need to shut down that system. So, all the VMs need to be migrated into the things. So, this migration is a major challenge or is one of the necessity what the virtual what we see in this type when we do any work with this when we think about the data centers or the cloud providers point of view right. So, in this talk or in couple of and next talk also we will see that what are the different way this VM migration can occur or what are the what is the basic philosophy behind it. So, we look at

the VM migration basics and the VM migration strategies right.

So, VM migration so, what we see that if you see that when we talk it is a process to move running applications or VMs from one physical server host to another physical server as we are discussing right. So, it was running in a host server and then I want to move it to the another server then maybe several needs we will come to that right. So, the once a process is running this is a this is a big challenge one way of is that you shut down everything put it to another switch on the things right. So, that is that is that is possible that we will see that that is one way that is possible and that may be the easiest way of migrating this type of thing shut down the process and re invoke or invoke the process in the other machine or invoke the VM into the other machine right. But the challenge is that always it may not be possible the consumer or the user may not allow you that that no that is that may not be feasible because my process is running for days together I may be running a simulator or I am writing a critical job and then you say that no no no you shut down.

So, that may not be always feasible right. So, then I am going to do a live migration when everything is on then you migrate on the things right. So, what we try to migrate that process are state right if it is a thing storage, memory, network connection are moved from one host to another right. So, process are state that whatever the storage things are there memory locations. So, that what is whatever in the so called RAM and then the network connectivity whatever is there if there are connectivity and it is going away some network level application is going on that need to be reconnected there right.

So, this the type of configuration out here in the machine one or the what we say source machine and the whatever there in the destination machine there should be a match there should be that connectivity because suddenly the if there are external connections and doing some network operation they will be in a for a spin. So, that is important. Why to migrate? Distribute VM load efficiently across servers in a cloud that is one of the major one of the reason what we are telling that load balancing so called. So, one because allocation deallocation of the VMs what will happen the some of the server will be more loaded than the others and then you may want the servers to be more distributed the loading to be distributed. So, that means, distribute the VM load efficiently across the servers in a cloud environment right.

Another is the system maintenance I may want to do some maintenance for a server and in that case I need to migrate those things right. So, this is that typical picture I think we that this one of the thing you have we have seen earlier also or in several literature. So, what is there the hardware then the VMM or the hypervisor is there over where differ different VMs are there if this is another instance another server right. So, I can migrate one to here right. So, that is the ideally one in one server one stack of VM is running and then I want to migrate on the thing right.

There are there can be either whole everything is migrated or some of the things migrated

there some other servers and things like that right. So, nevertheless this what we see that virtualization what we have looked studied in earlier means previously. So, I have VMs then I have virtual machines which have some case to S in this case Linux something net VST, Windows and over that different applications are running right ok. So, if we look into the more deeply in the need of VM migration one definitely the load balancing as we discussed that for fair distribution of workload among computing resources right. So, there can be computing several computing resources and I want to have a fair distribution of the workload of the thing.

Maintenance for server maintenance VM can be migrated transparently from one server to another right. So, the server is going for the maintenance the customer is not bothered about it, it wants that SLA to be respected and its application should be running as smoothly as was there. So, that time because you need to shut down a that server where those VMs has to be migrated here. Managers in some cases manage operational parameter to reduce operational parameters like power consumption, VM can be consolidated minimal number of servers, underutilized server will be put on low power mode etcetera. Like I have 5 servers and maybe in some cases the other way of the looking at the load balancing is that that 5 VMs are running in 5 servers, but it could have run in a 1 or 2 servers right.

So, I migrate them to the things and the other things I put it in a low power mode when the demand is low right. Then maybe QoS violation type of scenario when the service provider fails to meet the desired quality of service a user can migrate his VM to another service provider. So, this is not only the say one service within a service provider inter service provider like one server or one DC to another DC, it is something like one provider to another provider it wants to migrate right. So, that can be another situation fault tolerance in case of failure VM can be migrated from one data center to another where they can be executed. So, there can be other fault tolerant type scenario where we need to do some migration when there is when I am expecting when the system is coming down and there are faults and etcetera right.

So, there are various reasons as you see one maybe for load balancing, maintenance, manage operational, the user wants to be migrated somewhere and maybe some of the exigencies like or fault tolerant type of scenario. Now if you look at the migration things so, broadly there are two types right. One is cold or non-live migration right another is a hot or live migration. So, this is the broad category of VM migration right. So, one is cold or non-live another is a hot or live migration.

So, what is there in the non-live? In case of cold or cold migration the VM executing on the source machine is turned off or suspended during the migration process. So, it is during the migration process the VM is either shut down or turned off or it is suspended and then it is migrated and then switched on right. So, in other sense it is a non-live. So, it is not doing the on the fly when the things are running. So, it is it actually takes a some sort of a quote unquote downtime from the user they may have to pay some penalty or some cloud credit

as a result of that and that is migrated to the other VM right.

Another one can be hot or live migration right. In case of hot or live migration the VM executing the source machine continues to provide the service during the migration process. So, there is the this in case of hot or live. So, it goes on providing when the whole migration process is going on. In fact, the tri-grade VM is not suspended during the migration process.

So, when it is not suspended it is live thing it is going on. This is pretty tricky definitely there are lot of consideration out here and, but at the same time it gives a somewhat seamless performance for the user right. So, user practically does not know the migration process is going on. So, it was working and then migrated to the other thing and it is VM remains as the VM things itself right. So, that can be hot or live migration.

So, what we see that there are two broad categories one is little simple or easy to conceptualize, easy to implement that is the cold or non-live migration. Other is hot and live migration is there are this is a tricky thing and requires a defined process that or how it will execute right. So, if you look at the live migration. So, migrate an entire VM from one physical host to another all user of the user processes and the kernel state need to be migrated without having to shut down the machine right. So, what we are doing the VM from one physical host to another it is migrated all user processes which are running in that original VM before migration and the kernel state remain need to be preserved or should go on running in the next location without any interruption and you do not have to shut down the machine per say right.

So, in case of whereas, in case of non-live migration VM provides providing services remain suspended during the entire migration process right. Hence the large size VMs the service downtime might be very high right. So, there can be large downtime thing and there can be penalty for this type of downtime in case of a non-live or cold migration. So, for real time application non-live migration can cause severe degradation in the service quality which is not tolerable. So, if there is a real life application like something online broadcasting going on or some online some computation going on and things like for some real life applications and then you have a non-live migration then you have a severe degradation problem.

So, two main approaches one is pre copy and post copy approaches are there for this type of migration process right this live migration process. So, when to migrate to remove a physical machine from the service that is as again and to relieve load on the congested or less seen. So, VM migration if cost wise the cost associated with the communication overhead like if you are migrating the whole thing whole state and other things to the things there is a communication overhead and cost associated with the migration time downtime. So, this is the costing on the things right how much migration time, how much downtime and then effectively affects the overall performance of the cloud and at times it may violate the SLA for which the CSP has to pay for it pay the penalty for it right and also it hits its if lot

of downtime and my time times are there then it also hits its reputation right. So, these are different typical factor what we intuitively also try to do.

So, based on that so, the major concern are the minimize the downtime right definitely minimize the downtime. Downtime refers to a total amount of time services times the services remain unable to the user. So, downtime what we say that that overall time so, long the services are unavailable to the user that is the what we refer to as downtime.

Minimizing the downtime is one of the major challenge minimize total migration time right. So, migration time refers to the total time taken to move a VM from the source to the destination host right.

So, I have a VM running right and then the whole VM is run migrated to the another host another server and it is running faithfully. So, no nothing is left behind means like no memory no state and other things right. So, the total time requires whether you want to shut down and revoke or do it migration etcetera that is the migration time. So, it can be considered at the total time taken for the entire migration process right. So, that is why minimum the migration time better your performance because migration is always a overhead over the whole system.

So, migration does not necessarily disrupt the active services through resource contention that is CPU network bandwidth with migrating ways right. So, what it tries to do that your there can be resource contention CPU network bandwidth etcetera where you are where at the destination source and destination that need to be taken care. So, now what to migrate as we started the discussion with one is the CPU context of the VM like contents of the main memory these are the two things need to be migrated because as if you as we have shown the figures. So, it was running on a hypervisor right. So, and it has a this CPU context of that particular VM and the contents of the main memory of the things has to be migrated because now it is be running on another hypervisor too and those has to be instantiated.

Disk so, if there is a network attached storage that is accessible from the both host. So, that is one way of handling. So, both host are can connect to that particular NAS storage right or NAS server and that is accessible board host or local disk is mirror migrating disk data may not be critical. So, if the if it is on a NAS server. So, this disk data is still you just need to reconnect now the connection is from this particular port or this server and now it is from the other port that may not be very very challenging right, but still there are netigity is there.

Network assume both the host on the same LAN. So, if you assume both the host of the same migrate the IP, advertise new MAC address to the IP mapping via IRP reply and all those things right. So, if it is much easier or if it is on the same LAN or I need to do other things like proxying and other things has to be there. So, migrate MAC address, late switches, learn new MAC locations, network packets redirected to new location with transient losses and all those things will be there right. So, there are challenges, but there

are ways out in a sense that if it is within the same service provider they can extend the LAN and things like that can be possible right, but nevertheless network is an issue to be looked into.

Then I/O devices, virtual I/O devices are easier to migrate if it is a virtual direct device assignment to physical devices. VMs may be difficult to mean physical devices to the VMs may be difficult to migrate. So, that can be a if there are virtual devices then it is easier to migrate, but if there are direct devices that has means device assignment of physical devices to the VMs which may be difficult to migrate. Then you have to take care of that how this connectivity will be there even if it because the physical device needs to now connected to the other things right. So, if we look at the memory migration steps. So, one is that push source VM continues running while certain pages are pushed across the network to the new destination right.

To ensure consistency pages modified during the process are reset. So, what will happen the source VM is running along with that the memory pages are pushed to the destination while the source is running. So, whatever is getting modified out here or that it so, that has to be retransmitted right. So, go on retransmit it. So, stop and copy source VM stop pages are copied across to the destination where then the NIN VM starts right.

So, this is thing. So, you push it then stop and copy and then pull. New VM executes and if it accesses a page that is not found in the copied the page is faulted in the pulled across the network from the source. So, that means, there can be push there can be stop and copy or pulled the when the source is running when the destination VM or the migrated VM is running it does not find some page then it can be pulled across the network from the source VM right. So, that is say when the live migration is going. So, pure stop and copy simply simple, but both downtime and total migration time are proportional to the amount of physical memory located in the thing.

So, pure stop and copy means you just stop copy and then do then what will happen that how much migration time is dependent on the amount of physical memory allocated to the VM. May lead to an unacceptable outage if the VM is running via live services right. So, if it is something which is live service is running then this outage may be may not be acceptable to the consumer or the user right. So, this need to be taken care. So, if the life some services is providing then this type of things becomes a serious challenge.

Pre-copy phase it is carried out over several rounds right when we have a pre-copy phase the VM continues to execute at source while the memory is copied on the destination. So, this is the pre-copy phase right. So, it continues in a source and then go on copying to the destination while the memory is copied to the destination right. Pre-copy termination phase. So, when should I stop because if it is running then the some pages are getting modified and things like that and so, the stopping criteria for the pre-copy phase takes one of the following thresholds like the number of rounds executed threshold.

So, I say that there will be n rounds and it is over total memory transmitted exceeds the threshold like how much memory can be transmitted over the things it is above a threshold or the number of dirty pages in the previous round drops below a threshold right. So, when I am basically any modification then only I need to transmit if there is no modification then I do not need to transmit that to the destinations right. So, when the number of dirty pages is below a threshold then I can stop. So, some of the stopping criteria and stopping and copy phase in this phase execution of the VM to be migrated is suspended at the source then the remaining dirty pages the state of the CPU is copied to the destination node where the execution of the VM is then execution of the VM is resumed. So, what happened the execution of the VM to be migrated is suspended at the source.

So, it is suspended source then the remaining dirty pages CPU state are copied to the destination and while then the execution of the is resumed at the destination right. So, that type of that is the stop and copy phase. So, eternity pre copy live migration in case of eternity pre copy pre copy this phase may be carried out over several round as we are discussing the VM continues to execute at the source while its memory is copied to the destination active pages of the VM to be migrated are copied iteratively at each round during the copying process some active pages might be dotted right there may be change in the things as we discussed at the source host which are again resend to the subsequent round to the things right. If there is a update of the things so, that page to be re-transmitted pre copy termination what do we already discussed that either the rounds are these total this is basically it is more the same thing what we discussed in a more organized way and other total memory is transmitted exceeded threshold or what we see the number of dirty pages in the previous round drops below a threshold those can be there and stop copy already we discussed and there is a restarting phase the restart the VM at the destination server. So, what we see when we try to do a live migration this is a very pretty complex process and also what we are not looking at that what we are considering that the network delay is minimal right.

So, because if it is a far away migrating from one data center to another center there will be a serious challenge that the delay in the overall communication path right. So, these are the things what we look at and post copy live migration. So, what we these are stop phase stop the source VM and copy the CPU state to the destination VM right restart the destination VM on demand copying copying the VM according to the demand. So, the on copying VM memory right. So, on demand means as we discussed in the pool like I there is a in the destination there is a it found that one some page is not found then it is pooled or on demand it is copied from the source things.

So, in the post copy strategy when the VM is restarted the VM memory is empty if the VM tries to access the memory page that has not been copied the memory page needs to be brought from the source VM. However, most of the time some memory pages will not be used. So, we need only a copy of the VM according to the demand. So, that means, it is not

like that all are need to be copied. So, in reality what will happen you have that things which are there and whatever is on demand you are basically migrating those.

So, with this let us conclude our today's discussion and we will continue this migration things in our next subsequent talk and there are very some of these nice references you can have a look into the things right on VM primarily on VM migration mostly more critical on the live migration of the VMs right. With this let us stop our today's discussion. Thank you.

**Cloud Computing**  
**Prof. Soumya Kanti Ghosh**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**

**Lecture 46**  
**VM Migration - Basics Migration strategies**

Hello, continue our discussion on VM migration. So, last lecture what we discussed is that how what are the need of VM migration and why it is important and what are the different way or mechanisms to do that right. Today we will try to see little bit of analysis of that how to do some sort of a mathematical model or how to approach those problem in today's lecture right. So, today we will do some VM migration analysis and also we will be looking at the strategies. So, these are the keywords. So, just to quick recap.

So, it is a when we talk about migration is a process of to move some running application or VM from one physical server or host to another physical server right. So, process state, storage, memory, network connection, if there is active IO those has need to be taken care of. As we discussed in the last class, so if we look at that say the process state or sometimes we refer as a CPU state is a important factor we need to be taken care. Memory that means, the where the VM or the is working on that is the active memory or the main memory which need to be migrated.

And of course, in some of the cases networking because not some of the cases networking is important because now the your if you see the your connectivity is not longer in the in that VM it is somewhere here. So, some sort of a mapping of this IP to MAC and things like that has to be looked into right or some sort of a ERP related issues need to be resolved. And in some cases if there are active IO's, so that need to be looked into. So, that last class we discussed right. And also why to migrate that is one of the major reason is that if there is load balancing and system maintenance.

We discussed few of the things some of some may be the same thing, but it may be little there may be a wording difference. So, one definitely is load balancing right. So, when the load is considerable considerably unbalanced like if you have host where unbalanced loads and impending downtime often require migration of VMs right. So, in order to have a more faithful provide more faithful service by the provider this migration is one of the things which we may look into. Other things what we looked at as the power fault tolerant right.

So, is another challenge to guarantee critical service availability and reliability right. Failure should be anticipated and proactively handled rather in some of the reference you will find that what we say proactive fault tolerant. So, we if there are means failure then need to be proactively handled right. Power management, so switching the idle mode server to either sleep mode. So, the server which are idle mode or less load and sleep mode or off mode based on the resource demands right that leads to energy savings, VM live migration

is a good technique of doing that.

So, overall energy requirement of the particular data center need to be looked into where you can look into the idle server where the server having very less number of VMs that can be migrated to other servers and some of the servers can be made free. So, that they can go for a sleep mode or sometimes in a off mode that means, they will consume less power than expected while it is in a running mode right. Resource sharing, a challenge of limited hardware resources like memory, cache, CPU this cycles can be resolved by relocating VMs over from overloaded server to the under loaded server. So, that is other way around. So, if it is if there is a lot of punch that it basically goes on a some sort of a load balancing sort of things, but here the major challenge is that resource sharing.

So, if we need in some cases you require that these are the say CPU hungry operations. So, those may be shifted to some of the things which has more resources on that font and type of this right. So, there is a resource sharing of course, last but not the least rather one of the major requirement is the system maintenance right. So, a service provider putting I mean shutting down the system and for or taking down time for a long time is always a negative point right that affects its credibility and revenue as well right. So, physical system required to be sometimes the R graded goes for a system maintenance and so, the VMs in that physical system must be migrated to alternate server.

So, that the service downtime is minimized or you provide can be provided uninterrupted services. So, this we things we have looked into or seen this scenario right. So, these are the major things which we require when it is migrating and also we looked into two major way aspect that is one is pre copy approach another is post copy approach and just we have a quick little bit more deep in depth into the thing. But whatever we are talking about is live migration scenarios right. So, uses iterative push phase that is followed by a stop and copy phase in this pre copy approach because of iterative procedure some memory pages have been modified or updated or modified what we refer as a dirty pages right when it is going on transferring this page can be modified at the other end.

So, you know if you dirty pages are regenerated on the source server during the migration iteration. So, after migration iteration going on this pages can be at the source can be modified or updated and that creates this dirty pages. So, these dirty pages need had to be resend to the destination host in a future iteration right if there is a dirty page. So, need to be send to the things. Hence some of the frequently access memory pages are sent several times.

So, what we see that some of these memory pages are sent several time it causes a long migration time because it is once there is a dirty page and things like that. If you remember we called about that when it will when the dirty pages are below a threshold or ideally there is no dirty page then we say that the whole thing has been migrated properly. So, if we look at the first phase or the first step towards this pre copy migration is the all pages

are transferred while the VM running continuously on the source source right further rounds dirty pages are resend right. So, what we see that all in the step 1 all pages are transferred while the VM still running on the source source. So, automatic source source host.

So, automatically there may be updation of the pages. So, this dirty pages are resend right. So, second step is termination phase which depends on the defined threshold right at where you terminate this and put the CPU state on the other thing and then start working on the things right. So, the termination is executed if anyone out of this three condition we seen earlier also the number of iteration exceeds the predefined number of iterations right. I look for say n number of iteration and it goes on more than the defined.

So, that is what we say that it is crossing n max type of things. The total amount of memory that has been sent is above a threshold or the number of dirty pages is just previous round fall below a defined threshold. So, if the dirty pages fall below a defined threshold then we stop this step 2 or the pre copy approach the second phase of the things. And in the last stop and copy phase migrating VM is suspended at source server after that the move the processor state and remaining dirty pages to the destination server and then start working on the on this VM is initiated or started at the destination server. So, when the VM migration process is completed in the correct way that means, all the steps are there then what we do then the hypervisor resumes the migrant VM on the destination server.

So, the hypervisor resumes the migrating VM on the destination server right. So, if you look at KVM, ZEN, VMware hypervisor use this pre copy technique right this pre copy approach for live VM migration. This our popular hypervisor like KVM, ZEN, VMware this hypervisors they use this pre copy approach. So, this is in the pictorial form if we see in the flow chart form which is given in this reference is a nice survey paper which you can refer. Migration so, start of migration so, destination server selection that where is the destination server resource reservation as the destination site is important because the VM will be there.

So, the resource should be available capture the whole memory and assume it as a dirty. So, whole memory is transferred there, iteratively copy the dirty memory pages of the VM to the destination server right. So, initially the whole is copied and then as and when incremental things are going on there. So, stop and copy if it is still not below that threshold this first that iteration goes on if it is there suspend the VM and transfer the VM state that CPU register VM memory and so and so forth that state of the VM to the destination server right. So, resume VM at the destination server and start working on the things right by the hypervisor resumes the VM at the destination server and the whole migration process is done or committed right.

So, this is nicely given this flow chart where initially that whole memory is being transferred then iteratively copying the dirty memory and this iteratively from the source to

the destinations one that is there within the above the threshold then we suspend and go on continuing with the destination server. The other one is the post copy phase in the post copy migration the processor state transfer before the memory content and then VM could be started at the destination server right. So, post copy migration technique investigate demand paging, active push and pre paging there are other techniques also for prefetching of the memory pages at the destination server. So, as we say stop phase stop the source VM and copy the CPU state to the destination restart the restart the destination phase on demand copy copies for the on demand means copy approach copy the VM memory according to the demand of the things right. So, this is the what we look at the post copy approach right.

So, if we try to look at the same thing in the flow chart format. So, the migration starts right destination server selection that we destination resource allocation at the destination site right capture VM state like CPU registered if there are any IO state and type of things transfer the VM state to the destination server. Stands for VM minimum state to the destination server and then resume the VM at the destination server right and active push dirty memory pages of the VM to the from the source to the destination server right. So, initially whole memory is pushed into the things if there is a page fault yes the copy the full fully faulty pages from the source server right if it is no all memory pages of the VM transfer successfully if yes migration needs. So, it is somewhat what we say post copying of the memory from the source to the destination server.

So, these are two approaches what we try to look at. So, the major challenge is definitely this memory thing like if and it is getting dirty pages and how many iterations goes on and things like that it is a challenge and need to be looked into that need to be looked into the whole process to be completed and committed as a whole state is transferred into the things right. So, let us do some analysis for the things. So, this is this is with reference with some of the literatures and also done one of my one ex postdoc researcher Dr. Sourav who is presently at in IIT, Swarathkal.

So, let TMIG be the total migration type right. So, the if we consider the total migration time at the TMIG let T down be the total downtime right and so, for non live migration of a single VM the migration time TMIG can be  $VM \times R$ . So, VM is the size of the memory right or the memory size and R is the transmission rate. So, what we are considering that transmitted rate remains constant throughout the migration process. So, the transmission rate throughout the migration process right.

So, that is one assumption and that may be a fair assumption because taking a something average migration rate or transmission rate and with the type of connectivity etcetera that is the assumption right. So, what we say TMIG migration time is the whole memory migration divided by R right. So, this is the typical migration time. So, as we see that the memory is the major constraint of migrating right other things is a one time things you shut down CPU and go on.

So, this is the major constraint. So, rather in non live migration practically the downtime is same as the more or less same as the migration time right. There are other considerations like CPU state transfer etcetera, but practically downtime is the is dictates dictated by the migration TMIG migration of this memory right. So, in other sense if this is a non live migration what we say that  $T_{down}$  is more or less or basically equal to the TMIG right. So, this is the what we see if we have a when we have a non live migration type of scenario. Now let  $N$  be the total number of iteration in the pre copy phrase.

Now we are considering pre copy live migration in a pre copy cycle or pre copy pre copy approach right. So, let  $N$  be the total number of iterations what we are looking for right. So, much iteration required and  $T_{ij}$  represent the total time that  $j$ th iteration iterative transmit or  $j$ th iteration transmit  $i$ th virtual machines memory. So,  $V_{tij}$  so, the  $i$ th virtual machine in the  $j$ th iteration the memory is transferred. So,  $VM$  is the memory of the  $VM$  as we have seen  $V_{th}$  is the threshold for stopping the iteration right when this is the maximum need to be transferred.

$N_{max}$  is the maximum number of iteration which is allowed there is another threshold. So,  $V_{th}$  or  $N_{max}$  is dictates the threshold of the things and  $R$  because we have this page size and the dirty page and the transmission rate. So,  $R$  basically is the factor which dictates this controls this how much in iterations and how much this dirty paging things rate is there right. So,  $P$  that is the page size into dirty page divided by transmitted rate transmission rate  $R$  is dictates that this you can think of that this sort of a factor which dictates that the rate of this dirty paging right. So,  $T_{res}$  denote the time taken to restart the  $VM$  on the destination server right.

So, it is the time taken at the to restart the  $VM$  at the destination server. So, peak copy migration mechanism the  $VM$  memory can be migrated iteratively or rather  $R$  migrated iteratively we can compute the total migration time  $T_{i mg}$  of the  $i$ th  $VM$  as follows right if there are  $i$ th  $VM$ . So,  $T_{i mg}$  is equal to summation of  $j$  from 0 to  $n$   $T_{ij} V$  equal to  $V_m$  by  $R^1$  minus  $R$  to the power  $n$  plus 1 by 1 minus  $R$  plus  $T_{res}$  right. So, where it take as we told  $T_{res}$  is the time taken to restart the  $VM$  at the destination server and  $T_{down}$  is represented by this component  $R$  to the power  $n$   $V_m$  by  $R$  plus  $T_{res}$  right. So, some just to do some calculation like in the  $T_0 V_m$  by  $R$  whatever the memory was there it is in the  $R$  where in the  $T_1$  we have the dirty paging thing.

So, based on the if that is my base time. So, it is  $p$  into  $d$  by  $R$  into  $T_0$  and rest of the calculation comes from there and similarly if  $T_2 p$  into  $d$  by  $R$  that is the factor what we calculated as a small  $r$  into  $T_1$  and  $T_3$  is like that and so and so forth. So, when we have round  $T_n$  minus 1. So,  $R$  to the power  $n$  minus 1  $V_m$  by  $R$  right. So, through this analogy and the round  $n$  where we stop the memory copying stop this phase or what we say that step 1 step 2 and copy the things. So,  $T_n$  is given represented by  $R$  to the power  $n$   $V_m$  by  $R$  right.

So, if the total time we calculate and  $V_m \times R_1 - R_n + 1$  by 1 minus  $R$  which is basically represented out here right plus the time taken for them right. So, that little there may be little simplistic or some assumptions are there first of all these rate of transmission these are constant and we are considering that few factors like within this threshold it is working and type of things right. But nevertheless it gives us the idea that how overall that how much time or how much migration time may be required things like that right. Why this is important? This is important because we need to plan for it right migration is if it is a due to maintenance etcetera then we need to system maintenance and we are migrating with a predefined things like we will migrate on so and so date at so as so time then we have to predefine that how much time we will take then it will again active  $V_m$  will be active accordingly consumer need to be initiated. If there is a certainly due to some fault or something need to be migrated that also we need to know that how much time it required for migrating the things.

So, this is important to have an estimate of the things how is to be there. Things will become little complicated when we talk about this IO and other type of things specially the network level migration all those things like where your connectivity need to be repositioned and things like that ok. So, that becomes that may become little complicated, but nevertheless it gives you a broad idea that how things works right. So, there is another things we are not going much into the thing that there is a need of estimating that number of rounds right is required for the things right considering there is a threshold for the  $V_m$  memory right and also there is a  $n_{max}$  that how much we can go right. So, we need to look at that how much memory is there which is threshold by the  $V_m$  with a  $R$  is again coming as the factor of the dirty vegging dirty veg where we calculated and based on that we can see that it is the min of because whichever is minimum that  $n_{max}$  or this calculating  $V_{th} = V_m \log R$  of looking at that factor that is number of rounds.

Again there is a there may be little cross estimation in reality there will be different other factors which we need to be considered, but not, but it is a give a ball mark things that how can be migrated. Now where there is when we migrate multiple  $V_m$ 's right. So, when we migrate need to migrate multiple  $V_m$ 's as we seen that these are we considered that  $I_{th}$   $V_m$ 's and things like that then when we migrate multiple  $V_m$ 's generally two approach. So, generally multiple because when you are going for a system maintenance it is not like that one  $V_m$  to be migrated there may be a special case there is a there may be a some resource requirement where the  $V_m$  what you are working on is not possible to run on this particular server or the host and because of the requirement rise in the requirement the  $V_m$  need to be migrated to something. So, that you have a better resources to cater to the customer.

So, there may be reasons like that, but in most of the in other cases where you go for a system maintenance or you are expecting or detected a fault and want to recover the transfer this  $V_m$  to another thing. So, it is not one  $V_m$ 's one  $V_m$  rather is a multiple  $V_m$ 's

which need to be transferred to the destination host. Now typical strategies as intuitively come as one is going serial serial right one  $V_m$  then next  $V_m$  next  $V_m$  and type of things that is something one at a time which may be if we consider that looking at the other our calculation that a single  $V_m$  multiple times right it can be there or there can be a parallel migration. So, I migrate the whole thing in a parallel and it goes in a migrating way.

So, when you look at the serial migration. So, the first  $V_m$  that is selected to be migrated executes this pre copy cycle and other  $V_m$  minus 1  $V_m$ 's continue to provide services that is there right. So, I have say  $m V_m$ 's and then the first  $V_m$  gets migrated using this say pre copy cycle and type of things and then the rest of the  $V_m$ 's which are in minus 1 they continue give services as soon as the first  $V_m$  enters into the stop and copy phase the remaining  $m - V_m$ 's are suspended and copied to the first  $V_m$  completes the stop and copy phase until it copy this. So, the once it is p and copy the other are suspended and then that executes this first  $V_m$  executes its completes its copy phase. Reason for stopping the remain  $m - 1 V_m$  is to stop those  $V_m$ 's from darting the memory right otherwise what will happen during this portion what if it is a single bin what is expecting that nobody is darting this memory right. Once it is stop and copy phase we expect that nobody no one is darting this memory because there is a single  $V_m$  which was darting and now when say stop that thing it is not darting the memory right, but in this case it may so happen the other  $V_m$ 's may darting the memory right.

So, one assumption that is  $V_m$  that is copied at full transmission rate right. So, the full transmission rate is available as we have seen that R. The downtime for serial migration includes the stop copy phase of the first  $V_m$  the migration time for the  $m - V_m$  and time to resume the  $V_m$ 's at the destination  $V_m$ . So, that is those type of things are considered if we consider that this type of serial thing.

So, one when we look at the parallel migration. So, the major difference what we can see that all  $m V_m$ 's start their pre copy cycle simultaneously it is a big challenge right. So, all this  $m V_m$  because it is parallel. So, each  $V_m$  effectively we can consider get a rate of  $R - m$  for the transmission capacity right or actually there will be few more overheads. So, it is effectively less, but we can consider that  $R - m$  type of bandwidth they are getting or transmission capacity they are getting to transmit the that  $V_m$  to  $V_m$  1 to  $V_m$   $m$  to the destination right. So, here all we are considering that is transferred from one host or one server to another server on the other host or the server right.

So, as the  $V_m$  sizes as the  $V_m$  sizes are same and transmission rates are same. So, that is another consideration same the  $V_m$  begins to stop copy phase more or less at the same time and they end the copy phase also at the same time. So, it is little maybe I should not say strong assumption, but the a some we also assume that all  $V_m$ 's are of the same size. So, ideally they are having the same transmission rate and more or less though they are serving different things. So, they are other iterations stop copy phase will be at the same time right.

Since the stop copy phase is executed in parallel they consume the same amount of time at the downtime is in fact, equivalent to the time taken by stop copy phase for any V m added to the time taken to the resume the V m at the things right. So, as the if we look at the real downtime what it is happening is basically that because so long this memory is being copying the V m is active at the source end right. So, when we have a stop copy phase that time the downtime is being activated right. So, as we are considering that those are all in the same time. So, it is being the downtime in addition that the time it take to resume the operation at the destination V m right.

So, in case of a parallel things, but nevertheless the effective transmission rate it gots a  $H_V$  m is something which is  $R_m$  right effective transmission rate. So, that is that is the another that is that need to be considered. So, if we look at today's discussion. So, what we try to do? So, do some means a something a broad analysis or try to look at that what will be a some estimation of this our sort of a this migration time and downtime sort of things and also as that estimation of this how much iterations required based on the this dirty paging phenomena. And also we try to see or have a overview of this what is when we when we have multiple V m's more than one V m's into the things whether this parallel migration or vis-a-vis a serial migration how things are carried out right.

So, with this let us stop our discussion for today's today's discussion and we will continue with other aspects of this our this cloud computing in our subsequent classes. Thank you.

**Cloud Computing**  
**Prof. Soumya Kanti Ghosh**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**

## Lecture 47

### **Containers Container based Virtualization Kubernetes Docker Container**

Hello. So, we will continue our discussion on Cloud Computing. Today we will be talking on or discussing on container or container services or containerization services which are referred. This is an important component and if we see in today's world or today's IT world, the overall whole a major majority of the things is moving towards this containerization services right. There are lot of advantages as well as challenges in this sort of mechanisms. And so, we will today we will have a overview of the thing and then in the subsequent lecture rather we are planning to have a show you a demo of how container can be deployed to achieve a particular goal and how things works actually right using open source services right.

So, not today, but in this series of 2-3 lectures one small demo we will show you right. So, this is and when we talk about container which comes into our mind is the this big container which are being carried by this trailer or the trucks and which are being shipped over from one place to another put on the ship right and taken to one place to another right. The beauty or the overall mechanism of this container you we will come to the in context of virtualization is that when you look at it, it looks as a box right. It is a box which has a definite size and shape and inside whatever is there it is some sort of a self content right.

So, it is carried to the other place and it can be opened or explored at the other place and work on the things right. So, what is what it matters when you transfer this container is that you need to take care that it should fit into the things whether it does not matter whether you are transferring food grains or say four wheeler vehicles, two wheeler vehicles or anything, but so long it is fitting into the things. Definitely the size of the container is made such a way that it fits over this large trucks, it fits over it sits in the carrier of the this ships or it can be it has the crane mechanisms which can hold up to the type of container it shows on. Definitely there are other constraint like weight and etcetera, but nevertheless this type of container based things what is achieved that a large scale of interoperability. You do not bother that what is there what we are expecting that whatever is there inside is self content we do not have to give anything extra energy transfer.

So, anyway that is the thing which we are having in our common things and we try to see that how things works in this context right. So, what we will see today is more of a container and that virtualization and container and things like that a overview of popular open source implementation like Kubernetes and Docker services right. So, that we will try

to look at when we see this type of services. So, basically container and the if you look at this virtualization these are the keywords. So, if we just have a brief overview or introduction to the things.

So, what we have seen this virtualization helps in sharing resources. So, initially what we are having that server based systems where you are running more than one applications right and then the resources are shared and if a application is a resource hungry other may be starved of the resources and those type of scenario from there what we seen that this virtualization where we still have that infrastructure back end infrastructure it may be a one server or a cluster and then we have different virtual machine virtual machine may have their own OS right they are having their own OS over their things runs over the things. So, virtualization what it helped in share resources among many customer of cloud computing right. So, that is one of the things what we already seen right. So, what when you talk about container it is a lightweight virtualization technique.

So, it is basically a virtualization technique. So, there are similarity in that context and it is a lightweight technique right unlike our virtualization it takes a spell or the overall footprint is much lighter right. So, container packages what it tries to do packages the code right whatever the code you are writing and all is dependency. So, the application runs quickly and reliably from one computing environment to other right. So, it is it packages everything right what you are talking about when at the beginning of this lecture that when we talk about this what we say container in our real life that physical container right what we see you running over the highways etcetera by large trailer and trucks.

So, what we see that is that is everything is packaged inside right all whatever is required it is packaged inside. So, whether you are taking it to a truck or ship or in some other things it all it matters that overall it is safe size and weight right other things are self content. So, here also a container packages all codes and all is dependency. So, the application run quickly and reliably from one computing environment to another. So, it does not matter whether it is running in Linux based system you are putting into a say windows based system to some other flavor of Linux or Mac etcetera ideally it should not matter right.

So, it is self content type of things. So, two popular implementation or popular open source things what we are having are heavily used one is the Docker is an open platform for developing shipping and running applications and a Kubernetes where is an open source system for automating deployment, scaling, management of containerized application right. So, these are two popular things rather Kubernetes evolved from means Google which became open source exactly I do not remember somewhere in 2014 or so. So, it is now a open source things right. So, containers are packaged a packages of software that contain all the necessary elements to run in any.

So, we are repeating just to in build in our head that it is a self content thing right all dependencies etcetera there. So, it virtualizes the operating system and run anywhere right

practically anywhere from they say that from your own personal PC or laptop to cloud it is it can run anywhere right. From a private data center to a public cloud or event developers personal laptop it runs in the means you do not have to do anything everything is in means what we say encompasses in this. So, containers are lightweight packages of the application code together with dependencies such as specific versions of programming language run times and libraries required to run the software services right. So, to be more in a formal way.

So, these are lightweight packages of the application code. So, whatever the application you are trying to run together with the dependency whatever dependency such as a specific version of the programming language run times their libraries etcetera required to run that software services right. So, that is that is the container it makes it easy to say as CPU storage and network resources at the operating system level and offer a logical packaging mechanism in which applications can be abstracted from the environment which they actually run right. So, what we are trying? So, what we are trying to do live virtualization it is basically a lightweight virtualization approach only. So, where like virtualization it now it is easy it is say our CPU memory storage in a more efficient way and doing at the operating system level and something logical packaging mechanisms.

So, it can be from one environment we in which they actually run can be put into the things. So, as we as we just told offer a logical packaging mechanism in which application can be abstracted from the environment in which they run actually right. So, you develop somewhere and how things are there you package it as a logical packaging in the things and it can run in that same fashion. So, this so, what we in effect doing this we are decoupling the things right. So, running of the things and where it need to run this is now can be maintained.

So, what it helps container based application to be deployed very easily right because you do not have to break your hand in reconfiguring etcetera and consistently the way it is running in some in one deployment other deployment consistently regardless of whether the target environment is your a laptop or PC or cloud etcetera. So, overall what it the basic it supports this agile development that is one of the need of the things right. We require a right development from the developer point of view containers allow developers to move much more quickly by avoiding concern about the dependency on environments right. So, it is a possible to have agile development efficient operations containers are lightweight and allow to use just the computing resources one need this running of the application thus it basically run the application in a very efficient way and run anywhere as we tell that it is a basically along with the your application all this dependencies etcetera there it virtually runs anywhere. So, whatever platform you are putting so, it supports this container it supports your the application.

And of course, in we can see that from the previous discussion there are several benefits right one is separation of responsibilities right. So, we can quantization provides clear

separation of risk management. So, it is so, developers focus on application logic and dependencies IT operation teams focus on deployment management instead of the application details etcetera etcetera. So, the developer is more concerned about the implementation of the logic whereas, the this other team or the core IT team or deployment management team they are more concerned about this the deployment management and then it is they are not bothered because once it is containerized they know that things are going to be in this fashion right. So, what we are if we again come go back to our analogy when we are transporting something over a fixed container over the truck etcetera.

The other end that the shipping company is not actually bothered about what is this coming it understands that whatever is there it is self-contained and a box of this size and this weight is coming right of say this size and makes this weight is coming and we need need to fit it into somewhere right. So, it is so, it is they can work on their domain separately right. Workload portability another definite another major benefit that container can run virtually anywhere greatly easing development and deployment right it can be on Linux or different variants of Linux it can be on Mac it can be on Windows. So, it either on a VM or a physical machine or a developers machine data center on premise and public cloud etcetera. So, it is a enormous portability it keeps which is one of the major challenge in today's world or rather major challenge in IT world for from the beginning itself right.

Like if I if I want to transfer a application or a bunch of code from one to another then there are serious challenge of portability. So, it tries to address that issue so, highly portable and it becomes very what we can say that becomes very extremely useful and when we deploy something at the you want to deploy something at the customer end and where you do not have that expertise of make this tools and packages you need to reconfigure at your end the system etcetera right. So, if you have this sort of a container based services then you do not bother about this you just only you only bother to how to house this container and if that is standard and then you do not bother. So, once if it is housed things will run right. Application isolation container virtualized CPU, memory and network resources at the operating system level as we discussed right few minutes ago.

Providing developers with a view of ways logically isolated from the other applications right. So, it gives some sort of a whatever is here it is basically having a isolated from other things right. So, it is something which is they are on the for its own purpose only right. Again not a very may not be very good analogy in this case, but if I think that my container I want to ship something say I have a car company and to ship the cars. So, I only bothered about that how it fits into the things right.

So, it is all my things I do not care that whether it is separately transferred, how it is transferred, whether to or how many modes of things are transferred, what I am bothered about that how my things will fit this thing and how means what are the way I need to deploy that in inside that container right. So, what we see separation of responsibilities, workload portability and application isolation these are the major things which are serious

challenge right in any IT industry or deployment of the things. So, if we look at the three pictures so to say. So, this is our traditional deployment still there. So, or maybe slowly we are moving to the more of a virtualized world right.

So, this virtualized so in the traditional we have hardware operating system and apps over it right. In a virtualized deployment we have that infrastructure or the hardware which can be cluster of systems over that the operating system, then the hypervisor or the VMM virtual machine monitor and then we have different virtual machines and then run their guest OS on these virtual machines and run applications. So, initially the hardware which are not shared among the resources now it is properly shared right here also the application share this resources share the hardware resources in traditional, but there is no boundary between them appropriate boundary. So, in that case what will happen that some require more resources may resource hungry application may eat away some of the things which are other disturbing etcetera. There are issues of isolation and there are challenges of security and things like that those things are there.

And when we come to the container deployment then we have hardware operating system and the container runtime over that we have different containers right. So, they are self content they whatever the application etcetera is there. So, they are self content and run on this type of things right. So, this three thing so just to what we discussed this traditional deployment applications run on physical servers there was no way to define resource boundary for applications in a physical server like this traditional server. And then this cause resource allocation issue right as we are telling that some may be eating away more resources than the others and things like that.

Then the next thing is the virtualized deployment allows to run multiple virtual machine on a single physical infrastructure or the physical say machine right one or more things. So, physical servers CPU virtualization allow application to be isolated between the VMs because there are isolations. It allows better utilization of the resource in a physical server and allows better scalability right. So, what one is that application isolation and I can scale the things like I require more VM resources then I can migrate and things like that I can do different scalability. Each VM is a full system right required for the customer or the running application running all components including its own operating system on top of virtualized hardware and etcetera etcetera right.

Then next the next gene sort of thing what we are seeing today or what it is this IT world is going towards is the container deployment right. Container simulate to VMs in case of virtualization things like that, but they have relaxed isolation property to share the operating system among the applications right that is another part. So, that therefore, containers are considered lightweight containers has its own file system share of CPU memory space and more. The containers are decoupled from the underlining infrastructure they are portable across different cloud service providers or different OS distribution and things like that because as they are decoupled from the underlining infrastructure.

So, they do not have to be there. So, this things portability is much higher right than VM and rather in case of a traditional portability is a serious challenge. So, now as this containers and VMs are going hand in hand. So, instead of talking telling that containers versus VM let us put it as a container and VM like. So, VMs a guest operating system such as Linux and Windows run on the top of a host operating system. So, as we have seen in the previous picture.

So, there is a host operating system and a hypervisor and there are different guest operating system which runs over that. So, underlining host hardware right containers are often compared to virtual machines obviously, because both are virtualization techniques like virtual machine containers allow one to package of the applications together. So, it allows us to package these applications together with other libraries and dependency providing isolated environment for running the software services. So, what it allows? It allows the package the application together with the libraries and other dependency. So, it is not it is self content right and thus providing isolated environment for running the software services.

So, however, what we see? So, rather in a sense this is also somewhat is also true for other VMs, but in case of container offer a far more lightweight unit for developers and IT ops team operational team to work with carry a myriad of benefits and etcetera. So, it is practically if we will compare with the VM one of the thing is that it is extremely lightweight. So, what if we see containers are much more lightweight than VMs that is one thing. Containers virtualize at the OS level while VM virtualize at the hardware level right usually VM virtual means a VM virtualize a hardware level where container virtualize at the OS level. Container shares the OS kernel and use fraction of the memory required right.

So, it use the or exploit that OS kernel and use a fraction of the memory VMs required so in comparison to the VM. So, that is what we see it is a lightweight virtualization technique what we looked at. So, now let us look at two popular overview we will little look deep into the things in the subsequent lectures. One is Kubernetes right. Kubernetes is a portable, extensible, open source platform for managing containerized workload of services that facilitate both declarative configuration and automation.

So, one sentence tells a lot of things right. So, it is portable, extensible, open source platform for managing containerized workloads and services. So, it helps it basically able to manage containerized services that facilitates both declarative configuration and automation. So, it can facilitate your declarative configurations of the things and automation. So, it has large it has a large rapidly growing ecosystem so to say. So, it is a big thing what is the contraction is going on.

So, Kubernetes services support tools are widely available. So, that is widely popular and it is open source. So, what it says that the literature says that Kubernetes originates from a

Greek from Greek meaning some helmsman or pilot right anyway that is the why the name like that. Kubernetes operate at the container level rather than at the hardware level. So, it provides some application features common to past offerings right like platform as a service right where it is higher than the highest type of things.

So, some sort of a patch of a rings are as deployment scaling auto balancing and let us user integrate their logging monitoring and alerting solutions right. So, it is not at the hardware level, but can be thought of at a something as a pass level. However, Kubernetes is not monolithic. So, that these default solution are optional and pluggable.

So, it is not a monolithic structure. So, these are optional and things. Kubernetes provides the building block for building developer platform, but preserves user choice and flexibility where it is important. That means, it boasts provide building blocks to for the developers, but also give a choice flexibilities of the user things right. So, it is a what we say it is a more versatile service which is much lightweight and have their building blocks and also give. So, we will see little bit of Kubernetes in subsequent lectures.

So, if we look at the Kubernetes components. So, when we talk about Kubernetes installation it provides you a cluster right. So, Kubernetes cluster consists of a set of worker machines called popularly known as nodes that run containerized application. So, these are the nodes which run every cluster has at least one worker node means there should be any cluster should have only at least one worker node. The worker node or nodes host the ports that are components of the application worker workload right. So, this worker nodes basically host the ports which are basically application workload components right.

There is a control plane manages the worker nodes and the ports in the cluster. So, there is a cluster where the worker nodes are there which worker nodes host the ports which are the components of the application things and overall things are managed by this control plane which both the worker nodes host and the ports. So, in the production environment the control plane usually run across multiple computers and a cluster and a cluster usually run multiple nodes right providing fault tolerant and high availability. So, it is typically in a production environment what we say that the control plane runs over multiple computers and a cluster usually run multiple nodes which host different ports and it provides a fault tolerant and high availability. So, this is a big picture of the kubernetes where this control plane things are there and there are different nodes and if you look at there are different other components like scheduler, kubeproxy, kubelate and etcetera we will discuss those components in subsequent class.

Another popular open store platform is a docker or sometimes called docker container. So, name is something which is which can be docked and docking type of things. So, what we have seen docker gives a sandbox process that is isolated from all processes on the host machine right. So, it is a mechanism by which is isolated from the other processes container is a runtime instant of an image right one can create start stop move delete a container

using things like docker API and cl or cli and type of things. The container can run can be run on local machine virtual machine clouds and anywhere right that we have already seen.

So, docker again we will discuss a like kubernetes with the docker also little bit more. So, for to just to give you an overview the docker container image is a lightweight standalone executable package right software that includes everything needed to run that things code runtime system tools dependencies libraries settings etcetera everything right. So, this is very popular and becomes very handy for easy deployment and things like that. So, container images become containers at runtime right. So, we have container images which becomes container at runtime and in case of docker container image become containers when they are run on the docker engine.

So, there is a concept of docker engine over which these things runs right. So, available for both Linux and Windows application containerized software will always run the same regardless of the infrastructure right. So, that is the basic things what we started with right regardless of the infrastructure it can run anywhere and in the same fashion. Containers isolated software from its environment and ensure it works uniformly despite the differences in the operating systems and different instances of the development and things like that.

And if we look at that this docker engine. So, run on docker this container run on the docker engine which is standard docker created the industry standard and the containers they could be portable any anywhere right it is lightweight and secure applications are safer in the containers and dockers provide the strongest default isolation capabilities in the industry, but the docker they claim that they are a thing. Nevertheless it both this containers try to provide this isolation application level isolations which helps in different what we say securing the application from other applications ok. So, with this what we see in today's class discussion is that what is a overview of the container. We rather few lectures or sometime in the previous discussion we discussed about dockers and containers, but this time we in next couple of discussions we will be having little going little deep into the things and actually want to show you actual deployment of this sort of a container services and how things works in this type of environment right. So, with this let us conclude our things there are few references like Kubernetes and docker and some from the Google cloud things.

So, you can refer those documents and things like that. So, let us conclude here for our this basics of or overview of the container or container based virtualization. Thank you.

**Cloud Computing**  
**Prof. Soumya Kanti Ghosh**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**  
**Lecture 48**  
**Docker Container â€“ Overview Docker â€“ Components Docker â€“ Architecture**

Hello. So, let us continue our discussion on computing more specifically on container based virtualization rather we will be looking into one aspect, one very popular service that is the Docker right which is open source and widely used and it will help us in understanding the basic say philosophy, basic advantages and basic architecture of a typical container based services. Not only that in our sub-sick lectures we will also have some demo using this Docker to show it how applicable it is. Though we are looking with the Docker, but it is generally true for any type of containers right. So, we have this Docker container which will be discussing and keywords this as we have seen that as we looking at the Docker container it will be Docker container and architecture component, basic architecture advantages and so called benefits and key components of Docker component container what we will be covering. So, this is a very popular picture if you look for Docker which also implicates right like yesterday last class we were discussing regarding this containerization service where we see that it is quantanize all things are which are self content or with all it dependency a container which can be shipped from one to another like it can run to your laptop to server to cloud anywhere.

So, it is a self content application. So, which has the that application running along with dependencies and other things right. So, it is more like a as we discussed that is shipping through a some material in a container. So, the whatever you put in the container should be a self content.

So, it is from the source to the destination it will go and there it will act in between whatever the way it is shipping or how it is managed etcetera, it is based on the how the overall container thing. So, it is not only created a revolution in our overall transportation systems of goods and things that making a standardized approach it also what we see is making a revolution in the software world. So, today we will be more concentrating with Docker a open source platform. So, it is a platform that allows to build, ship, run any app anywhere very straight forward and direct thing. So, it is a platform which will allow you to build, ship and run any app any application anywhere.

So, it irrespective dependency of the ways etcetera. So, you have this type of Docker container. So, it is considered to be a standard way to solving one of the challenging aspects of the software not only development, sorry deployment, development, deployment and also we will see that it helps in your testing and things like that right. So, it is this is containers attempts to address this problem. Additionally, traditionally what we see that any software development process is a pipeline typically involved combination of various

technologies for managing the movement of the software such as virtual machines or VMs we have seen configuring this management tool right, package management, complex web of libraries, dependencies etcetera right.

So, this is those who are in the software industry or those who are working with the software and seen or even looked into our traditional software engineering approaches and documents you see that these are the things which come into play when we go for the development. So, all these tool needs to be managed and maintained by specialist right, specialist or trained personnel or what we say trained engineers many of them or most of them having their own way of approaching the problem right. So, what it makes more difficult to interoperate right. So, when you ship one to another suppose you develop something in a variant of Linux say Ubuntu and put it on red hat or sent ways or something and then it creates a problem you need to re engineer the whole thing not the whole thing you need to put your hand on the things to see that how it need to be customized on that environment right. This is a big hindrance in deployment right in especially large case deployment where you expect that the other end you cannot expect that the other end to be versatile enough to handle all those things right.

And especially in today's world when this IT service glorification is in a big way starting from our banking to everything what we talk about in our life is now in this is going in that way and that is serious challenge right. If you need something to be configured at your end for getting that service it is a challenge right. So, this docker try to address this things rather container based services tries services try to address this challenges right. So, docker or containers has changed the traditional approach. So, everything goes through a common pipeline to a single output that can be used on any target there is no need to continue maintaining something that array of configuration mechanisms right.

So, it is a thing. So, in another sense if I have a application self content say I have a database of say customize things say MySQL and so like that and I develop out here and give it to you to run on based on your other application like it talks to your web server or so. So, it should work right I do not have to go and look at the things that how MySQL to be configured etcetera. So, MySQL may be a more what we say more versatile or more widely used. So, it may be configure as an may be easy, but there are if there are customized things.

So, it may be a challenge to do that right. So, at the same time there is no need to throw away the existing soft stack if it works for you right that is that is another important thing something working for you for a soft stack for a good amount of time and you do not want to throw that right you can package it up in a docker container as is for others to consume right. So, this is another way of looking at it when we make a containerized services right. So, this is another important thing. So, something a soft stack which is working fine for you and not only that in some of the cases it may gone through some testing and some certification or some sort of a standardization process and then you do not want to throw it out and you want to make it as a containerized so that you can be delivered to other to use

right.

So, additionally you can see how these containers are built. So, if you need to dig into the details you can right. So, in other sense that additionally if you want to see that how this overall container is built you can look into the things right ok. So, this is the big picture. So, before going further let me refer to this book this is a very nice book it is good to read that Docker in Practice by Ian Mill and Aidan Hopson.

So, this is a nice book. So, rather most of the things what we are discussing are referring to this book. So, I kept that reference here, but I encourage you to have a look in this book. So, it is a good to have a overall feel of the things very nicely written. So, what we see that before Docker.

So, we have different tools right. So, like if you see that the top one where this is the inputs this green is the input to the system requiring manual maintenance fewer inputs are here mean less in the maintenance burden right. So, these are the inputs and that these are the tools this blue colored that use the input to create environments for the software development right. So, these are the tools and stages of software development requiring the environment to run in right. So, this is development, test, life type of things and here also testing release to life.

But what we see that overall this with the Docker we can reduce this middle time in big way rather what we can see that some of the configuration can be standardized which may not be so much thing right. So, it becomes quote unquote life little easier for software specifically for software deployment where for using Docker or other any containerized services. So, one analogy or which it comes from like if you look at the Docker is typically a person or a so called quote unquote laborer who move the commercial goods in and out of the ship when they are docked at the ports right. So, it is docked at the port and the Docker. So, these boxes are items of different size, shapes and experience Dockers actually make that fit the goods in a appropriate way right cost effective way in the sense packaging is time right.

So, at it is first of all costly in getting those people and making these things are challenging right. Now, though we do not have that pretty early traditional things, but in something shipping what we are doing we are doing more using cranes and etcetera. Nevertheless, if there are variable number of sizes of box making then appropriate things forget about ship if I if you have a variable number of say gift boxes even say proper means regular shape rectangular type of things. Then also if I if you want to put it on a big box it is always a challenging because you need to optimize that thing in a big way. So, this is something what similar what we do in our software development.

So, much time and intellectual energy is spent getting the metamorphic this odd shape software. So, into different size ships or carriers full of other odd shape software etcetera

etcetera right. So, how to metamorph this type of things into this odd shape and size software into the thing. So, it is somewhat similar. So, that may be a reason where the name came from.

So, if we look at the benefits or why Docker so to say. So, before Docker deploying software or to different environments required significant efforts right. Even if you are not hand running scripts or provisioning software to different machine you would have to handle the configuration management. Like even if you are not doing the software development, but at least the configuration management is also a challenging. You might have seen that if you have a working in organization or like even your institute.

So, once you put one to put a software by a software and put on the things and then they are expert engineers comes and configured based on your available resources right. So, and it gets a configured on the things. It is not may not be very true for your normal software when we download our laptop and to so I can still if you want to do a customized configuration right other than your default configuration then you have to need to think right that what is your requirement, fitting to your requirement and the software things is again a challenge. So, those challenges are there too. So, even this things is little eased out when we went for VM type of solution like virtual machines and things like that.

So, we try to ease out this type of challenges, but even with this encapsulation in VMs a lot of time was spent managing that deployment these VMs waiting for to boot, managing the overhead resources they use and things like that right. So, though the VMs tries to ease out this make some sort of a standardized way of approaching the things, but still there are challenges in handling this steps. So, again referring to this particular book if we see. So, there are if we look at the life or not before it is without a traditional way of looking at it this is very much existing out here now. So, there are typically three step to manage the deployment right.

One is that install configure and maintain complex application right and then install configure maintenance complex application in test server and install configure maintain application is live server right. So, if you want to manage it is like this these are the things which you want to do whereas, in docker ideally there may be a single effort right install manage maintain complex application in docker image. So, if you can make this docker image ready or things what you find in a lot of standardized image docker images in the docker hub those who are using or you can look at the docker hub. So, you can docker run on development laptop test server and live server or what you have the production server type of things right. So, it makes definitely life easier for the specially for the deployment engineers who break their hand on lot of time on this type of things.

Docker benefits with docker the configuration effort is separated from the resource management right and the deployment effort is trivial right. So, docker is separated from the resource management. So, it is something like run your docker the environment image

is pulled down ready to run right consuming fewer resources content and content. So, that it does not interfere with the other environments right. Like I was telling that if I have docked into a MySQL into a docker and put it into the things with a properly configured then there it goes on running on the things.

So, what we require is that the docker engine out here or the where the platform any other containerized platform which can house this docker right. So, rest of the things are need not to be looked into other dependencies and like this right. So, there is a big advantage you do not need to worry about whether your container is going to be shipped into say a Red Hat machine or Ubuntu or centways as long as the docker on it right. So, I do not much worry about the where it is going to house or the what is what we want to say that it is whether it is any flavor of this ways we if that docker as long as the docker is running on it will run on the things right. Now we like to have some of the key advantages or of the things or key reasons one is replacing I should say not only advantages key features may be the appropriate word.

So, replacing virtual machine docker can be used to replace VMs in many situation right. If you do not if you only care about the application not the operating system docker can replace the VM right where because in these days this with wide scale deployment of the of this IT services this applications are becoming major thing. So, where docker may help not only in docker quicker than VM to start up it is more lightweight that we have discussed in the last class also the lightweight to move around right. So, if it is lightweight so, shipping to one to other is much easier and due to its layered file system right. So, it is approach a layered approach you can more easily and quickly say changes with others it is also rooted in the command line in and scriptable all it is possible to script it.

So, it is scriptable. So, prototype typing software if you want a quickly experiment with software without either disrupting your existing setup or going through the hassle of provisioning a VM etcetera. So, it is a docker can give you a sandbox environment right to do a almost instantly right. So, it gives you a sandbox environment almost instantly where you can prototype your software for other than now otherwise the before prototyping the setting up of the whole environment takes a lot of time which can be effectively saved when you use the docker platform. Other features are is that packaging software because docker image has efficiently effectively no dependency right. So, a docker image do not have any dependency it is great way to package software you can build your image and be sure it can run on any modern Linux things right.

So, anything it can run on the so, packaging is much easier. Enabling microservice architecture docker facilitates the decomposition of complex system to a series of composable parts right. So, that is a thing right. So, where it you can decompose your complex system in a in a set of which can be composed to into this making this this complex system enabled right which allow you to reason about your services in more discrete way that is this allow you to restructure your software to make the parts manageable and

pluggable and type of things right. So, in other sense if you have a complex system dock dockers do facilitate make them into manageable parts which can be separately managed, but once it packaged it goes as the things right.

So, which means ideally it enables a microservice architecture. Modeling network it is so, once you have say several hundreds or more of isolated container can be initiated in one message modeling network can be done efficiently. So, that also it is helpful. So, modeling is network enabling full stack productivity when offline right. All parts of the system can be bundled into docker container you can orchestrate this to run in your laptop and work on move even when the thing.

So, in other sense on offline mode also if all the things are there you can basically bundle it and run work on the work on move right or this is has different type of advantages one is that this offline mode is a full stack productive environment otherwise testing something and doing something on a offline mode. Also helps in reducing debugging over it right complex negotiation between the different teams about software delivered in a common place within the is a common scenario in any industry right. So, docker allows you to state clearly even in script form the steps for debugging a problem on a system and known with known properties right. So, making bug and environmental products in much simpler etcetera. So, it facilitates this debugging things and thus reducing the overall overhead cost of debugging right.

You can clearly specify the steps of for debugging a problem and things like that. And also overall this particular platform or the approach is helps you in documenting software dependencies right. By building your images in a structured way ready to be moved at different environment docker quote unquote forces you to document your software dependencies explicitly from this from the base starting point right. So, as it need to be can be portable to anywhere and everywhere as we start our discussion with that statement. So, in a sense this docker allows you to or not allow is somewhat force you to document your software dependency explicitly from the means starting point that is the step 0 type of things.

And also it enable continuous delivery that is one important feature for any software thing that you want we you may like to have a enabled a software continuous delivery is a paradigm what we say CD sometimes referred to as CD is a paradigm of software delivery based on a pipeline that builds the system on every change and then delivers to the production on life through an automated or partly automated process. So, that means, it is a this CD is a paradigm where approach for where based on a pipeline that rebuilds the system on every change right. So, I have a I have some change request change the thing rebuild the systems and then we will delivers to the production for the life system right and it may be automated or partially automated. So, this is a challenge and this is a feature to be addressed in a software deployment which dockers efficiently supports or it enable this CD right. So, docker builds more producible and replicable than traditional software building

methods right.

So, these are or more so, docker if we look at the docker builds or docker itself is more producible and replicable right that is the basic philosophy we started with right if you core philosophy is that build ship and deploy anywhere. So, it is something which is it can be reproducible replicable and then our traditional building methods. So, this makes implementation continuous delivery much easier right. So, this is the if we look at the key concepts of the docker.

So, if we see there is a layered thing. So, this is the things which are on the host meeting is stored that say this particular the Debian layer or the Linux variant layer then my application code layer may be there is one application say v1 and there is a application v2 layer right. So, these are the two layers what we are having. So, this images and image is an collection of file system layers and more and some meta data right taken together they can be spun into a docker container right. And this layer is a collection of changes to files right collection of changes to files the difference of between v1 and v2 of the my application are stored in this layer right. So, I can have different type of say like my application version 1, my application version 2 and type of things which are in that those layers.

Now this is on the your stored on the disk right. Fine running the process I basically instantiate this right. So, that the container is a running instance right of an image. You can have multiple containers running the same image right. The same image can be run on the multiple container they may have my application container, my application container version 1, 2, my application container v1, 3 and these are the multiple applications. And you have these my application that there may be again one or more of v2 right.

So, the other version. So, what we see that you basically create that image and then deploy it as and when it is required right. So, if you have a new version of the things with some more applicability and changes etcetera you create a v3 and which can be deployed into the things right. So, these are the things as it says that it looks for a registry or a namespace where things will be maintained that how do you know that where things are there those are the things are there. And if we look at the key commands. So, one is docker build, build a docker image, docker run, run a docker image in a as a container, docker commit, commit a docker container as a image right.

So, and docker tag, tag a docker image. So, these are the basic commands those who are working on the things some of the things will be showing you during the demo that how things works right. So, that will be that you can see in the in our subsequent lectures at least one or two we will see show you the demo. So, basically having rather what we are talking about is basically somewhere encroached into the architectural thing, but more formally telling. So, if you have a docker on your host machine is split into two parts right. So, this is the traditionally right one is the daemon with a RESTful API and a client that talks to the daemon right.

So, we have a daemon with a RESTful API and a client with talks to the daemon. So, a private docker registry is a service that stores the docker image right. So, I was telling that how do you find like you need to be this need to be a catalog or registry. This can be requested from any docker daemon and has the relevant accesses definitely you have to have the registry is on the internal network and is in publicly accessible.

So, it is considered to be private right. So, that what are the different variations and other docker images are there that is there in the registry. So, one we see one invokes the docker client to get the information from or give instruction to the daemon right. The daemon is a server that receives request and returns the response to the client using the HTTP protocol right. So, or piggy back on the HTTP protocol. In turn it will request to other services to send and service images also using the HTTP protocol.

So, our basic underlining what we say that transport protocol you should not confuse with the transport layer of the network stack. So, the protocol which is piggy back here is the HTTP. The server will accept request from the command line client or anyone else authorized to connect to the things right. The daemon is also responsible for taking care of your images and containers behind the scenes whereas, the client acts as the intermediary between you and the RESTful API right or between the your application and the RESTful API right. So, this is the overall working how it works and if we look at the same thing in a picture which is there in that reference book.

So, what we have that here your the host machine on which the docker is installed this is the host machine. The host machine will typically sit on the private network right. So, you invoke the docker client program to get the information from an instruction to the docker daemon. So, the docker daemon receive request and returns the responses to the docker client using the thing. So, the docker client is there HTTP request to the docker daemon and it is a request respond type of things right over HTTP.

The private docker registry stores the docker images. So, it consult the docker images and the docker hub is public registry run by the docker in commercial as we are talking about docker hub you can on the internet these are the docker hubs which can also be interacted with the things right. Other public registry can also exist. So, there can be other type of public registry. So, what we what we is dictating here that one is that how your docker are built, but where this dockers how the metadata and other connective information these are in the registry. One is your within your thing that whatever the docker you are managing other can be the by the which is maintained by the docker that is the docker hub or you can have other public docker registry from where you can pull the things right.

So, this is the broad picture though a broad picture, but it gives a overall idea that how things work on the in overall mechanisms right. So, today's class what we try to discuss that more little more on the docker. So, that and any other containerized services are more or

less in the same philosophy their way of working may be different, but their philosophy is same. So, what we try to see that what are the key feature or advantages using docker and what is the broad way how things works right. So, in our subsequent lecture or may be lectures we will see a some demo very small customized demo, but try to see that how docker what we have conceptualized how it works in how it can be made to work in a say very in a lab type of scenario though synthetic, but in a real life scenario ok.

So, this is the references. So, the one is that docker main site other is very nice book I encourage all of you can go through this book docker in practice specially if you want to work on docker this is a good thing to start with right. Thank you.

## **Cloud Computing**

**Prof. Soumya Kanti Ghosh**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**

### **Lecture 49**

#### **Docker Container - Demo**

Hello. So, we will continue our discussion on cloud computing rather if you recollect we were discussing about container or container based services and today we will have a demo on this a particular container like docker container right. So, that you can have a feel that how container works and also we will try to make it in such a way that you yourself can try out at your end. So, that will give a what we can what you can say some feel of that how whole things works right. So, let me introduce I have one side Mr. Subobrata Anand and Mr.

Anupam Khan from our CS department. So, they will be primarily running that demo showing the demo. So, before handing over the control Mr. Subobrata and Mr.

Anupam. So, let me have a couple of a preview of the whole thing. So, that it will be easier to follow the demo. So, it is primarily what we will be looking at is installation of the docker engine and running this docker container images. So, if you remember so, what in while discussing with container specially for the docker and if you remember that that basic architecture when we discussed.

So, it was a it was the docker for the running a docker image you need a docker engine which is which need to be installed on your system. So, if it is installed so, ideally you do not have to do anything when you have any docker image in the things right we will tell you what the things. So, typical keywords as usual the docker, docker engine and docker container image. So, we will start the demo before the actual hands on demo we will have a some few couple of slides on the things. So, container as if we see in a what we say very abstract way or the basic thing it is the standard unit of a software right.

So, the best part of it, it packages of the code and all its dependency right so that means, it is self content services right. So, you only have to bother that how this how you take care of this whole container image right for that you may require a container engine in this case a docker engine to house that thing right. As the analogy we are several times we are taking or from where the concept came like container which is moved from one place to another through ship or other trucks and things like that. So, it is only the we need to take care of the box right inside what it is there it is there along with its dependency. So, application runs pretty quickly and reliably.

So, quickly in the sense you do not have to install the all the dependencies before running the application and reliably means you now you know what you want to run it is running on the other end right. So, if I have a container for a particular application and then though that container is transformed from say source to destination or multiple destination or even hop

from one destination to another. So, we know that whatever the data or whatever the things I want to run that is running at the other end. Otherwise in other sense what you have to do you have to tell that what is the configuration you have done at the other end and things like that and specially where you do not have adequate expertise at the other end right. So, the that is at times becomes difficult this gives a much say easier way of executing.

So, when we talk about the docker. So, it is a docker container image is lightweight that means, it is lightweight stand alone and executable packages of the software includes everything needed to run the application right. So, basic concept is that the docker engine should be installed where at the destination system or systems. So, this docker image can be placed over it and it is just start running on the thing right. So, in this demo also we will try to show the things are how it runs.

So, there is a docker hub we mentioned about in our earlier discussion. So, from the docker hub there are standard images which are available you can create your own things what we will do here and rather we will download that basic docker image from the docker hub do some modification or adding some data etcetera and see that how I can send it to the other end. So, this overall objective or overall the demo what we are like to show is basically using two docker images one is MySQL and phpMyAdmin which are being downloaded from the docker hub you can also download directly from there. So, what we will do this MySQL say we will download in a say Ubuntu Linux system and do some modification like creating table etcetera do some operations on the things and want to transfer that for to a window system and see that without installing anything out there it start running right and similarly same thing for the phpMyAdmin. phpAdmin is more of the looking more working as a front end to this MySQL.

So, we will not be doing much change over the phpMyAdmin you can as well download it from the docker hub straight way. So, the MySQL widely is a open source and I think most of you or many of you or have used this for database applications. So, it is a relational database package which is open source and widely used similarly phpMyAdmin is provides a GUI it is a web based GUI and connects to MySQL database and again widely used with my backend MySQL database. So, it is allows you to display. So, or put the data in appropriate way.

So, our objective is that with a backend database and front end MySQL phpI run the want to do so, from the destination I want this whatever I configured or whatever the data I put on the MySQL the destination thing should run as is varies and things run start running there at the things right. So, this type of applications finds lot of there are lot of application areas specially when you are dealing with some that some users which who are more interested in running the application rather configuring and they do not rather bother about the configuring the things. So, there you can have this docker engine at every end which is not a big deal then once you transfer this docker image from one to another and then it start working right away. So, you do not have to configure and things like that. So, just to give a quick analogy if we look at this our standard systems what many of you are used to it

requires a if we want to do the same MySQL php.

So, separate installation for MySQL you have to some web server like Apache, php and phpMyAdmin. So, these are the things installation is required and once you want to transfer from one to another. So, it again require a separate installation at the other end backup the data from the old MySQL transfer restore the backup data at the new MySQL server or the destination MySQL server right. So, these things are required. So, there are lot of logistic thing and number of times you will find that this installations etcetera require some sort of expertise I do not say that you need to do rocket science for that, but nevertheless you need to be used to that where how to install, but on the other sense if I say now I have a packed image you just take it there and run right it start running right.

So, that is the thing what we want to achieve when we look at the docker based installation. So, import MySQL from the docker hub in this case import phpMyAdmin from the docker hub for both the things we are putting taking from the docker hub because that is a standard place where lot of applications are there lot of images are there. The it can be done from other places also right if you have other repositories or if you maintain a repositories that can be done from here. So, in this case we did that. So, bundle the package on Apache php my admin right.

So, this is the thing done at the source end right while transferring no separate installation required at the other end or the destination end right. Build a docker image in the source image includes everything needed to run the things like that the application along with all the dependencies it load the docker image in the target or the destination machine and it start running. So, that is the bottom line of the things right or that is the means what we say ease of running application using docker. Now so, more specific for the demo what we try to show you is that how to install this docker engine and then import latest MySQL and phpMyAdmin container from the docker hub. As I mentioned that if your application is different and given by somebody else so, that also you can do in this thing.

So, load the containers at the this that docker containers connect MySQL from php. We want to show a create a table and update some records. So, you can see at the destination or the target system also that same thing works right save docker image of the modified MySQL at the source end. Transfer the docker image to a separate machine or the destination machine and load the image at the destination machine and again connect MySQL from phpAdmin running on a different it can may run on the different machine also right that is that whether it is the same machine or other machine right. So, this is the whole thing we want to show.

So, what we will try to do that initially Ubuntu OS will make this source or the prepare this container and the modified image of the container and then in transfer to a windows machine which are not here, but somewhere in this network. So, the Ubuntu is particularly the laptop or the machine what we are showing from and then show you that how things

works right. Also we may not able to means if it is possible we will also try to show that we can put it to some cloud like AWS and see that how things works. So, that is that if time permits and also if there are no challenge in connecting to the directly to the cloud due to other constraint of the IIT Kharagpur network like firewall etcetera. So, that will try, but these two things will initially, but it is it works at the same way if you have a VM in some other any cloud in this case we will try with AWS and if the your these are running this docker engine is running it should work seamlessly.

So, the philosophy of whether it is your laptop or any other server or any other desktop or cloud it works seamlessly. So, that the thing we want to show and it will be you will be in a such a fashion that you can replicate at your end. So, we will now start with the demo which will be primarily taken care by Mr. Subabroto and Mr. Unupam for showing the step by step how things works.

So, and directly on the system it is not on the presentation rather directly on the system. So, you can have a direct feel of it ok. So, we will start the demo shortly. So, we will start the demo and Mr.

Subabro to and Mr. Unupam will show. So, as I mentioned so, initially we will be having a this installation of the docker engine and running the thing PHP MySQL and MySQL server and PHP my admin in the in our Ubuntu based systems and then we will we will show you that how do some modification and put it to the windows system right. So, Subabro to you may please start. Yes, thank you sir.

So, let us start. So, in this system first we will install docker. So, if you see when we use this command docker it might not recognize because docker is not installed here. So, let us install docker in this system. So, it is a Ubuntu based operating system.

So, we are going to install. So, let me install docker this is the command I am using to install docker. Yes. So, after this docker gets installed we will fetch some images docker images from docker hub and we are now installing docker engine. So, in this demo we are using two docker images one is MySQL docker image and another is PHP my admin docker image and those two docker images will be run as containers in this system. So, PHP my admin is front end and that connects to the backend MySQL database and so, after running these two containers we will be able to add some databases create new tables and add records into the tables and those things.

So, let us first install docker engine. So, we will be installing docker engine first in Ubuntu. So, for that I have gone to this website docs.

docker.com. So, here we need to first install this docker engine. So, let me use these commands. So, yeah yes first is sudo I have to get update. Ok. So, we are basically using these commands to this update apt package and install packages to allow this apt to use

repository over HTTPS.

So, yes let us install little less. Then we need to add a docker official jpg key. So, let us wait for this command. Yes, let us use now this command.

Alright. Now, we are using this one. This is basically needed before installing in this docker engine. Right. So, let us now do one more time this sudo update. Ok. Now, we are using this command to install docker engine.

Alright. So, let us wait for some time after this installation is complete we will check whether docker engine is installed or not using this command docker run. And if everything is fine this command like docker run hello world will run a container that will be giving us some message like hello. So, if this is successful then we can work with other docker images. And this when we use this docker run call then container image name this basically pulls the image from the docker hub. So, in this case this hello world docker image will be pulled.

If in the current system that image is not present then it is pulled from the docker hub otherwise the system the from the local system it will run the container. So, let us wait for this command to complete. See yes. So, likewise this documentation is also helps in you know further if we need to uninstall also.

So, those things are there ok. So, in the mean time let me show the docker hub basically docker hub this in the mean time it will get completed and different container images are present in docker hub right. So, those command is completed. Now, for example, if I search over here let us say MySQL it will show me the docker images like this is the official one MySQL docker image yes. So, like that different images can be obtained from there. So, let us now test whether this docker installation is successful or not for that I am using this command docker pseudo docker and hello world.

So, let us see let us first pull that hello world docker image from docker hub. Yes. So, let us try to run that hello world docker image right. So, this message is given like hello from docker. That means, this docker container has been installed successfully in this system and we are able to further run different containers and also we can means also do different operations on those docker containers yeah.

Thank you. Ok. So, what we have seen now is basically that how to install docker engine right and that you can refer as Mr. Subhamprotha has shown from the docs.docker.com how to install the docker in your say in this case a wundo system. So, that means, your basic platform is ready docker engine is ready.

Now, you can put this docker images to run the things. As of now what we did is a running a hello world docker image which is pulled directly from the things. So, the rather the latest

version is pulled from the docker hub. Docker hub is a repository as again he has shown you. So, some images if you are looking for you can look for the docker hub right.

This is this is good because this is initially at this stage is good because those those are the images which are which are tested and official images are there. So, which you can run directly. So, that is the first step to run make your docker engine ready. So, it now the next step will do what will be there is now as we mentioned MySQL and phpmyadmin pull from the docker image and running on the can will be placed on this. Now, the docker engine what we have done what he has done now on this this local wundo system and then show that how things run we do some modification in the sense that database creation and populating some records etcetera.

Then again packaging it transferring to the things is the next step ok. So, say within in like in the next session we will be showing that. Thank you. Thank you.

**Cloud Computing**

**Prof. Soumya Kanti Ghosh**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**

**Lecture 50**

**Docker Container - Demo**

So, we will start our next session on the rather continuing session on this Docker demo. So, what we have seen that how to install a Docker image in a Ubuntu platform, it is as good as you can install in other any other platform. So, then what we will do next is to for particularly for this demo, we will download that MySQL server and PHP my admin into this Docker images from the Docker hub to this Ubuntu platform, but just sometime back Mr. Subbaprath has shown you. And then we will do some update on the MySQL database that putting some records and things like that. Then again is package it and we will share it and or rather transfer it to another this Windows based Docker, Windows based system where the Docker engine is already running, we can we have already installed, but you can install yourself in other things.

So, while for transferring any other mode like FTP anything you can use in this case we will be using to drive like Google drive we will be copying and putting it down right you can use any way transfer it right. So, I request Subbaprath to continue. So, let us first pull the Docker image. So, we are pulling the MySQL image.

So, in this demo we will have this MySQL database and we will have PHP my admin that is the front end and this PHP my admin will be connected to this MySQL container. So, in that front end we will be adding means we will create a database and we will create a table and inside the table we will add some records and after that we will be able to save the images like this MySQL image with that new database right. So, then that image can be transferred to other system and those can be run in that system. So, it may be Ubuntu, may be Windows, may be CentOS, Linux ok. So, this MySQL image is getting pulled from Docker hub right.

So, it is present and now if we see I mean it is pulled if we see Docker images right this MySQL image is present. Now, let us pull PHP my admin. So, now we need to also write a Docker file. So, a means a Docker file is basically if we specify some comments over there that thing is that some things like those things will be saved over there and those things can be you know build and we can have some new image new container. Like in this case we will write this Docker file.

So, we will do this we will use this MySQL image that is being pulled before this step. So, on top of that we are basically copying the varlib MySQL directory to our customized directory

that is volume 1 varlib volume 1. So, and data directory is basically we are specifying this one this customized directory. So, this helps in basically saving the data database when the container is removed then also the you know data will be saved in this directory and we can make a new image based on this container running container. So, we will be using this Docker file.

So, let us yeah. So, let us wait for that comment oh this comment is completed ok. Now, in this where we are running in this path we are keeping the let us say we are creating a new directory we are creating this Docker file there right. So, this Docker file is saved here. Now, let us build let us first see what this whether this image is a present or not ok.

So, this MySQL container and image and my phpmyadmin image is present. Now, let us go to this directory and let us build the this Docker file. So, you will be doing some record changes and data changes right. Yes. So, basically we will be using this image this new image and using this image we will be creating new database because in the existing MySQL image which is being pulled from Docker hub in that one it is not customized which means data directories by default given, but we need to custom create a custom data directory.

So, that we can keep the data persistent ok. That is why we are building a Docker image on top of this MySQL specifying the customized directory and we will be using this image new image right. So, it is building the this Docker file right. So, this is the last step. So, now if we see Docker images yeah.

So, this new image of MySQL is created. Now we will use this image and we will run those as container and this my phpmyadmin will also run. So, let me first run this one that this thing that your MySQL image ok. So, let me write this Docker run let me give a name and then yes after giving this one we need to specify the image name also right and yeah that is it.

Yes. So, if you now write sudo docker ps ok. So, that container is not running let me check I think some parameter is missing ok. So, that this environment very well we need to give right. So, let me run once again. Yes some error is coming.

Let me change the name. So, let me yes. So, this is the command let us try to run the container. So, if we write sudo docker ps yeah it is showing that this MySQL container is running. So, now we will be running that phpmyadmin that container.

So, let me give a name phpmyadmin and yes we need to link that one like that MySQL container with this container ok. So, for that we are using basically that I think this link command is there. So, if we go there just let us see yes. So, this link we need to do like minus minus link is the command that we need to use over here and we need to write this container over here. So, this is the database that will be used over here and we also need to do this port forwarding over here right.

So, if we run now let us see what happens this container should be running ok right. Let us see what is there yeah sudo docker ps is giving me that these two containers are running ok that is fine. Now let us go to this web browser and let us go over here 8080 localhost colon 8080. So, if we see over here that phpmyadmin this interface is here. So, what you have done you have taken this MySQL image and the phpmyadmin from the docker and you are running in the docker engine installed here right.

Yes. So, that means, you do not have to now install MySQL as such anything out here. Right. Right. So, it is from the docker hub he has downloaded that MySQL server and phpmyadmin and then running right away.

Right. Right. So, here now it is. Now your phpmyadmin is connected to the backend database right. Exactly. And MySQL database which are both two containers right.

Yes sir. So, these two containers we have created and now we can see that this phpmyadmin is up and it is connected to that MySQL server database. Now let us try to create new database and new tables over here. So yeah. Matthew actually this is a dv tb. Yes, yes we are first creating a new database and inside that we are creating some new table.

I think you give npTEL tab you can give it something like. Yeah yeah. NPTEL tab. Yes, let us take three columns say for timing.

Ok. So, yeah let us create. Name of the table may be course. Yeah. This column is course id.

Right. Say let it be integer. Yes. Say 1 0 and then 4 ok fine. Yeah and course name.

Right. And then credit. Yeah. Course date. I want the credit also.

Ok. So, that is yes. 4 ok. Yeah. So, if we now save this schema will be ready. Now we can insert some record in this database. Right. So, let me insert some data.

Say course id 1001. Yes. And what is course name is say cloud computing and credit is 3. Right. Let us also add some other records. So, that cloud computing should be within.

Yeah yeah correct. Let us say string right and let us say this. 102 network only networking say 4. Correct.

So, let us try to add this two. Right. Yes. Yes. So, these two are inserted. So, you can see that thing.

Yeah select star if we do correct. So, these two are there. Now we can do one thing. We can

save this running container into an image new image. So, yeah. So, for that we will use docker sorry yes for that we will use docker commit this command and this is the container id and let me give a new name yes.

So, this will basically save a new image with this name MySQL underscore v3 and it is basically saving the current container state basically with this new database new table those things will be there. Yes. Now if we see pseudo docker images first of all in ps it is running because we have not closed I mean those things. The pseudo docker images if we go here we can see this has been created.

Yeah just few seconds ago. Now we can do one thing. We can basically move this image as a tar file to another system. Ok. Let me create a tar archive.

And this backup is created run time means. Yes. It is a means we have not stopped the MySQL and all. So, it is created run time. Right. So, for saving into tar image we will be using that command pseudo docker save. Pseudo docker save right yes pseudo docker save then this one I think the syntax let me see in fact in here also they have shown docker save docker docs dot docker dot com they have given here the syntax yeah.

So, we need to give like that that a container name sorry image name then the tar archive. Right MySQL v3. Correct yeah pseudo docker save MySQL v3. Better than something.

Yeah some image some archive name. My tar. My image let us say. TAR. Yes tar archive yes. So, in this directory this tar file will be created because we are currently in this directory yeah it is getting created over here yeah it is getting created. Then for transferring you will be using in the in this case Google right to transfer.

Yes yes. We can use other things also direct FTP etcetera, but here we are using. Yeah. So, the tar file is being created. So, that will be transferred to a we will show you into a windows system.

Yes. So, it will take some time. Right. So, because it is having the this MySQL instance along with all its dependency etcetera. Right. And in that new system this MySQL image will be run as a container. First we will be loading that this tar archive to Docker image and then.

Yeah. Yeah then we will be running this container and PHP my admin will be pulling from Docker hub itself. So, at the destination as we mentioning we will be using this updated MySQL updated in the sense there with the. This change is. Data loaded means that table created etcetera and whereas, in the destination system windows system the my PHP my admin will be downloaded from the Docker hub. We could have transfer that also from here, but as there is no change.

So, using the latest thing from the Docker hub may be a good thing. Yeah. So, it is completed. Now, let us upload this one in a shared directory.

Do care directly empty table or something? Ok. Here will upload. So, this is basically only transferring the file. Now, we download from a this one windows system.

Yes. So, that docker shipping that is the. Correct. So, shipping that off the docker and it will be used at the other end. Yeah. Like analogy is that you are putting on the ship and transferring that container to other destination and then opening up the container at the other end.

Right. With the changes this thing. With the updated things. Yes. So, you need to share the location with the. Right. Destination. With the destination.

File size is little high that is why it is taking. Yeah. Some time. Right. Yes. So, now we will be looking downloading the thing in a windows system and where docker engine is already running in the same process how it is been done in the ubuntu you can see that reference in the docs.docker.com and now we will be showing that the updated MySQL updated in the same data has been updated that you can run on the things ok.

Please. So, here we got the tar file that has been generated and it is placed in this as a my image 1.tar. So, let us first download this one. So, here we are using this Google drive for shipping the container, but it may be something else. So, we can transfer it by maybe running scp command or maybe through the FTP, but the concept is that we have created some container at the distance place and we have updated that data and then we are trying to we have shipped that container and now we are running into a separate machine.

So, that container the content within that container should be as same as the at it has been shipped earlier. So, it is getting downloaded and here in the in case of windows. So, if we install the docker. So, windows provides a docker desktop. So, through that that is a graphical interface through which we can monitor what are the images running and what are the containers running and what are the images available.

So, once we install that docker we will get a default image and that default container. So, that is running that will be running in port 80. So, we have stopped that one.

So, that we can run the other things. Now, let us see whether it is yes. So, now it has been downloaded. So, I think it is downloaded and yeah. So, let me. So, this is the image I got.

Now, next part is to I have to load this image into our docker. So, what we are doing is we will open a command prompt and then we go we will write this command docker load we are loading this container this container the name was my image 1 dot img1 dot tar. So, if you can see that here at present there is only one image. Now, once I run this one. So, I have

to first go to that directory and then load.

So, it is getting loaded. Now, you can see that in this docker in this machine. So, there is one container one image is there. So, once it is completed. So, you will get another image will be coming.

So, this my SQL v3 that name that we have provided there. So, the same name. So, that image has appeared here. Now, it is loaded. Now, our job is to run this my SQL. So, shipping part quote unquote shipping part has been done now the image has been shipped from the source to the SQL.

Docker run minus minus name the name that we have provided. So, one name we have to provide it here and later on if we want to connect it from the phpmyadmin that name we have to specify there. So, that is why we have to provide a custom name here. So, let me put it as my SQL 1 then say minus d then it is my SQL and this is the image name. So, this image name it is my SQL underscore v3 underscore v3.

So, we are running the my SQL. So, now, it is completed. So, if you go now in the container you can see that this my SQL 1 it is running as my SQL 1 and this my SQL v3 that image is running as my SQL 1 container. Now, it is part of the phpmyadmin. So, we have shipped only the my SQL because the some table are updated and some table we have created and then we have inserted some data. So, there is a change in the my SQL, but there is no change in the phpmyadmin. So, that is why we have not shipped that one we are directly downloading from the docker hub.

So, here also we are providing a name that phpmyadmin it will run as a container name will be phpmyadmin then we are linking that my SQL which was running. So, that name is my SQL 1 it is running as db so colon db minus p that is for port forwarding. So, we will run it the phpmyadmin will run in 8080 port then we are pulling the phpmyadmin and the latest version. So, from the docker hub we are pulling my phpmyadmin and giving a name called phpmyadmin 1.

Yes and that should connect the SQL 1 container. So, as the phpmyadmin package is not available in the local. So, now it will first connect to local and then if it is not found then only it will go for downloading. So, here you can see that there is two images. So, now, the third image will come and the third container will also come as running. So, so now the new phpmyadmin appeared here and if you go so this phpmyadmin is running in port 8080.

Now, let us go to this phpmyadmin which is running locally in port 8080. So, if we go there so here this phpmyadmin is running and let us log in so here you can see that this mydb1 that was created earlier it is there the nptel tab the table which have created it is existing here and the columns those three columns course id course name and credit all those things

are here available. So, let us see what are the data's are there. So, here this cloud computing 3 1001 and 1002 the networking 4. So, those records which are updated in the in a other machine and which has shipped as a container in my machine.

So, the data is still present. So, we have preserved the state we have shipped the container and we are importing that here, but without we are not installing any separate MySQL or PHP or Apache anything. So, we are just placing that container on the top of a Docker engine and the things are running. So, this is a so hope this demo will help you in understanding the basic philosophy of container and though we have used this Docker to demonstrate it and as it is a open source and there are lot of documentations under the docker.

com or docs.docker.com I encourage you to do it on your end and see that how things works it will give you a confidence on how things work. Thank you. So, we will end our demo session today and we will continue our discussion in the subsequent lectures. Thank you. Thank you.

**Cloud Computing**  
**Prof. Soumya Kanti Ghosh**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**

**Lecture 51**  
**Dew Computing**

Hello. So, today discuss about some of the computing paradigms more specifically today we will discuss about due computing paradigm which are some of the related cloud computing related computing paradigm. So, in this thing what we will discuss about that overview or what is the basic concept of this due computing, what are the basic features, this different type of applications and a broad architecture of the due right. And definitely also we will end up with the type of challenges due computing is facing right or what are the different challenges out there right. So, if we see your today most of us are used with different type of services like one of the popular services is Dropbox right where we use as a storage primarily for storage purpose. And we have also experienced whether the whether there is a internet connection is there or not your Dropbox folder is primarily synced is available for your purpose and at a local level.

Or in other sense it is something as a storage as a service which is synced with that in the with the cloud or storage where wherever the Dropbox is there, but it has also a local instance right or what we say which is which can be accessed even if the connection is not there. So, rather if you those who are working with cloud computing somewhere other we have all have experience of this cloud compute thing. So, one of the major challenge is the this your reliable or continuous uninterrupted internet connectivity. If the connectivity is lost then you cannot able to do something right do your work.

So, due is try to have. So, Dropbox is an example of due computing where you have a local instance for that matter what we say some sort of a local sort of a due virtual machine is installed in your systems which not only allow you provide you services whatever the services is provide for in case of Dropbox the type of data level services. And also it sync with the cloud as and when it gets the connectivity right. So, if you look at the overall big picture of the computing paradigm taken from this particular paper. So, the internet is the binding things right.

So, though the major this paradigm is the cloud computing, but there are different allied paradigm like mobile ad hoc computing, fog computing, edge computing, cloudlet computing, mobile cloud computing, mobile edge computing and of course, there is a concept called due computing right. So, we will be looking into little bit more into this due computing paradigm how things works right. So, but nevertheless for all these scenarios what we require is primarily is this reliable or constant network connectivity. If the connectivity is lost then we face problem in getting the services for that particular purpose

or particular activity application areas. Now, due computing also referred to as this as different perspective or you can see that if you find the literature there are different way of looking at it right though the basic philosophy is same, but different.

So, it is a computing paradigm that combines the core concept of cloud computing definitely with the capability of the end devices right. So, personal computers, mobile phones etcetera. So, which comes at the end devices try to leverage on the this capabilities of the end devices right. It is used to enhance the experience of the end user in comparison to only using the cloud computing. So, say I have like if you again refer to this our Dropbox example.

So, the storage could have been in the cloud I get the connectivity I work on the things right. So, that has requires a constant connectivity and, but I have that thing if I have a some sort of a virtual instance of that Dropbox in my system or having the. So, I can what I can do I can still work on the thing whenever this whenever the connectivity is there is sink right. So, it goes on sinking with the backend cloud infrastructure right. So, it enhance the experience.

So, it is rather gives more smooth experience to the end user right and the and the as the name suggest also from the cloud is far above then we come down to something called fog and then when we come to the dew we are basically at the ground level right or at the device level. So, dew computing as we are discussing attempts to solve one of the major challenge in case of a cloud computing thing is that reliance or on this network access. So, constant or ubiquitous uninterrupted network connectivity that we can get rid of right or it we should not say get rid of we it even it works when the connectivity is not there right. So, other let us see that what are the different way of looking at it. So, it is a computing model for enabling ubiquitous pervasive convenient ready to go plug in flexibility and power personal network and include single hyper hybrid peer to peer communication link right.

So, I can have one communication link with the rest of the world or the cloud, but however, I can work out here. So, primary goal to access a pool of raw data equipped with metadata that can be rapidly created, edited, stored and deleted with minimal internet management effort that means, can be also done with a in a offline mode right. So, that is one of the primary goal. So, also to utilize this all functionalities of cloud computing network users are heavily dependent on the network connectivity of that time which dew computing tries to facilitate that even without network connectivity things can work. So, little bit jumbled up means too many too many lines on this slide, but let us go little slowly and try to have the whole thing in a bundle right.

So, it is a paradigm which is user centric right and highly flexible and personalized supported applications are prioritized right. It is located very close to the devices right or end devices and it is the sort of a looked as a IoT for cloud continuum right. So, it is if we see the IoT for cloud continuum. So, it is the first thing from the user end we which you attempt

to interact with right. So, it is a microservice based computing paradigm rather in similar literature they primarily focus on this aspects with vertically distributed hierarchy right you have a hierarchy things on the up in the ladder.

So, it comprises of smart devices like it can be smart phone, it can be smart watches, it can be tablets and even other any our laptop and other devices located at the edge of the network to connect to the end devices right. Collect and process IoT sense data and offer other services right. The services in due are relatively available and it is not mandatory to have permanent interaction. So, the due services are mostly always sort of a always available and you may not have the always the network connects TBT right. So, in other sense this it allows you to work when the connectivity is not there it is not to be seen that without connectivity you still sync etcetera.

So, that it requires like those who have used things like drop box, google drive and type of things you have seen that it still require you need to sync the data right. So, this is a microservice based which means that it does not depend on any centralized server or cloud center cloud data centers right. So, it is one of the other thing is that it is a microservice based operation. So, it is does not depend on any centralized control or DC cloud DC type of things right. Though it requires to sync with the this cloud infrastructure or cloud data center, but it is not dependent on fully on that.

So, DC does not rely on centralized computing devices nor a permanent network connectivity to work on right. So, typical example as we are telling drop box is an example of the dew computing paradigm as it provides access to the files for the folders in the cloud in addition to keeping copies in the local device. So, whenever the connection is not there then also user can access those right. Allows the user to access files you during times without an internet connectivity when the connection is established again the files folders are synced back to the cloud server. So, that is the thing and those most of are have are used to this type of system.

So, we can have this type of paradigm which allows you to work when the connectivity or the cloud connections is not there right. In some cases it so also so happen that it is not only that connectivity is lost it may save your energy in a or overall costing in a big way. Because you do not we do not want to connect you want to work and then you connect when you get a favorable connections or high bandwidth connections things like that right. So, you do not have to be always connected with the things right with the cloud backend. So, it is obvious the key features are independence and collaboration right.

Like if you see in case of a drop box we are falling back to the example of the drop box because we are used to it and we see these features directly we have used those features right. One is that it is independence like from the connectivity and things like that. So, means that the local device must be able to provide service without continuous connection to the internet. So, it is independent like you have some file in the drop box you work on

that file even the connection is not there still you can work. And collaboration means the application must be able to connect to the cloud service and synchronize with the data when appropriate things are there right.

It is should be done on a means sort of a automated mode quote unquote automated mode you do not have to forcefully sync you may opt for that options like forcefully sync. Additionally there can be a lot of other things which we can enable like what we say sharing of the data based on some role based control right. So, we can share with some or say role or rule based accesses right. I say that this for this share at all those things. So, as we discussed you reflects the natural phenomena clouds are far from the ground fog is closer to the ground and dew is on the ground right.

Analogically cloud computing is remote service fog computing beside the user and dew computing at the user system or user premise or user end itself right. So, that the it attempts to give those type of service. So, there are several model and some of the typical applications like one is infrastructure as a as dew like local device is dynamically supported by the cloud services like iCloud and things like that which is a infrastructure as a dew sometime known as IAD software in dew right. The configuration and ownership of the software as saved in that particular platform. So, that is that is saved in the means cloud whereas, this we have those things at the like typical example are your Google play store or app store and these are the things which are there.

Platform in dew a software development suite must be installed on the local device with the settings and application data synchronized to the clouds services. So, GitHub is one of the one such example which are we are using this sort of infrastructure. There are other service model like storage like what you in dew like storage of local device is partially or fully copied to the cloud right. So, one is the drop box, Google drive these are some of the examples and web in dew a local device must possesses a duplicated fraction of the WWW or world wide web that is the web in what we consider means is the web in dew right. So, these are the several applications or several service models which dew computing paradigms try to leverage on right.

So, what we see considering this. So, I have several services like storage, platform, software, infrastructure along with that web in dew or you can have your own other type of applications which we can have that on realized on the new platform. So, now looking at the broad architecture. So, what we see one side what we require is the cloud right. So, which need to be synced with a either a single hybrid P2P communication link which whenever the connection is there it is there.

Other side you have IOT devices and sensors where the data are being collected right. And you have a dew framework in your host system often referred to as dew virtual machine or DVM right. In some of the literature you will find that dew virtual machine which must have a dew server which is the core of the thing along with that there is a dew client program

which interacts with the client. There can be a few DBMS which have database which helps in storing the data in a more organized way and there are few client services which provide services to the thing. There can be some of the other things like in some places you will see that there are some of the means intelligent tools and techniques which takes care of the things.

There are because it need to learn that what sort of the data access and configure the things at the at the few end right. So, there are different module AI module or other things may come, these are the things what we expect that the few will have. So, one is that interfacing with the client services another is that server do sync with the cloud server based on the connectivity and the type of data being exchanged and this need of the data of the exchange etcetera. So, to establish a cloud few architecture on a local machine a few virtual machine is needed right some sort of a this few virtual machine or few that particular few system need to be installed like what we do we need to install the drop box at our end right and DVM is an isolated environment for executing the few server on the local system. So, it is a isolated independent environment right to run the few server at the local system.

It has the capability of syncing with the other end cloud system right. So, if you look at the basic function of the few server to serve user with the requested services through that client services module and things like that and to perform synchronization and correlate between the local data and remote data right. So, to perform the synchronization and correlate between the local data and the remote data through these are the things which are by the few server. So, attempt to achieve three goals one is data replication, data distribution and synchronization right. So, these are the things it attempt to look at it.

So, need to serve the user end and need to synchronize at the other end right. So, this is the one of the means basic goal of the things. There are other housekeeping work like based on if the data is being specially in case of a storage it is the data is being changing or being updated deleted things. So, how overall organization will be data organization within the few database or data repository has to be looked into right. So, those are the things which the server takes care.

So, now let us see some of the categories or broad application areas right. So, one is the web in the few right possesses a duplicate fraction of the worldwide web or a modified copy of the fraction to satisfy the independent feature right. So, because this fraction synchronizes with the web it satisfies the collaboration between feature of the few computing right. So, this is one of the aspects which is which need to support like web in the few is one category like you have a fraction of the worldwide web into your local systems which you need to work on it. There are storage in the few already we have seen that the storage of the local device is partially or fully copied to the cloud right based on the things need to be copied.

Since the user can access the files at any time without the need of constant internet connection or connectivity the category this particular category or application area meets the independence feature of the new computing right. Storage in new or SID also meets the collaboration feature because the folder and its content automatically synchronize with the cloud service as we are discussing. So, it is both independence and collaborating features what we see is there whereas, in the web though it synchronize with the web in the WID also it synchronize with the web and it satisfy the collaboration feature of the new. The other thing is the database in new the local device and the cloud both store copies of the same database right. So, one of these two databases considered the main version and can be defined as such the database administrator which is the primary version of the database and where other one which need to be synced.

This service increases reliability of a database as one of the databases can be act as a backup to the other. So, by default it has a backup server DB server of the system it works. So, we have web in new, storage in new, database in new right. So, these are the typical scenarios what we are there are few more one is software in new as we are as we seen in the couple of slide before. So, the configuration and ownership of the software are saved in the cloud right.

So, examples include this Apple App Store or Google play where the application the user installs are saved to their accounts can be installed on any device linked to their account right. So, this is the way it works right. So, it is a software in the new there can be platform in the new like software development suite must be installed on the local device with the settings and application data synchronized to the cloud service. It must be able to synchronize the development data, data deployment, the system deployment data and online backups. So, like one example those who are working with GitHub may find this thing very natural what you are doing right like it should be must be able to synchronize development data, system deployment data and online backups.

So, that is that what we say platform in new right or PID. And finally, the infrastructure as new or IAD local device are dynamically supported by the cloud services right. IAD can come in different form, but there are some of the popular forms the local device can have exact duplicate DVM like what we have discussed that is new virtual machine instant in the cloud which is always kept in the same state at the local instance whenever the connectivity is there or it can be the local device can have all its settings data saved in the cloud including system setting data and data for each application. So, it depends on how things works, but we can have that infrastructure as a new. So, what we see that the there are different categories of new computing paradigm starting from infrastructure platform software and also things like that web in new, storage in new and database in new.

So, there are several categories what we can see into this type of applications. So, what are the there are challenges or there are several challenges thing one is definitely the power management right. So, what should be the if based on your different categories and

application area how the power management is done is definitely a serious challenge right. Other thing is the processor utility like as on the local system. So, it is using the processor along with other services.

So, what is the processor utility is makes a makes a important aspects right. So, other is the data storage like because you need to store the data and also need to sync with the cloud. So, there are different issues like specially how much data you have subscribed to and also how much data you require from the local infrastructure or local system. So, those are need to be appropriately looked into not only having the subscription on the cloud storage, but also you need to have equivalent things you may be in the local system right. As such it is not very challenging if you have a laptop or desktop, but it may be challenging when you are using the same thing in some of the low resource devices right.

Like it may be even mobile devices or other type of devices when you are using then it may be a say smart watch or so that may be a challenge. Viability of the operating system. So, based where the your dual virtual machine is running. So, what sort of operating system underlining you it supports or it can run on is there. Programming principles that is also another challenge that what is how much configurable or how much what type of services you can provide based on the thing.

And also it comes as it also definitely a security on information or database security is a concern because it is now exposed to the that system and other services can whether it can access this data and others means due data the data stored in the due or not those are the things so that is there. So, these are some of the challenges where due computing paradigm which what the due computing paradigm faces while going while providing flexibility in terms of independence and collaboration things like that right. And now if we again come back to this big picture where the due enabled computing may be helpful one. So, if you look at the whole stack. So, at the top we have the cloud server then we have actually we can have fog servers or here we refer to as edge servers.

So, fog devices then we can have edge devices, smart devices and smart IOT modules right. So, these are the different layers of or hierarchy of this different type of category of servers and systems right. So, due enabled computing can be realized at the lower hierarchy starting from edge to smart device to smart IOT modules. So, at this end we can enable this thing in the due computing. So, those are those still work when the connectivity is not there right.

So, based on that last synchronized data it can work and you can update and work on the things and get the services whatever is there and then sync with the upper layers as and when the connectivity is available ok. So, this what we see that is the overall how do you play a big a platform on the things. So, what we try to look into today's discussion is that whether we can have we can bring down the computing already we brought down to this fog we have seen that how fog and edge. Now bringing to the ground that means, to the

device level where it can provide some independent services or provide some independent microservices irrespective whether the connectivity is there or not right. So, it facilitates the end users with a continuous service provisioning right that is important like it provides the services even if the connectivity is not there and on the other end whenever the connectivity is there it allows to sync with the cloud infrastructure.

So, that is one good part of this or one advantage of this due computing and it can be it can be enabled at the lower levels with this due with this particular philosophy of due computing paradigm ok. So, with this let us end our discussion today's discussion there are few references so, you may have a look on those references right. And so, let us conclude our discussion we will continue with other aspects of cloud computing in our next class. Thank you.

**Cloud Computing**  
**Prof. Soumya Kanti Ghosh**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**

**Lecture 52**  
**Serverless Computing - I**

Hello, continue our discussion on Cloud Computing. Rather if you remember the last session or the last class we discussed on one of the computing paradigm. So, rather some of the this correlated or associated computing paradigms we are discussing in this couple of lectures. So, today what we will take up is a serverless computing right. There is a very emerging area and what we say that it is a buzz word which is there in this computing domain and having lot of interest you will find that lot of conferences and other journals papers are coming up on serverless computing. So, we will try to see what it is and rather in next lecture also we will see little more detail of the serverless computing right.

So, we will be covering mostly the serverless computing and what comes back to back with it is a function as a service even sometime a term called backend as a service right. So, these are the two keywords. So, what we have seen in case of cloud computing. So, what we see that computing as a utility right.

So, all our things is a computing as a utility and also what we have seen that we have different models like one is your IS, PAS, SAS these are the predominant models right like infrastructure as a service and say platform as a service and software as a service. So, this has different manifestation right. So, and what we try to do you from your on premise computing infrastructure you are hiring the computing infrastructure, but if you look at in several cases we need to still provision the thing right like provision a VAM, provision that what sort of a OS etcetera underline you need to look at when the development is going on and things like that or rather whether it is a your own server or a cloud server you still are linked with the server primarily I mean to say the developers are still linked very much linked with the server and need to have a lot of say quote unquote investment in terms of time and expertise and sometimes this money that how the it can be configured to use it on the things. So, this new way of looking at it new paradigm what we look at it the serverless computing. So, let us see that in some of different aspect how try to have a feel of the thing that what is tries to say.

So, it is a computing method of providing back end services on as used basis right as I am using or as I go on using the thing it has been built or charged based on the things. So, the back end services right. So, the serverless provider allows users to write and deploy code without the hassle of worrying about the underlining infrastructure right. So, the infrastructure management is being taken out as by the serverless provider right. So, we will see that how this FAS or BAS type of things which is coming into.

So, serverless architecture simplifies the code development and deployment and eliminates the need of system administration allowing developers to focus on the core business logic what it is looking what he or she is looking for without creating additional overhead like instantiating the resources such as instantiating VMs or containers what we seen couple of classes back containers in monitoring the infrastructure right. So, what it tries to do that the that core developers should be involved in developing the job rather than looking at the that back end services or the back end management and monitoring issues right. So, if we can segregate this or if we can provide this type of a service things will become a much what we say beneficial to these developers or for developing and deploying their applications. Now what we see in this model the developer execute their logic in form of functions. So, what we try to do that their logic is some sort of a function and submit to the cloud provider to run the task in a shared runtime environment right.

So, more developed the logic as a function. So, what we are looking as a going towards that a means giving as a function as a service and submit that to the this service provider to in a shared runtime environment. So, cloud provider manages the scalability needs of the function by running multiple functions in parallel right. So, it one may one way of looking at it that the if I want to scale. Now suppose you want to scale in other way you need to look at you need to think about that the whole thing right whole base of the back end etcetera right how things will scale and things like that right.

So, you still the provider will provide in as a as a IS service provider or PaaS or SaaS service provider, but nevertheless you need to look at the what we see that scaling factor. But here the provider basically take the scale up things based on the type of by like one is that running multiple function in a in parallel right. Rather we will see that it is something which is going little bit overlapping of or something what PaaS type of service provides right, but we will see that where it goes. So, following the wide scale applications so, one thing what emerged out in recent times is that wide scale proliferation of this containers right or the services provided by the in form of containers that one the things which is what we see in last say couple of years or so right. So, wide scale application of the containers we have also seen that how easily you can take the container from a from say windows to Linux to say any cloud platforms what they say from your desktop to laptop to cloud it works in a seamless fashion.

You do not have to go install the things of the of installing the things, but nevertheless you need to load boot and if there are orchestration of the thing multiple things need to be orchestration that need to look at. Like if you remember that our previous couple of class when we back when we were discussing about containers especially the example scenario where we have taken up two container containers one from one is that your MySQL and PHP main from this docker hub and then we can run if the docker this engine is running. So, you can run upfront right, but still you need to run that docker engine right. So, that is there, but nevertheless this container has given a different dimension to deployment of

different services and specially looking at the microservices and things like that. So, following the wide scale application of containerization approach the cloud service provider adopted to offer a better fitting containers that require less time to boot right.

So, it is something the container needs to wake up. So, or what we say time to less time to load or boot and to provide increased automation in handling orchestration that is another challenge containers on the behalf of the client right. So, there may be challenges of orchestration of different services. So, how it can be it can facilitate that is another challenge which need to be managed by the why we need to be realized by the or the client right client want to have that orchestration of services to basically realize a particular outcome. So, serverless computing promises to achieve full automation in managing fine grained container right.

So, what we coined the word called fine grained. So, this our serverless computing promises or attempts to achieve this full automation in fine grained containers in quick loading or booting up the things or automatic or handling better handling or in automation in handling the orchestration or containers orchestration of containers services and things like that. Now so, if we again try to look in that in a little bit whatever we have discussed so far. So, serverless computing is a form of cloud computing. So, we are trying to realize another a form or model of cloud computing that allows users to run event driven and granular applications right.

So, triggered by the event so, event driven and granular application without having to address the operational logic right. So, underground whatever the backend operational logic is there user are not bothered, but it can trigger this event driven and granular application. So, serverless as a computing abstraction if we look at. So, with serverless developers focus on high level abstraction as we were discussing like function queries events like that and bit application that the infrastructure operates map to concentrate resources and supporting services. So, what it is the developer is trying to do? It only it basically focuses on the business logic or what we say that high level abstractions of the things like at the function level or you can look at the query or event level and build application that the infrastructure operator or the service providers map to the resources and the supporting services right.

So, you build your functions and instead of bothering that how it should run on the thing the service provider or the serverless service provider so to say, they take care of this mapping to the actual resources and this other supporting services if there are any. So, developers basically focus on the core business logic and on ways to interconnect the elements of business logic into complex workflows. So, whenever we do have a particular application in mind so there is a definite workflow right. So, the developers bothered about the overall business logic and how to interconnect this workflow to achieve something right. So, that is the that as such the developer doing along with that you do not have a you do not have to manage the back end so.

So, service provider ensured that the serverless applications are orchestrated that is it puts into put into container deployed, provisioned and available on demand. See this is interesting right. So, this that the back end it takes care of how it the containerization has to be done, how it need to be deployed, how need to be provisioned and how it can be made available on demand. So, while billing the user only for the resource used right whatever the resource is used that may be the billing anyway that is a billing procedure how things will work that is one thing, but the back end thing is taking care of all those things ok. So, this is the way we want to say what I can say that segregate or isolate this infrastructure management of a cloud like say VAM and other things along with that application development.

So, or even container management things etcetera taken care at the back end. So, one two terms come into play we will see that one is a function as a service and another is a back end as a service right. So, when we look at the function as a service. So, client of the serverless computing can use as a can basically use this function as a service model. So, what when you talk about a FAS is a form of serverless computing.

So, it is a form of a serverless computing in which the cloud provider manages the resource lifecycle and event driven execution of the user provided functions. So, the users or the client provide this function and the cloud provider or the serverless computing providers takes care of the resources needed, lifecycle of the this process and event execution of the this sorry user level functions right. So, those are taken care by the cloud provider. For like we can so, with function as a service. So, user provides small statelet functions to the cloud provider which manages the operational aspects of running ring function right.

So, as if you can realize that to realize your overall applications you have number of functions which are you know the workflow, but you are not bothered about or not concerned about how they the containers will be made, how it will be deployed, how this overall based on the workflow this overall orchestration of the backend things will be there that you are not bothered right. So, one example available in this reference like considered that x camera. So, it is considering some camera application which uses cloud functions and workflow to edit, transform, encode videos with low latency and cost. So, the whole application that in this so called this cam this video application or camera application what is what the application is named as x camera. So, what it is doing it is basically in wants to edit, transform, encode videos with low latency and say minimal cost.

So, that is the overall objective of this application and which is being taken care by this serverless things right. So, a majority of task if you see in this operations what we if we look at even x camera or any type of applications can be executed concurrently right. Some of the things I can we can execute like if the if there is a transmission of some videos or something I can do parallel transmission or several things. Suppose I do some frame by frame operations some of the things we parallel frames wise or a group of frames and

things like that we can do right. So, even encoding etcetera so wherever the parallelization is possible right.

So, can be, but there are lot of opportunity of parallelization executing concurrently allowing the application to improve its performance through parallelizing this task right. So, the if the if we have that type of function as a service then say executing multiple things by the route provider may improve the overall performance of this particular applications right. So, though the picture is not good you will find in this the references whatever given at the end. So, if we look at that evolution of serverless computing starting from 1960s when we are having this IBM systems. So, in 70s this virtualization other things were coming up then one we see that in 80s this RPC implementations and RPC things were in the high node that is that remote procedure installs then in 90s what we see that what that private this virtual private servers are all those things are coming up in a big way.

And along with that there are other things like on the other side CGI if you see that resources code function naming and registry services functions execution flow and different type of aspects. And what we see in 2000 is that this containers line at containers and other things where where AWS cloud all those things came up Google App Engine and the all those surfaced. And 2010 onwards what we are looking at that more proliferation or a big way proliferation of container services like Google sorry Docker came in some 2013 and then Kubernetes and so and so forth. And other things like IS pass SaaS type of cloud which services cloud based services came in a big way in a what we say what we can what we say that is in a something which is deployed are being used right. So, generally it was more of a organizations own uses, but it becomes as a public services right.

So, public at large or city means different organization can have this type of third party services of the of services right. On the other hand like if you see on the events so that event driven architecture also 2010 onwards we see that lot of activities there also. So, taking together what we look at that orchestration this container orchestration and microservices is one part which are coming came up or what we see it is a booming thing in these days and which basically lead to this formation or realization of a function as a service model. And on the other hand we see that event driven workflows right. So, we have a business workflow or my application which is event driven.

So, this things what with function as a services trying to realize this event driven workflows or this together is basically lead to this serverless computing paradigm right which we see right. So, though it is came up from the is a say model based on this cloud computing paradigm and leverage on different thing this cloud computing things, but nevertheless what we see that it came up in a this event driven workflows and other side this function as a service type of things. So, if we look at that in serverless computing cloud provider dynamically allocates and provision servers right. So, servers are there right. So, only thing you have a abstraction of the things and try to realize that thing right.

The code is executed in almost stateless containers that are event triggered or event driven and maybe ephemeral right like last for one invocation and things like that right. So, if we as we discussed the serverless attempts to cover a wide range of technology primarily grouped into two categories one is that backend as a service another is a function as a service we have seen or we have discussed something on the function as a service right. So, in this brought to categories what we try to see. So, as the backend as a service at it the name itself implies enables to replace the server side components with of the self services right. So, instead of looking at the server side component we are more looking at that services in terms of say functions and things like that right.

So, bash enables developers to outsource all the aspects behind a scene of an application. So, that the developers can choose to write and maintain all application logic in the front end. So, in other sense this all this backend load is now taken care by this backend as a service. So, I require this type of environment for my code it takes care of the things. So, typical examples as example scenarios are remote authentication system, database management, cloud storage, hostings which are going at the backend of the things right.

Rather those who have used that Google Firebase you may have seen that is a fully managed database that can be directly used from an applications right. So, that the Firebase Google Firebase I think it takes care both means NoSQL and relational model both you please look into it I as I remember what I am telling. So, Google Firebase is a, but it is a fully realized database right where you can quickly connect to other applications and if means quickly connect to specially this web applications and things without very seamlessly right. So, in case of this Firebase as a say backend services as a database and things like that. So, manage data components on the you on users behalf right.

So, whatever the data etcetera is there this Firebase takes care of the backend services right. So, this is the thing which we look at when we look at the backend as a service. And the function as a service already we discussed. So, the serverless applications are event driven cloud based systems right. So, it is we are considering that over and above means basically my cloud system is in place, where application development relies slowly on combination of third party services right client side logic and cloud hosted remote procedure call.

See we are not talking about any server etcetera here right. So, it is a third party services client side logic the basic business logic and the cloud hosted RPC or remote procedure calls right. So, FAS allows developers to deploy code that upon being triggered is executed in a isolated environment. So, it is as if is executed on a server mode.

So, that the things it tries to ensure. So, function are so, each function typically describe a small part of the entire applications right. So, I have the applications what we realize is as a bunch of functions and each function takes care of a chunk of this application. The execution time of the function is typically limited right. So, like if you look at AWS lambda it

is typically I think 15 minutes or so. So, it is based on the type of things it is the execution time is limited.

So, function are not constantly active instant FAS platform listen for the events that initiate the function right. So, function are instantiated whenever that things are that platforms are this it is triggered right. So, thus the function are triggered by events such as client request, events produced by any external system, data stream and other things like that right. So, it is triggered by the events right like in the camera application and I do not know that how only it works, but it can be event by a movement of a any object of the things it is getting triggered. It is not like that every time it is capturing and edit a filter means capturing, storing, filtering the image, videos etcetera, but nevertheless it is triggered by something right.

It is it can be triggered by a some object movement, it can be triggered by something that predefined time of switching on the cameras and so and so forth. So, the FAS provider is then responsible to horizontally scale function execution response to the number of incoming events right. So, based on the number of incoming events the FAS this your FAS provider will scale up the things right. So, like having different parallel execution of the function and like that. So, there are if we look at there are not all good things there are few there are several challenges or there are several what we say not so good things about this serverless computing.

So, one is that asynchronous call like asynchronous call to and between the serverless functions increase complexity of the system right. So, it is a asynchronous. So, usually remote API calls follow request response model and easier to implement with synchronous call. So, if there is a there is been synchronous call it is easy to implement, but if you have asynchronous call then there is a challenge in it. Function calling other function that is another challenge that complex debugging lose isolation of features extra cost a functions are called synchronously as we need to pay for two functions running at the same time that is a billing or cost wise thing, but nevertheless it becomes a more complex situation.

There are issues of shared code between functions. So, might break existing serverless function that depends on the shared code that is that is changed right. So, it is based on the on some shared code and that the shared code has changed now it has to do means it has to execute the thing. So, risk to hit image size limit like typically AWS lambda the 50 MB is the image size limit or warm up time the bigger the image longer it takes to start or load or become active. So, there are issues like that.

So, uses of too many libraries right. So, increase space used by the libraries increase the risk to hit image size limit and so and so forth. So, too many libraries adaptation of too many technologies right such as different libraries different frameworks languages now you have become very flexible right you are telling that lot of things need to be all my back end work is maintained by the provider I only write the function based on my business logic

rest it be executed. So, it becomes too many technologies may be involved. So, add maintenance complexity and increases skill requirement for people working with the within the project right. So, that who are working in the project or who are at the maintaining the things required more remains skill full people right.

And at times there may be too many functions though that that already we discussed creation of function without reusing the existing one. So, non active server less function does not cost anything. So, there are a temptation that you keep create new function instead of alternating the existing function to match change requirements right. So, decrease maintainability and lower system understandability right.

So, what happened a running system if you are reusing. So, what happened that it goes for different change requires and go on updating. Now in this case if you if you see that so long the function is not instantiated. So, you executing. So, what we have we have the whatever it is not you are not paying for that right. So, while a new functions or new change request or something new is coming.

So, it is a there may be a tendency of updating this function not to update create a new function and deploy because you do not want to disturb the previous functions and type of things right. One may be that one psychology may be that that if sometime it is coming to use you can use. So, even too many functions comes into play and it creates a problem of maintainability and of course, also overall understandability of the system specially when you are hand over into some other things like that. So, these are some of the issues or faced by the server less computing in this paradigm. So, just to look at there are there are popular things there are many more.

So, like AWS have a what we say AWS lambda platform whereas, Google has a like Google cloud functions or cloud functions for Google in the Google cloud and Microsoft Azure is a Azure function. So, these are the popular things which are there and which and there are several others right means environment, but these are pretty popular things there are few open source stuff also that which you can use for this server less computing. So, Google if you see that we have virtual different virtual servers these are the functions and there are databases and also different type of storage services right which is the there ok. So, what we is we at what we try to see that what is the over overview of the server less computing paradigm how it helps say developers to develop and deploy their applications looking at the core business logic in a more efficient and first method and the is being build or charge for the instantiation of that function instead of the whole infrastructure what it has what it has taken down. So, with this let us conclude our discussion today there are few references some of the other things are taken from the internet maybe.

So, these references are you can have a look on the things. So, let us conclude our discussion today we will continue our discussion on server less computing in the next session. Thank you.

**Cloud Computing**  
**Prof. Soumya Kanti Ghosh**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**  
**Lecture 53**

**Serverless Computing - II**

Hello. So, we will continue our discussion on Cloud Computing or rather the different computing paradigm what we were discussing. If you recollect in the last class, we discussed about serverless. So, in this class, we will have a little more discussion on serverless computing specifically emphasizing on some of the popular serverless platform which are available at present and you can also use those you can basically handle on those platform to have small experimentation most of them give some free credits that is a provider. So, we will see those platforms right. What are the basic philosophy and what are the basic architecture though they are basic philosophy is same, but they may differ in some minor aspects.

So, what we discussed about is we will be discussing about the serverless computing little whatever we discussed little recap of the things and then primarily we will be taking up three things AWS lambda, Google Cloud Function and Azure Functions. So, which are serverless paradigm which are provided by Amazon AWS lambda is a Amazon is the service provider cloud service provider Google provides this Google Cloud Functions and Microsoft Azure comes up with this Azure Functions right. So, these are some of the popular there are several other things even there are few open source platform also, but these are popularly used and you can easily create a login and start using it though that will be chargeable, but they do have a some free credits. So, these are the things what we will be discussing.

So, just we let us continue our discussion little bit recap. So, if you recollect so, serverless computing primarily it hides the servers by providing programming abstraction for application builders that simplify cloud development making cloud software easier to write. So, what it is telling that a developer or a programmer is not bothered about the servers or server related configurations right. So, it is basically hiding the servers from it and then providing a by providing a programming abstraction for the application builders or application developers right. So, that this cloud level development becomes much easy.

This is becoming important these days because most of the cases we are moving towards this cloud based platforms right. So, that the huge purification of this cloud based application and cloud based platforms basically regards the need having a basically a platform where I do not have to bother about the backend configuration management and etcetera. So, here the focus or the target of the cloud computing if you see was system administration and system administration is one that to make that overall system administration is much easier type of things right. Like I configure a cloud computing and then I this basic system administration if I maintain then if people can leverage on those things right. Whereas, serverless basically targets towards the programmers.

So, they are basic focus is not the administrators they are the programmers right. So, in case of a cloud computing the if we look at that type of personal we are looking at the system administrator here the programmers right. There are lot of other features definitely. This change requires the cloud provider to take over many of the operational responsibilities need to run applications right. So, if you want to give a platform to the programmer.

So, that they can easily develop or programmer or developer or builder so to say. So, that they can easily develop their application. So, the responsibility of the cloud service provider is not only now providing VMs etcetera. Now, they have to give them operational responsibility of application something it sounds like PaaS right. So, they are also we try to provide this sort of a environment where they can develop their application etcetera.

Yes, that is there are overlapping, but we will see that we have seen that there are some differences. So, serverless computing to this in order to look at the change of focus from servers to applications right. So, the change of focus is from servers to applications this paradigm is known as serverless although remote server etcetera there right. So, in order to say that this is not some of the literature mention them as a server full computing like our traditional or that existing cloud computing framework as server full computing right. Though that those are that will create confusion.

So, we are not want to give names like that, but the serverless wants to emphasize that the now the focus is on the applications not the servers right. The this phase or this new phase of cloud computing or new paradigm or the variant of this cloud computing will change the way programmers work as dramatically as cloud computing change the operators works right. So, if you see or if you are you could have if you are managing your system administration of your or some sort of a system administrator of your organization and looking up a system etcetera. Now, once you move to the cloud the way things has to be done has changed right. So, similarly when we look at the serverless the way the programmers will work will change or programmers or coders or developers will work will change dramatically when we look at the serverless computing right.

So, thus serverless applications are one that do not need any server provisioning right, but ideally there is no need of server provisioning and do not require to manage servers right. I do not have to manage servers managing servers means managing their not only different type of configuration OS network etcetera. I we do not have to do that I need to focus only of my so called business logic for which for the logic of the application that is required here right. Now, in the last session or last class if you remember we discussed some of the major players right. So, three major players here AWS cloud Google cloud and Azure there are other things other still major players like IBM and so and so forth right.

There are many many players, but primarily we are these are some of the popular things

and they are they are a lot of programmers, lot of coders, developers use these platform that is why we taken up there is no specific reason for that you can the philosophy still remains the same right. So, if we see the virtual server so in case of a AWS so we say that as AWS instances or virtual machines and VM instances in case of a Azure. So, the function or what we say that cloud functions or function as a service, but where which basically enables the server less computing in case of a AWS we call is lambda AWS lambda service. In case of a Google Google cloud functions and in case of Microsoft Azure it is Azure functions right. And they do have their own databases though they can connect to other databases like here Amazon DynamoDB is their basic database which basically have a storage of a back to back storage of S3 Amazon S3.

Whereas, in case of Google they have a Google cloud data store and they Google do have a cloud storage and in case of a Azure, Azure have that Azure Cosmos DB right as we Azure Cosmos database and in the storage they have S3 they fall back to the S3 or they can utilize the S3 and type of things right. So, rather our main today's discussion hovers around this three functions right. So, we will see that the philosophically they are same because all are trying to provide or provision server less computing platform and, but there are there may be certain difference of their working path. So, we start with AWS lambda. So, AWS lambda is an event driven right server less computing platform provided by Amazon as a part of Amazon web service or AWS right.

So, it is a event driven server less computing platform right. So, thus if you look at you need to you need to worry about thus one need to worry about which Amazon resources to launch or how to manage actually you sorry there may be there is a typo that you do not need to worry about that AWS resources to launch or how to manage them instead you need to put code on the lambda and runs right. So, the somehow one that not is missing. So, one need not worry need not to worry about there that how this it will be configure where to launch and where the resources how to provision etcetera need to more worry about that logic of the code and just put the code on the lambda and it runs right. So, in AWS lambda the code is executed based on the responses of the events on AWS services such as.

So, as it is event driven. So, that AWS it basically the codes are triggers or codes are lambda code is executed based on the response of the events in the AWS services like some like it is add delete files in S3 bucket or HTTP request from Amazon API gateway. So, means there can be lot of type of things which are going on means events which can trigger right. So, however, Amazon lambda can only be used to execute background task as of now what we see that whenever. So, this background task come into things we want to execute those tasks right.

So, what we have seen it is a event driven serverless computing and some sort of a this function is executed in response to some of the some events right. So, that is another thing what we see it here right. So, it can be something in the S3 bucket or HTTP request of Amazon API gateway etcetera. Now if we look at that big picture or the block diagram and

there are components 1, 2, 3, 4, 5 different components marked we will see that what are the role. So, AWS lambda function helps you to focus on your core product right that means, or on your business logic and instead of managing operating system access, OS patching, right sizing, provisioning, scaling etcetera which I otherwise you need to bother about when you go for our traditional way of cloud computing right.

You or your that cloud administrator needs to bother about, but here you just put your code and it start running right. So, if we look at the picture so, one is that so, this is the overall block diagram we will see what are the different components. So, first thing you need to upload your AWS lambda code in any language supported by the AWS lambda that means, whatever the AWS lambda platform supports you can the code it can be Java, it can be Python, Go, C hash and other some of the languages there are supported right. So, these are which are popular language which are supported you can have other means other whichever the language supports by the AWS you can use it. So, these are some of the in the two if you see this the down layer or if you go up this one like Amazon S3, Amazon API gateway, DynamoDB, Amazon SNS right that Amazon kernels and etcetera different things are there you will see the different type of operations you can have that different AWS services which are there.

So, what it is there it is basically allow you to trigger AWS lambda that means, as we as SIG and that is event driven. So, this events are generated by this things right. So, AWS lambda helps you to load code. So, this AWS lambda platform code and the event details on which it should be triggered right. So, some like as we seen that two things we are discussing as we add delete files on S3 bucket things will be triggered or HP3 request, HTTP request which is from this Amazon API gateway that can be triggered and there can be other triggering things also right.

So, these are the AWS services which can trigger this the AWS lambda code which you have uploaded right. So, execute AWS lambda code when it is triggered by AWS services, AWS charges only when AWS lambda execute code. Now, the charging is only this whenever the this AWS lambda execute the code then the billing will be done not otherwise that means, whenever it is not triggered and not executing deciding it is not there otherwise right. So, it is not charged. So, if you go to this what are the different components which we see when we execute a code one is that or this AWS lambda component when you want to execute the code one is the function a function is a program or a script which runs in AWS lambda right.

So, lambda passes invocation events into your function which processes an event and returns to the response file returns its response right so that means, the code you upload. So, there is some event generated by your AWS services and that is good. So, there are some runtimes, runtimes allow function in various languages which runs on the same base execution environment right. So, it helps to configure your functions in runtime it also matches your selected programming languages right. So, it is the things whatever the programming languages you are using for the coding.

There are event source and event source is AWS service what we have seen that AWS in a SNS can be one simple notification service or custom service that this triggers function helps you to execute its logic ok. Lambda layers, lambda layers are important distribution mechanism for libraries, custom runtimes and other important function dependency these are the lambda layers. And you finally, need the log stream allows you to annotate your function code with custom logging statements and which helps you to analyze the execution flow and performance of the your AWS lambda code or functions right. So, that is the log stream for analysis purpose but we do. So, these are the different concept which comes with lambda.

If you see that many of these things are there in there, but in their nomenclature where you go to other type of service provider. And running or executing is straight forward if you have many if you have a login in the AWS is great otherwise you can create one login. And then basically sort of a edit your code and execute nothing is required everything is everything will be managed by the backend this your AWS lambda backend. Next one we would like to see this Google Cloud Function as we mentioned Google also provide the serverless computing. So, Google find function is a serverless execution environments for building and connecting cloud services right.

So, that is the environment provided. So, with Google Cloud Google functions you write simple single purpose functions right that are attached to events emitted from your cloud infrastructure and services right. So, this is this you can write this in the Google Cloud Function is triggered when an event being watched in the is being watched is fired. So, when again this is also once event is fired the cloud function is triggered right. So, the code execute in a fully managed environment there is no need to provision an infrastructure or worry about the managing the server.

So, you see now we are looking at a higher level than the at the server like I that more specifically on the code level right like whatever is request whatever the dependencies etcetera are taken care by this platforms that whether it is Amazon AWS lambda or Google Cloud Function or as Amazon sorry Azure function. So, these are taken care by those things. So, this is one very simple overall example or scenario of the functionality. So, we have cloud services this is the cloud Google Cloud platform cloud platform. So, which emits the events like in case of Amazon you have seen that AWS services right and various services like Google Cloud Storage, Google Cloud Publication and Subscription Pubsub Stackdriver, Google Data Store etcetera those are in that different cloud Google Cloud Services part right.

And they all have events that happen inside them for example, if a bucket has got into a new object uploaded into deleted from the metadata it being updated etcetera. So, those are the events which are being these are the event generator things and in case of the function that is the cloud function say an event is generated or fire or emitted the event data associated with the event has information on the basic event that is the metadata sort of

things event is associated with that event metadata for that event. If the cloud function is configured to be triggered by that event if your cloud function is configured that it will need to be triggered by that event then the cloud function is invoked or run or executed right based on that thing right. So, as a part of its execution the event data is passed to it. So, that it can decipher what had caused the event that is the event source get meta information for the event and so on and do this processing right.

So, this is at the part of that cloud function as a part of the processing it might also may be invoke other APIs like it can be a Google API or other external API right. So, while processing the things it is possible that you want to invoke other APIs right not only Google API is external API that is possible right. It can even write back to the cloud services that is another thing right. So, if required it is not a one way traffic it can write back to the cloud services if required right. So, when it is it has finished executing its logic that is your code the cloud function mentions it or specifies that it is done.

So, it is basically once it is done it flag that it is done. So, multiple event occur occurrences will result in multiple invocation of the cloud function if there are multiple events. So, multiple obviously, multiple invocations will be there this is all handled for you by the cloud functions infrastructure right. You focus on your logic instead of the functions and be a good means keeping your functions in a single purpose use minimal execution time etcetera. In other sense you manage your function in a such a way it is basically executed in a more efficient way do not bother about the environment much right.

So, it is only need to work on that type of programming environment which is supports and rest is taken care. So, it indicates that your model based in a stateless fashion right. So, if you look at so, things like it is a stateless fashion right. So, it is even when the event is there if your function to be triggered by the event then it is triggered right. So, rather you can maintain your state out of means outside this framework right.

So, you can maintain the state in outside the like in say memcached or somewhere you can maintain the states. So, what we see that events and the triggering. So, events if you see that they occur in cloud platform services for example, file uploaded to a storage a message published in a queue direct HTTP invocation etcetera this can be that several events and TIGERs you can choose to respond to a event via a TIGER right. So, whether you respond to a your functions. So, TIGER is the event plus as related with the event based on the event and the data or the metadata information with the events the TIGER can be fired.

Event data this is the data that is passed to your cloud function when the event TIGER results in your function execution. So, once it is executing so, the data to be passed to your function to that can that is the event data. So, even providers is there that all that what we have seen right different services right and that TIGERs and the data is passed to the cloud functions that Google cloud functions where your code are running. So, there are some of the event providers like one is the HTTP that is the one of the popular things invoke

functions directly via HTTP request right. So, that is the provider what we see pretty popular not only here in other context also there can be cloud storage cloud pub sub as we mentioned there is a firebase those who have used firebase you might have seen that is a database analytics authorized and all those things are maintained by this firebase stack driver login cloud fire store Google compute engine BigQuery.

So, these are the different cloud services Google cloud services which are code unquote event providers right. So, now finally, let us look at that Azure functions right. So, the it is a serverless solution that allows you to write less code maintain less infrastructure and save on cost right instead of worrying about the deploying and maintaining the servers the cloud infrastructure providers provides all the up to date resources need for needed for your applications. So, if you look at the basic philosophy remains same right. So, concentrate on your code and less maintain less infrastructure and less configuration issues and type of things right.

So, that is that is the bottom line right. So, user focuses on the piece of code and Azure like others serverless platform handles the rest of the things right. So, a function is a primary concept in Azure function right a function contains two important piece your code which can be written in a variety of languages and some configuration that like function dot JSON which are stored in the function JSON and things like that right. So, for compile languages this config file is generated automatically from the annotated annotations in your code for scripting language it must provide the configuration itself right. So, that depends that whether it is a compile language or a script what you are running right. So, Azure function helps you in build your own functions.

So, there are different resources are available like use your preferred language write functions in C hash, Java, JavaScript, Python etcetera etcetera or you can use a custom handler to use virtually any other language right. So, those this is there, but these are the things which are supported and maybe the more efficiently handled. Automate deployment from a tool based approach to using external pipelines there is a huge amount of deployment options available right good amount of deployment of options available. Troubleshoot a function use monitor tools and testing strategies to gain insights in your own app flexible pricing options with consumption with the consumption plan you only pay while your functions are running while premium and app service plans offer features of the specialized needs. So, that means, you pay when it is running same thing if you remember we discussed when we discussed about Google or Amazon same thing was there you pay only for the execution of the this AWS lambda when it is there right otherwise you are not charged for the things.

So, common serverless architecture plan or architecture pattern is that the serverless APIs mobile and backends right event and stream processing, IOT data processing, big data and machine learning pipelines right. And so, we can have different type of backup like serverless API or mobile and web backups or different in processing event and stream

processing right integration and enterprise service bus to connect line of business systems publish and subscribe to business events automation and digital transformation of the process automation and so and so forth right. So, it is so different type of architectural options are there different types of patterns are there right. Like if we look at some one two common scenarios like web app backend what we are looking at slightly tell scenario where we have request made in a web app request queued in the service bus and then go for this function a function process the request sends out the output to the Cosmos DB and things like that right or like mobile app backend thing backend where we have a financial service scenario right. Like we have HTTP API call from mobile app call processed by a function output data goes to this Cosmos DB and data transfer TIGR second functions right.

So, if you want to do a data transfer thing is TIGR functions which sends notification to the that notification hubs etcetera right. So, there can be several other scenarios where you have something where the event is generated and the this function is TIGR and then output are based on the type of output the either it is stored or it is means going for like here output stored in the Cosmos DB data transfer TIGR second function and type of things right. So, that based on that what sort of applications or what sort of serverless applications you want to run right. So, with this let us conclude our today's discussion. So, these are some of the references from where I have taken not only the this it is different figures for this our discussion purpose and also we have we that helps us in coming up with this discussion document.

So, what we you can go through these documents every provider have their own things and there are lot of other documents which you will get over the internet. So, let us conclude our discussion today's discussion here we will continue this looking at that different computing paradigm in the subsequent classes. Thank you.

**Cloud Computing**  
**Prof. Soumya Kanti Ghosh**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**

**Lecture 54**

**Sustainable Cloud Computing - I**

Hello. So, let us continue our discussion on Cloud Computing or we what we are discussing at different computing paradigms around this cloud computing concept right. Today we will discuss one aspects one which is coming up in a big way is that sustainable cloud computing right. So, we will see that what are the different aspects of this sustainable cloud computing and why it is important for consider as such right. So, as such when we talk about sustainable computing. So, what we try to mean that which the computing which is sustainable considering the different application, environmental constraints and in some cases even the social aspects also.

But if we confined with our basically computing and the environmental constraints. So, there are lot of things come up right specially we will see in the cloud computing right that energy dissipated that whether it is using renewable energy or not and still giving appropriate reliability and services those aspects will come into play. So, we will look into aspects like sustainable computing and sustainable cloud computing and this keywords like sustainable cloud computing, cloud data centers, energy management, carbon footprint all those things will come into play ok. So, let us start our discussion more specifically to cloud sustainable cloud computing.

So, what we have seen in this era of cloud computing or in this cloud computing paradigm that there are service providers we who have large data centers geographically spread and they is to give virtualized services right. It can be something at the highest level infrastructure as a service or it can be pass or platform as a service, software as a service, data as a service, sometimes databases as a service and different aspects of the things right we have seen in length and we are now used to those type of service enabled computing paradigm right. So, what happens this CSPs or the large service providers like if you consider Amazon, Google, IBM, Microsoft Azure and many others right. They basically run on their large data centers or sometime we call it CDCs right cloud data centers to support the user needs and different type of user demands both with respect to computing, infrastructure, application and things like that right. So, if you look at that this user needs are not only ever increasing they are having different type of nature right.

Some are means peak at some other some places some have peak in some other play other time and there is a lot of computing requirement type of things. In order to support that or in order to support reliably what one side what we see that there are SLAs like service level

agreements which need to be guaranteed right. On other side the cloud service provider need to providers need to minimize their energy requirements, their overall maintenance cost within the limits. So, their return on over investment is at the expected level or it able to maintain some level right. So, there are a there are constraint of servicing a user community which is very demanding and any fall in the service may be there are penalties will be enforced.

On the other hand need to minimize this energy requirements, minimize the overall carbon footprints and things like that right. So, the financial and carbon footprint related costs running such large infrastructure by the CSPs cloud service providers negatively impact the sustainability of the cloud services right. So, what we want to do? We want to sustain this cloud computing paradigm right. So, though positively is that more people are using or more user community that means, there is a good return on investment. On the other hand you have a problem of huge power consumption carbon footprint which goes against say environmental laws social laws and type of things not only that sometimes getting that type of energy requirement at peak time may be questionable right, maybe there may be constraint on the things.

So, those things actually negatively impact on the overall this sustainability issues right. How long it will sustain things like that right. You can even think of that if you have a larger if you have a system in your office or at even at home which requires a AC environment, requires a standard power supply is this uninterrupted or very regulated power supply and things like that or along with that network connectivity and so and so forth. So, in order to maintain and sustain that is a very challenging job right. So, we need to look at optimizing or taking care of all the different aspects.

So, focus on minimizing the energy consumption and carbon footprint and ensuring reliability in the of the CDCs right. That is one of the primary goal of means of means primary what we say goal of sustainable cloud computing or that is the what is the requirements for towards going for sustainable sustainability of the cloud computing aspects right. So, one is the definitely energy consumption and carbon footprint in order to do this the best way of one of means not may not be the best way the one of the standard way is that when your say servers are less loaded then you put it them in sleep mode right. So, when the request come it wake up and give services right it wakes up and give services. The problem is that if there are delay and things like that it may violate the SLAs of the thing right and if there is a simultaneous peak of number of users or good demand high demand for number of user then there may be a problem in attains means maintaining the SLAs with the SLAs with the users right.

So, this type of scenarios may be has need to be it is not a straight forward solution it need to be looked into in a much deeper fashion right. So, what we see cloud computing paradigm offers on demand subscription oriented services right over the internet to host applications and process user workload right. So, what it does it is a it is a basically on

demand and subscription oriented right you need to subscribe to this cloud service provider and things like that. So, ensure availability and reliability of the services the components of cloud com CDCs like such as network, storage, servers are to be made available round the clock. So, that is a requirement.

So, if we see the basic core components like they feed the computing resources is the first other components also like network devices, storage devices and even this environment itself like AC power supply all those things are to be available 24 cross 7 right. And if all those are available at their peak performance or at their full operational way then that may be a lot of consumption of energy where practically in number of times there may be a lot of consumption of energy practically no loading at that time right. So, creating processing storing each bit of data also add to energy cost right it increases carbon footprints and even what we say it warm up this system right. So, it requires some cooling things which again it is energy processing. So, like whatever we do with the processing one is that you supply energy for the processing and then in order to have with this reduced form factors and huge volume of computing resources like servers, network devices, communication devices other storage devices etcetera in a very compact form what we end up in dissipating lot of heat right we need to be cool down right.

So, we put again energy there right. So, there are a serious challenge in maintaining or sustaining this environment over large or over time right. So, how to sustain this over time? So, this is one scenario one projected energy consumption graph taken from this references what is maintained here. If you see the amount of energy consumed by these clouds data centers or CDCs is increasing regularly and expected to a something 8 k terawatt hours by 2030 that is a what it is looking for right where it is going for. So, it is a huge requirement of energy right.

So, in order to see if it is if we see that 2014 was near 0 that means, very less such that of a wide scale cloud requirements environments or cloud service providers available whereas, by 2030 it may go up to 8000 terawatt hours right. So, there is a pretty high requirement. So, as we as we are mentioning like if we look at the sustainable cloud computing. So, the some of the components are consuming good amount of energy right other than the servers components like network, storage, memory, cooling systems of the CDCs are consuming huge amount of energy right. So, other than the servers where virtual machines etcetera are host right.

So, to improve this energy efficiency so, overall energy efficiency CDC there is a need for energy aware resource management techniques right for management of all resources including servers, storage, memory, cooling system in a more holistic manner right. So, we require some of the energy management strategies or technique into the thing. So, due to the under loading or over loading of the infrastructure resources energy consumption CDC is not efficient right sometimes it is consuming energy, but you do not have a client to work on it or. So, you spend things without the return right. So, you can do you need to keep it

alive right.

So, it is not efficient in and what we see that in many cases energy consumed while some resources like network, storage, memory, processors are in idle state right still energy is in consumed because you cannot put them into say sleep mode or off mode and things like that right. So, which adds to the increasing overall cost to the cloud services like even if you see if you even if you have your own system means even we come out of this cloud paradigm and you think that even if your system at home or office it is how much amount it is being utilized right. If you there are several systems which are not being switched off right. So, 24 hours it is may not be you may be used for around 7, 8, 9, 10 hours right. So, 50 percent more than 50 percent of the time the system is in idle state right where neither the processor is used nor memory etcetera nothing is used and it is away the means it is away the power and power and energy and then you pay for it right.

And not only that not only you pay for it, it creates a requirement at the adding station to generate that part of the thing. So, in other sense the whole ecosystem is consuming more energy into the thing. So, CPCs are finding other alternative methods specifically something what we say renewable energy sources to reduce carbon footprint of their CDCs right. So, this cloud service providers are looking at alternative energy sources like carbon footprints and so forth to so that they can reduce the carbon footprint right of their data centers. So, this major players like Google, Amazon, Azure, Microsoft sorry IBM are all looking forward towards running data centers with renewable energy or something which is more thing is that more popular is the hybrid like you part of the thing renewable part of the things is generated energy things.

So, what it is seen that future CDCs are required to provide cloud services with minimal with minimum emissions of carbon footprints and heat release in the form of greenhouse gas emission and type of things. So, what it is seen that increasingly these CSPs like with future data centers or CDCs right. So, they will have need to minimize their carbon footprint and heat release that means, what we say know that there is a warming up or heat release from the whole process in form of greenhouse emission and things. So, now, you see to enable sustainable computing data centers can be relocated based on very difficult task relocating data centers because this has a huge installation overhead along with requirement like specific environment and data center again are distributed over geographical spread. So, that they do not fall in the same power supply line or in the same network or even sense seismic zone type of things.

So, to enable sustainable cloud computing data centers can be relocated based on that opportunity of waste heat recovery type of things like whatever the heat is being lost whether you can recover and replace something accessibility to green resources right instead of all our standard energy sources renewable energy sources and things and also we try to make it to proximity to cooling resources. So, that the cooling that after cooling that cool air if it is passing through a long pipe and long duct. So, it will basically lose its

temperature or in other sense it will be go on means dissipated and things like that means, actually it will go high and then the cooling effect will be reduced right. So, it is near the cooling stations. So, idea of the data centers are let it be near cooling stations, let it be near power stations.

So, that the overall power to be transmitted is with over a less distance and things like that. Now if we look at a big picture or what we say that what are the different components or what are the different modules so to say of a typical sustainable cloud computing platform right. So, this is a thing which has been design which has been you will find in this particular reference. So, it has been taken from that reference. So, nevertheless if we see that one is that the cloud architecture right which comprises of is the standard cloud model IAS, PAS or SAS right.

So, these are the standard cloud model which are being they are which is definitely there right and the other three modules are one is remote CDCs right. So, there is a concept of remote CDC that means, the CDCs are at means at the remote locations and this data centers. There are cooling manager right which are taking care of the cooling aspect of the things and there is a power manager right. So, how much power say starting from how much power to be over this renewable this from the renewable sources how much power for from the fossil fuel or the standard energy source state what will be the switching mechanisms of the things and things like that. So, these are the four major component for a typical sustainable cloud computing installation or when we have the sustainable cloud computing paradigm or want to look into the as the aspects of sustainable cloud computing.

So, what we see that conceptual model this is the what we looked as the conceptual model. The conceptual model for sustainable cloud computing is in the form of layered architecture right which offer holistic management of cloud computing resources to make cloud computing more energy efficient and sustainable right. So, you what it tries? It tries to do a overall management of this cloud computing resources. So, that the this overall your cloud computing paradigm is sustainable and energy efficient. So, first one I have seen that cloud architecture component divided into three as, as, as already we know that.

So, that one is infrastructure one is the platform and one is the software as a service right. So, this is one component which is the major thing for the cloud and the other component is the cooling manager where the where which basically gives thermal alerts if temperature is higher than the threshold value and so on so forth. So, this takes care about the this cooling aspects right. So, about the threshold value heat controller will take an action to control the temperature and minimal impact on the performance of the CDC. So, what happened as this these CDCs are packed with servers and other components like storage, networks and other the different other components.

So, this if it is not properly cooled. So, it the there may be a chance of failure, there may be under performance of the things right. So, this are pretty much required this cooling

manager. So, other aspects is the power manager which takes care of the overall management of the power, it controls the power generated from the renewal renewable energy sources like it may be from say sun or some other type of sources like we can have those type of renewable sources and the fossil fluid like grid electricity and things like that right. So, if there is execution of the deadline oriented workload, then this grid energy can be used to maintain the reliability of the cloud services right.

So, what we say if there is a there is a stringent requirement of the workload. So, then we cannot rely we want to rely more on the whatever energy available in a normal fashion that through the grid and type of things right. Whereas, if there is that no such requirement, then we can fall back to this our this renewable sources like one may be your as we are telling that solar cells other may be something we windmill and type of things right. So, those are all green sources green energy sources right. So, power distribution units is used to transfer electricity to all the things.

So, what it happen is automatic transfer switch ATS be used to manage the energy coming from the both the sources right and power distribution unit transfer the electricity to all the series in the cooling devices and things like that right. So, remote series is virtual machines and workloads can be migrated to remote series is to balance the load effectively right. What happen once you are getting more load on a particular server, you transfer the load some of the idle server which get some workloads and overall this loading on that this what we say that so called parent VMs or parent servers is reduced right. So, this balancing may help in reducing the overall power requirements sorry. So, what we see this there are issues of reliability and sustainability right.

Suppose I ideally if I switch off when nobody I do not find anybody then switch on switching on may take lot of time and then we end up in end up in losing SLS. So, that is another problem. So, there is one consideration of energy to reduce energy consumption of the cloud center to reduce under loading and overloading resources which improves the load balancing to minimize heat concentration and dissipation of the cloud centers right. So, these are the with the respect of the energy to reduce the carbon footprints to improve bandwidth and computing capacities to improve storage management in the disk etcetera this is with respect to energy. Now, if you look at the from the reliability point of view definitely SLA violation is one of the things where the CSPs wants to avoid to identify system failures and their reason to manage these the risk right to protect critical information for security attacks and things to provide secure VM migrations like whenever you are migrating VMs and to reduce the turn on investment suppose you put it on sleep and then you turn on there are challenges right.

So, right. So, there are there is a balance between this energy optimization and maintaining the reliability at the basic level or above the threshold right. So, there is a serious challenge into the thing. So, implication on reliability and sustainability. So, improve energy utilizations which reduces electricity bills operational cost right. So, that is with respect to

the lead to sustainability.

So, you have a lower footprint less bills and things. So, in order to provide reliable cloud services the business operations of different cloud provider are replicating services which need additional resources increase energy consumption and things like that. Now if you have a high demand. So, you have a replicated server more replicated services and then it increases more demand on energy.

So, these are something which goes conflicting. So, we need a tradeoff between this energy consumption and reliability to provide cost efficient services right. So, that there should be a tradeoff between the things. So, if you look at existing energy efficient resource management techniques consume huge amount of energy while executing workloads right which decreases the resource release from the clouds data center presently or many of the data centers which are not looking are energy concern things like that. So, dynamic voltage and frequency scaling as we popularly known as D V F S based energy management techniques reduce energy consumption, but the response time and service delay due to switching of resources between high and low scaling things may affect adversely on maintaining the what we say reliability of the service or maintaining the SLA. So, what we see the reliability of the system component is also affected by excessive turning off and turning off right.

So, if you want to turn off when it is idle state and turn on when the request is there. So, there may be a chance that you take time in turning on session and it is the reliability is challenging. Power modulation decreases the reliability of server components like storage device etcetera those are the challenges. So, reducing by reducing energy consumption CDC we can improve the resource utilization, reliability and performance of the servers by having appropriate mechanism of taking into the reducing the things. So, thus we see that there is a need of energy aware resource management right.

So, I am doing resource management energy aware resource management techniques to reduce power consumption without affecting the reliability that is the primary goal as we were discussing that. So, that I cannot as a service provider as a CSP I cannot afford to have poor power management things right or what we and also on the other hand I cannot afford to means afford to say violate the SLAs. So, if we look at the sustainability of the cloud computing or sustainable cloud computing. So, there are several components right one if you see these are the major components or I should say different aspects of the thing one is the application model right one that VM or virtualization issues there are waste heat utilization things like thermal aware scheduling, renewable energy, resource targeted energy management right. So, energy we are managing energy, but at the resource target it is not a general energy management.

So, say if some resources idle whether I can do more finer grain energy management and capacity planning right you need to plan a priori that capacity that what you want to

what you want to means based on your expected return that how much consumption, what should be the your data center size or CDC sizes and how what will be the growth rate of the things user demand and accordingly the CDC accordingly how your server provisioning etcetera to be done right. So, those are the aspects of the things. So, what we see there are different components which plays a important role in shaping this achievement right. So, rather what we discussed today what we discussed today is basically looking at that in order to achieve sustainable cloud computing. So, what are the different aspects we want to look at right.

So, two one is that the performances right how to say support the SLAs, how to have a minimum say SLA violation a high level of reliability while catering to the customer to the customer or the user. On the other hand how to appropriately manage the resources in terms of energy, carbon footprint and things like that right. So, this is there is a we need to we need to look into this tradeoff between these two components right. So, whether without compromising the reliability how we can look into the things like this low power consumption, less carbon footprint, utilization of more energy more renewable energy and things like that what we are seeing that waste heat utilization or waste heat management when you have your large C D C's there are lot of heat generated which need to be managed right. Either it should support the existing system in some time or in some form or the heat need to be transferred and utilized in some other fashion right.

So, and also the thermal aware scheduling of the job. So, how I job suppose I have a say n number of servers and if my workload can be run into a m number of servers which is m is less than n then whether it is wise to distribute the things into a into number of servers in a in a some particular fashion or it is more concentrate into that minimal number of servers. So, that this my consumption of energy etcetera are there. So, to one is requirement of how much energy wise how you are optimized other requirement is that say heat generation or thermal aware scheduling of this job and workloads and of course, on the other side is the SLA's and reliability of the service provider is there is has to be maintained right ok. So, with this big picture let us end our discussion today and we will continue from this slide itself in our subsequent or in our next lecture right looking at few more aspects of this sustainable cloud computing paradigm. Thank you.

**Cloud Computing**  
**Prof. Soumya Kanti Ghosh**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**

**Lecture 55**  
**Sustainable Cloud Computing - II**

Hello, we will continue our discussion on this Cloud Computing. More specifically we are last class we are discussing about Sustainable Cloud Computing. So, we will continue our discussion on Sustainable Cloud Computing right. So, what we see the concept of primarily the sustainable cloud computing thing and along with that we will look at that what are the different components taxonomy for needed for this sustainable cloud computing paradigm or sustainable cloud computing as per means model of the this cloud computing. So, this is we are continuing from the last class just to start with a quick recap what we have seen in the last class. This CSP is what we see that heavily rely on this CDCs or the data centers which are there to cater this ever increasing need of the user.

Especially in today's world with so many or a millions and billions of IOT devices are in place and so this overall computing requirement is increasing exponentially we may say right. So, to cater to this, this cloud service providers are increasing their capacity like in terms of increasing number of their distributed data center, packing more servers into the data centers and which in turn lead to more heat generation, more power requirement which in turn have a more cooling mechanisms into the place right. So, all those things brings about a problem or a challenge for the sustainability of the computing system in our case cloud computing system right. If you see in a more holistic manner, so if you see that over years or over decades this the overall size of this our chips and semiconductor devices has reduced to a great extent whereas, the computational power has increased right.

In other terms if we see that other than there is a limit for the particular say technology like VLSI technology or whatever you look at it that how much compactness or how much packing you can done. On the other hand once you make this type of very compact circuits is generated tremendous amount of heat right which need to be dissipated right. So, that also in if we see in the thermodynamics term I am not a no where expert in that area, but in, but intuitively we see that that is also a limiting factor the how much heat you can tolerate right. So, if you see that say GPUs or HPC type of architecture, so there is a enormous amount of this so called cooling mechanism heat sinks which is required to cool the systems, so that it can perform at its optimal level right. So, rather in modern day data centers specially housing large servers or HPC type of servers what we see that the somewhere the cost of the this IT infrastructure and cost of this environment like heating mechanism power sorry cooling mechanism power and say environmental other control things is basically leading to the same cost like if it is cost x this IP infrastructure this is somewhere it is if you look at it is another x type of things right.

So, that means, in order to make this sustain sustainable or in order to this cloud computing sustainable what we want to look at we need to look at all the factors right that starting from this computing why is how much how what is the how much we can go on increasing along with that how to handle this say cooling mechanism what to do with this what we say this excess heat or sometimes what is referred to as waste heat right what to do with this waste heat thing and there are definitely things like power and type like that right. So, the way we take whole this whole mechanism like computing plus if we say this per say this overall environment to handle the whole system into together the making it sustainable is always a challenging job. So, what we see that financial and carbon footprint related to the cost of running of such infrastructure negatively impact the sustainability right. So, at some point it will basically say that this is the you need you cannot cross this things right like if I have a say 10 feet cross 10 feet a room or maybe a particular size x x cross y a room with SCE of so much SCE and things like that. So, the number of server even the space allowed is something related with the how much heat generated and cooling is there it is not only that how was your space or power allowed also the how much heat it is generating and how whether there is a way to cool it down and type of things that is one aspects.

If I say that the cooling mechanism things are there then means actually all these factors this cooling mechanisms heat generations power computing they need to be looked into a more integrated fashion right what this sustainable cloud computing or sustainable computing in general tries to address this. So, the focus is to on minimizing the energy consumption and carbon footprint and ensuring reliability of the these cloud data centers. So, this is the goal or one of the major objectives of the sustainable cloud computing. Now if we look at this picture also last class we have seen. So, this is a typical model the conceptual model proposed by this in this particular reference in this particular journal or paper.

So, what we see one side this cloud architecture is there where we have IAS, PAS and SAS right. On other things what we have other models other modules right one is that this power manager which takes care of this power and once we talk about the power it can be from our conventional sources right which takes which basically based on this fossil fuel etcetera or we can have renewable sources right like it may be solar it can be wind things like that right. So, that is the things and once we have this type of stuff then there are lot of things that when to switch to conventional to renewable and things like that and there are pros gone in running the IP infrastructure at the in the whole stuff like initially we may start with the conventional sources then switch with to that renewable sources and if required come back based on the load and type of things right. So, that is another thing is a cooling manager like temperature control, heat exchanger and type of things which takes care of that so called air conditioning infrastructure and there are remote CDCs right where that data centers there are remote data centers where we may require to given all those things constant power and cooling mechanisms other things constant. So, I can I can have with the load increasing or the user load increase I can basically migrate the VM from one to another

and overall say power requirement or heat dissipation I can we can manage right whether that is a possible right that type of thing.

So, if there are remote CDCs which to migrate and migrate what should be my migration plan and things like that will come into play. So, all those things should hand work hand in hand and if you look at that pass thing on the left side of the things there are there are definitely other pass things other than there are things for remote CDC manager, VM resource manager, green resource manager, thermal profiling, cooling manager these are different type of applications which are running at the infrastructure itself to have a more holistic approach. Now, this also we started at the end we basically shown this. So, if you look at that overall what are the different components or what are the different aspects of this CDSDC, SSCC rather sustainable cloud computing. So, one will is the application model right the what sort of application how we have tuned the application the same application may be more energy friendly if you if you have a more optimized way of planning it.

There are things for virtualization that because of virtualization is one of the major aspects of cloud computing. So, how you handle that whether VM migration, storage migration things like that there are things for waste heat utilization there are waste heat because which are heat generated either you throw it out whether there is a possibility of using this heat right like there are concept of vapour evaporation right once you evaporate then basically it cools down and things like that whether it is can be used for this type of things thermal aware scheduling right. So, scheduling mechanism architecture which are thermal aware right. So, whatever architecture the scheduling whatever we are following in our mechanism whether these are thermal aware like the one is that doing efficiency in different terms and other way looking at it that whether it can be thermal aware. There can be renewable energy which plays a big role like energy source, storage device, switch location etcetera and these are the renewable energy there can be renewable energy and once renewable energy is there.

So, we require some sort of energy management not only renewable energy any energy staff we require a energy management right. So, energy management to the in first IT infrastructure itself like processor, processor, memory, storage or the cooling mechanism or network or even when to switch from renewable to the our standard energy sources and things like that right whether if the load is increased whether it is good to have work on with our standard energy sources which are which get energy from this fossil fuel etcetera which may be which may be more quickly to deliver power or renewable energy that those things are to be worked out. And of course, make the make this whole so run properly we need to do a capacity planning right that is required for any things because now if you see this cloud computing it is not only putting some infrastructure and state is set it runs. So, it is much more than that why we are doing that? So, that we can make this sort of infrastructure sustainable right. So, capacity planning so, power infrastructure IT sources workloads based on that we need to plan the capacity things right not only that we

need to have predictive models like what we are heading to if this is my requirement down the line if we going to next stage of things whether the whether there is requirement increases and what sort of requirements are there and things like that.

So, that is that is important for capacity planning and doing for looking for the in something which is something futuristic manner. So, as we discussed so, application model the application in this the application model plays a vital role and the efficient structure of the application can improve the energy efficiency of the cloud data centers right. Application model can be data parallel, function parallel, message parallel the what way we are trying to look at whether it may be data parallel it can be function parallel that means parallelism in executing the function or it can be message passing and things like that. So, based on type of application we can have different application model and this model can be what we say energy aware or thermal aware type of things right how much consumption of the energy how much heat generation things like that. Another aspect is the energy management as we are talking about.

So, energy consumption of the processor memory storage cooling of the things as some from some reported work as the document we are following I have shown the reference in the previous slide. So, what some study shows that for processor it is 45 percent next comes the cooling a 20 percent and this memory storage networking these are the 15 10 10 type of things right. So, we need to look into this when we look for the energy management who is basically main tickers or the consumer of the things appropriately we need to look into the things right 10 percent cooling requirement if we can reduce it may reduce more than the same 10 percent if we want to reduce for the storage etcetera right. So, power regulation approaches increase the energy consumption during workload execution which affects the resource utilization of the series right. So, if we have a power regulations across the thing without so, when at the peak time the overall performance may be affected right.

So, there are things like we have last class also we discussed that DVFS attempts to solve the problem of resource utilization, but switching of resources between high scaling and low scaling modes increases the response time and the service delivery delay right. So, too many switching may have effect on the SLA per se right. So, again putting server in sleep mode then wake up things when demands come also is a challenging thus improving energy efficiency in looking at different aspects may have a better way of sustainability and keeping the resource utilization reliability performance of the server in a within the threshold or in a optimal way. So, thermal aware scheduling as we discussed. So, component of thermal aware scheduling are the architecture and scheduling mechanisms right.

So, can be different architecture single core multi core. So, heating problem during the execution of the workload reduces the efficiency of the cloud data centers right. To solve this heating problem of the CTC thermal scheduling is designed right. So, what we see that it present day thermal aware scheduling or whatever used is mostly focused on reducing

power uses efficiency right. And but PUE reduction may not reduce the total cost of ownership always anyway that is another stuff to look at, but nevertheless thermal aware scheduling is again in need and becoming popular for sustaining this cloud infrastructure.

Virtualization because the whole cloud is primarily concept is based on virtualization. So, during execution VM migration perform to balance the load etcetera right due to lack of one side renewable energy workload can be transferred here. So, this overall now once you transfer the VM from one to another. So, there are lot of implication to that right we have seen that how VM migration can be looked into. So, though this is also one major component to in to be looked into in realization of the sustainable cloud computing.

Capacity planning as we discussed this cloud service provider must innovate must involve an efficient and organized capacity planning mechanism right to attain expected return on investment. Because the if you see the CSPs point of view. So, they want to look a whole thing as a investment right one major component is the investment and there is a return on investment. There are definitely SLA violations to be avoided customer satisfaction and customer what we say likingness for the infrastructure or the for the CSPs also all those factors, but nevertheless there should be a return on investment right that whatever is there. And investment in most of the cases are has to be done upfront and the utilization slowly grows means grows up right or slowly the growth utilization slowly increases and things like that.

So, they have to need appropriate planning or appropriate capacity planning not in terms of hard resources also in terms of that what we say regular resources right like power, how to handle this waste energy or waste heat, what should be the cooling mechanism. So, it should be more adaptive and things like that and we need there is a need to do a full-fledged capacity planning. So, there is a need of effective capacity planning for data storage and the processing also right. So, it is not only the computing things like if we when we talk about IT infrastructure we want to include storage, networks and things like that. Renewable energy as we know that either solar, wind and things like that.

So, energy storage device and location and location like whether it is offsite, onsite, how much with and those things becomes important. So, carbon uses efficiency CUI can be reduced by adding more renewable energy, but there are a flip side that renewable energy getting when there is a huge demand whether my this renewable energy source can handle this demand. So, it need to be looked into that how to switch between the renewable energy and our standard energy sources and things like that right. So, cloud CDCs are required to place near the renewable resources to make cost effectiveness and if you see different renewable energy sources is required say large span of area whether it will look at say solar power thing or solar cells. So, it is it requires a sufficient space or even wind energy where you required need to place those all those wind in a different places.

And wastage utilization the cooling mechanism of the heat transfer model plays an

important role to utilize the wastage things right. So, what is happening that lot of heats are being generated by the infrastructure. Now this heat either this what in our standard practice the cooling mechanisms basically support this heating thing basically not I should not say support basically take care of this heating mechanism and cool it down whether there is a way to utilize this waste heat right. So, this vapor absorption based free cooling techniques can help in reducing the cooling expenses right. So, there are different techniques which are used.

So, energy efficiency CDC can be improved by reducing the energy using the cooling. So, it can be if the waste heat is utilized in somewhere other than it may help in better cooling mechanism. So, now we with this basic modeling concept we try to see there in another taxonomy the overall taxonomy of the sustainable cloud computing and with respect to the mostly with respect to the this different aspects and components of the features what we just discussed right. So, needless to say that which huge clarification of IOT devices and means extreme uses of or this our IT infrastructure for different other purpose like several location based service or mapping services, video conferencing, online feeding of live streaming of different events and games etcetera. So, it became a huge need of computation and there is a with cloud computing in place there is a there is a obvious fall back to this cloud computing infrastructure right.

So, in order to make it sustainable right. So, this cloud computing must be energy efficient and sustainable to fulfill the ever increasing user needs right. So, there is a ever increasing needs of the user in this cloud computing should be sustainable. So, research initiative on sustainable cloud computing can be categorized as follows one is that application design, sustainability metric what are the different sustainability metrics are there starting from its something which is performance centric to cooling to power to and different aspects of the things right. Capacity planning as we are discussing energy management, virtualization, thermal aware scheduling, cooling mechanism, renewable energy and waste heat energy utilization right.

These are all important because most of this CSPs and data centers are working around the clock right. So, it never stops and the though the user need may be the demand may vary, but the system need to adapt with the things right. There are there is a need of predictive model that what we are going to get in future and plan accordingly. So, next couple of slide we look into the different taxonomy again presented in that reference which is given below right. So, those are taken from there I am referring that and not only that I rather encourage all of you to have a read on this particular paper or rather whatever in the references also we mentioned I mentioned.

So, if you look at the application design. So, there are distinct component one is that looking at that QoS mechanisms that overall application model, work load type, architecture and there are other components like if you quiz time, cost aspect, work load management like what are the different critical in like critical interactive mechanisms or there are batch

style or what there are different type of mechanisms right. So, what we try to look at the design of an application plays a vital role and effective structure of a of there is a small typo of an application can improve the energy efficiency right. And different type of application may need different type of what we say designing right. Say something a parallel computing type of application may be different when we have something called a streaming application and things like that right.

So, the resource manager and the scheduler follow different approaches for application modeling. So, the resource manager and the scheduler follow different approaches for application modeling to make infrastructure sustainable and environmentally eco friendly there is a need for green ICT based innovative applications right. So, we need to look at more applications models or designs which can cater to this means energy efficient or thermal aware type of things. Next one is the what we want to discuss is the sustainability metric there is a big one right. So, one part if you see that more of the performance related metric which can be related to the cloud provided like throughput, application, security and so and so forth.

Here is the cloud user satisfaction, response time, correctness, reliability etcetera what the user is looking for right. So, this is the performance and other things like we have at the bottom mostly that heat waste heat utilization, cooling mechanism right, thermal aware scheduling metric, energy management metrics and these are the things right. Whereas, there are issues of virtualization metric like which looks for that different virtualization techniques and things what should be the say if we have any virtualization any migration issues or other things how need to be handled like that. There is a thing for capacity planning like we already discussed like looking at the capacity and also predicting that say uses of the infrastructure in terms of memory, in terms of storage, in terms of CPU, GPUs and things like that and there are renewable energy matrices for bringing down the carbon footprint and so and so forth. So, these are the different sustainable metric which you need to keep in mind right.

So, once we employ or deploy some mechanisms which are which support the sustainable cloud computing you need to look at that that how effectiveness of that mechanisms by using looking at the different metrics right one or more of the metrics which are there. Capacity planning as we discussed is important things like with respect to what are the different components like IT, cooling, power infrastructure etcetera with respect to IT workload, application model, things like auto scaling, utility, different utility functions those are things which are required for this capacity planning right. Then, we have energy management things like as we discussed it can be overall management of the energy. So, energy a sustainable is an important factor because the energy need to be as at a optimal level. So, that the performance are not disturbed and SLAs are not violated or and use the minimum energy right.

So, improving energy use reduces electricity bills and operational cost essential

requirement of sustainable computing are optimal software system design, optimized air ventilation and installing temperature monitoring tools for adequate resource utilization which overall may increase the energy efficiency. So, if you see there are things called static way of management static management which mostly of the system level and the CPU level and there are lot of dynamic things like resource management, resource consolidation, configuration components which looked into the different aspects of energy management, but nevertheless is one of the major factor. Virtualization we have already discussed is an important part for the for not only for the sustainable of cloud computing or CDCs, but it is the core of the philosophy of cloud computing right. So, this once we have this virtual machines if there is a virtual migration, VM migration, VM elasticity, VM load balancing, VM consolidation, fault tolerant and scheduling all those aspects need to be made in an efficient way. So, that the overall energy requirement can be minimized thermal heating can be minimized and so forth if including waste heat type of things right.

Thermal aware scheduling as mentioned earlier. So, consist of a the if you look at a CDC or any data any for that matter any data centers or any rack. So, it has a chassis racks to replace the servers to process the IT workload right. To maintain temperature of the data centers right server produce heat during the execution of the IT workload the processor is an important component of the server and consume most electricity right processor things. But both cooling and what we have seen few slide before the cooling mechanisms or other type of computing mechanisms also take or other IT infrastructure also have a huge amount of this power generation sorry power requirement and thermal heat generations are there. So, there are different aspects like architecture, heat modeling, thermometer, scheduling, monitoring ionization, simulations and things like that.

Then we have the cooling management. So, the increasing demand for computation networking storage expands the complexity right size and the energy density of CDC's exponentially which consumes a large amount of energy and reduces the huge amount of heat right. So, we have appropriate cooling mechanisms which need to be in place in order that the everything works faithfully. To make CDC's more energy efficient and sustainable we need an efficient cooling management system which maintain the temperature of CDC's right. And then we have the renewable energy right different sources like it can be solar, wind or in sometimes like hydropower what are generated energy which has a near zero carbon dioxide emission things like that that can be employed. And of course, if we need to these then we need to have appropriate workload schedule there.

So, what switch sort of workload to be scheduled, what are the different sources on energy, whether it should be location aware, whether we can have this storage device, how to maintain with this renewable energy and things like that. Then finally, we are having this waste heat utilization right. So, reuse of waste heat is becoming a solution for fulfilling energy demand in the energy conservation system. So, the waste heat means heat generated by the system itself which are cooled down, but if we have a mechanism of

utilizing this, reusing this waste heat then we efficiently can basically overall what we say like overall this green computing aspect we may put more emphasis on this right. So, that one way of what is seen in the literature people are trying for vapor absorption based cooling system as we are discussing in the beginning of the thing can used waste heat and remove the heat while evaporating right.

So, vapor absorption based free cooling mechanisms can make the value of PUE ideal by neutralizing the cooling expenses. So, it will have a not only monetary wise, but also you can have utilization of the things right. So, if we look at the overall sustainable cloud computing. So, this ever increasing demand with IOTs and different type of sensor a very findable sensor even this huge proliferation of this drone level drone operations and capturing different images and other things. So, we get a huge volume of data right apart from that there are streaming data and things like that.

So, we need to have more efficient. So, which which dictates this some sort of a ever increasing form of cloud infrastructure which is may not be feasible up to a limit. So, what we need to do we need to have this all this component together and make a sustainable model right. So, that it sustain over time span. So, this what we see this next generation or the today's cloud computing must be energy efficient and sustainable to fulfill this. Sustainable computing CDCs are powered by renewable energy sources kept more towards a means this near the power generation things may be more in a cool more cool environment right.

There are other issues of environmental issues right. So, that that overall carbon footprint is reduced right. So, sustainability with high performance and reliability is one of the primary goal of this cloud computing or today's cloud computing infrastructure. So, with this let us discuss in our discussion today there are very nice references some of them I have put here. So, I encourage you to go to this references that will be most of the things we basically taken from this different references, but I think that reading the differences will be good for you. Thank you.

**Cloud Computing**  
**Prof. Soumya Kanti Ghosh**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**

**Lecture 56**

**Cloud Computing in 5G Era**

Hello. So, we will continue our discussion on cloud computing rather we were looking up going into the aspects of what new things are coming up or which came up in last couple of years. So, one aspect one as we all are all know or all getting used to it that that 5G is on the card actually it is already start rolling and in our country also formally things may come up in this country in a in this year or so in a big way right. So, we will have some a overview or a discussion on that how this cloud computing will be seen with 5G in place right. So, or I can say that cloud computing in 5G era what way it is beneficial, what are the different impacts on cloud computing whatever we are doing on 5G with the this deployment or large scale deployment of 5G right. So, that we will try to see definitely it is a 5G is a big area to discuss and if you see or 5G and beyond.

So, what we can see that it is a itself is a separate maybe couple of forces. So, here we are not going into the 5G communication per say or what we are trying to look at that keeping that view in the 5G network whether this cloud computing make sense to in some aspects of the things right. So, we will look into the 5G network and little bit of some aspects of 5G network which will be needed out here and then cloud computing in 5G. So, comes more specifically like when we talk about mobile edge computing or per say edge computing itself and mobile cloud computing in 5G what are the different aspects right.

Again this is a very very high level overview what we are looking at ok. So, if we recollect this picture or rather when we discussing from cloud to things continuum right with some case studies. So, what we see that the at the edge of at the edge of the or at the edge of the network or that here is the IoT devices right and at the top is more of the is the cloud that is the this at the level 0 or something what we mentioned it in between there are fog layers and there is a edge layer right. So, our basic intention was to bring the computing down the line to the to the edge as far as possible right. So, why we are trying to do that because first of all cloud is a something what we say infinite resource in terms of computing storage and type of things right what we look at it thinking that the power of cloud will be immense right.

The challenges was that any data travelling from this to this IoT or this sensors or this down the line to the what we say that devices to the cloud and come back if there is a sensor come actuators are there then it say the major challenge is the data load one side another challenge is the latency right. So, if huge billions of IoT devices which are increasingly being deployed in every field and every domain right. So, what we see that

there is a huge volume of data being generated and once you want to push them to the cloud for any type of so called analysis or like it is looked into business analysis analytics and intelligence. That means, any type of decent type of things then you have to need to travel this data there and if there is a actuating involved then it may come back to the device level or it may travel to other type of things right. So, what is happening there is a round trip delay which is the latency is pretty high and most of the cases we need to rely on this mobile networks because that the cloud is somewhere else and you are this devices are like some sort of a having some wireless capability and so once bringing down that at the fog layer I can do computing or at the edge layer one computing things are there right one level of computing.

So, I can basically the round trip delay etcetera. So, that was one challenge latency and data load other thing other challenge what we have seen that the mobility of this devices right it can be the sensors which are there or the what we sometimes call as user equipments or user end devices which may be mobile the edge devices may be mobile like if you are using your mobile phone in the edge and if we are if there is may be a mobile even the fog can be mobile or if something usually fogs are what we look at it is more of a stationary things, but nevertheless I can switch from one fog device to another fog device and things like that. But so if we look at if we look at that when we bring down this cloud which becomes a small part of the cloud or what we say some sort of a if we can think some sort of a mini clouds in the things. So, there is a mobility also come into play right. So, what we look at things like one side this latency, next is your massive data and other side is the mobility this is the aspects what we look at.

Other things what the challenge is coming out is that what we say more of real time applications right it can be a gaming applications, it can be something a healthcare applications, it can be different type of say traffic applications like where it is real time or near real time right where the response time matters though this latency etcetera are at the back end are challenges definitely, but this real time applications demands high availability of the things right that is or quick response of the things right. So, this is one of the challenges and of course, this different level of QoS when you are having different type of traffic it also have a lot of things. So, what we see this if we look in a overall things that this cloud do have challenges in say going to its next step or expanding in a particular fashion is one is this network delay, data load, how to handle real time things, mobility aspects, maintaining QoS and of course, security and other things are also coming into play right. So, keeping this mind we just see that why what this 5G a very overview type of things right. So, 5G what we see is the fifth generation mobile network.

So, it is a it is becoming a new global wireless standard after 1G, 2G, 3G, 4G networks right. So, what 5G enables a new kind of network that is designed to connect virtually everything everywhere everyone together including machine, objects, devices etcetera. So, it is something which enable this high availability and connectivity between anything which can connect right. So, 5G wireless technology is meant to deliver higher multi GPS peak data

that is huge high very very high data rate, data speeds, ultra low latency that means, very low latency, more reliability, massive network capacity, increased availability and more uniform user experience to more users right. That means, what we mean to say that it tries to address the aspects what are the challenging what we are having whether with cloud or without cloud right for of that huge data rate.

This your latency very very ultra low latency, more reliability into the things, massive network capacity that means, some sort of the network itself is a high capacity something what we can say that say network as a service type of things or as a platform type of thing. Increased availability and more uniform user experience that means, it is it is it maintains some sort of a quality of service based on the this based on the traffic or based on the application type of things. And improved efficiency and power new users experience and connect new industries and type of things right. So, these are the things which promises which 5G promises right. So, if you look at the quickly the generation of the things now why we are talking all those things because our baseline of this cloud computing or success of the cloud computing depend that how it communicates right.

If you remember that at the beginning lectures what we started with and every iterated maybe every time or every other lecture that we require a ubiquitous connectivity right. Unless this ubiquitous connectivity is there the whole paradigm will not work right. So, that is important and as the mobile network is now becoming more or less de facto for communicating anywhere everywhere specially with this IoT's and things edges and things like that right. So, it becomes extremely important how your backbone mobile network is right. So, if you look at the 1G which is somewhere in 80s.

So, it delivered a analog voice right. Then we come to 2G say early 90s that is introduced digital voice right like using CDMA technology and things like that. And then we came to a third generation that early 2000 that is the 3G brought this mobile data into play right with the 3G network we have there are incremental things right that between 2G, 3G, 3G, 4G. Then we came what we are they are predominantly is the 4G or 4G LTE what we have seen in 2010 like in the next decade broadband type of thing. So, we had a mobile broadband with better connectivity and things like that.

And all these so called 1G, 2G, 3G, 4G led to 5G which is designed to provide more connectivity than that was ever available. So, it is a it is something what is looking at as in a different level or together. So, 5G is a unified more capable air interface it has been designed with an extended capacity to enable next generation user experience right. Empower new deployment models, delivery new services etcetera right. So, with high speed superior availability and negligible latency that means, very ultra low latency as we are mentioning 5G is all set to expand the mobility mobile ecosystem to a new paradigm or new realm.

So, what we see this is this now becoming a perfect fit for our next level of cloud computing right. So, this all the things what we see with the 5G is basically is win-win

situation for this cloud computing different paradigm right. When we talk about cloud computing cloud fog age and whole paradigm itself. Now, this is just to just a continuation slide. So, what we see that in 1G the something 2 kbps and then 2G we had 384 kbps type of things then move on to 3G then 56 mbps then 1 gbps type of speed up to 4G and 5G things about 10 gig type of things right, but which is something tremendous right.

And similarly what we see that a that your latency also decreased in a in a substantially right. So, practically near supports all real time type of things. So, again just to have a comparison because always it comes into play. So, I kept this slide. So, all those references from where I took is at the end of this lecture.

So, if you want more details. So, you can go through these references. So, we have taken of those references for this our academic discussion. So, one is the latency wise, one is traffic wise rather it is 1 millisecond is the best possible things which is the 4G you can offer rather it is much more than that. And if you see peak data rate, connect the connection density architecture and things like that.

So, what we see that 5G is a different ball game altogether. So, if we look at that what is the utility of 5G right. So, 5G is that it is designed for forward compatibility this is one aspects like it is it is something which is which will be support forward compatibility will be interoperable with the with something in the future is stick services right. So, used across three main type of connected services like enhanced mobile broadband right. In addition to making our smart phones better 5G technology can be assign in new immersive experience like virtual reality, AR with faster more uniform data rates, low latency and lower cost per bits and type things like that.

So, this is one thing. Mission critical communication 5G can enable new services that can transform industry with ultra reliable high level low latency links like remote control of critical structure vehicles medical procedure like some of the examples you will see that.

Now, like if in the medical terms you need to specialist or physician they are location dependent like they are here you need to go and even you want or the this health service wants to reach you there are things are difficult. But now things will become that you can do something real time stuff like these days also are there, but this will become more ubiquitous in nature. Massive IoT so, 5G is meant to seamlessly connect a massive number of embedded sensors in virtually everything through its ability to scale down in data rates power mobility providing extreme lean to low cost connectivity solution. That means, it can connect anything to anywhere with different type of data rates scaling up at the network level itself.

That is why you are telling that some of the literature it is they were they are even referring as that network as a platform being going to delivered right network as a platform type of service. So, features in the same line actually one is that it has enhanced mobile broadband what we was just looking at the enhanced indoor and outdoor broadband enterprise

collaboration augmented and virtual reality which we are looking at. Massive machine type communication also referred to as MTC that is EMBB, IoT, asset tracking, smart agriculture, smart cities, energy monitoring, smart home and different type of application. You see that it is a it encompasses from agriculture to health to smart home to industry and things like that and ultra reliable low latency communication URLLC that the term which is there if you look at the 5G literatures. So, it is a autonomous vehicles smart grid, remote patient monitoring in the health sector, telehealth, industrial automation things.

So, these are the some of the key features which it tries to support and if you see that most of them if not all is something which helps cloud computing go to a new stage right. So, rather going to a new era type of things right with 5G into place. So, if you more specifically look at 5G and cloud computing. So, it is a perfect companion to cloud computing definitely both in terms of its distribution and diversity to compute and storage capability right. So, on premise and edge data center will continue to close the gap between the resource concern, low tendency device and distance cloud center leading to driving the need to heterogeneous and distributed computing.

So, what we have seen that on premise that is and data centers or what we say edge computing. So, it goes to the close to the resource content IoT devices which are low latency and cloud on the other hand can have that can handle heterogeneous data or multimodal data sources, distributed computing and the 5G comes in between to bridge their gap right. So, this technology will support that in that way. So, in this evolving computing paradigm service provider should look into the provide full end to end orchestration. So, that is the one of the major challenge what we see that service orchestration right.

So, it is not only that you requiring one services. So, there are now say a bundle of services which need to be orchestrated so that you realize your business goal right. So, if you look at the workflow of your particular say objective of or a particular business goal so to say then that there are different services which need to be evolved which need to be orchestrated and in a particular fashion right. So, which is which may be dictated by the orchestration engine or work flow engine and type of things. So, 5G is something it is trying to give as a network as a platform for enterprise services.

So, where was the major bottle neck of this huge volume of data latency. So, 5G comes with a boom for this type of services and on the fly orchestration and things like that. So, service orchestration will play a key role moving forward enabling industrial application to interact with network resources in advanced way such as selecting a location, quality of services, influencing type routing to deliver application on demands and type things like that right. Nowadays what we see that with real time things like starting from the streaming videos and on demand things different type of services right. So, you need to have different traffic routing or what we say classification of the traffic and they are based on their QoS's giving that type of allocation things are needed and 5G which is enormous capacity which it promises that can be a platform to realize this type of service and services or service

orchestration.

That is why the term you will see in some of the literature as network as a platform what you looked at. 5G and cloud computing there are other features like two key aspects in the relationship between the 5G technologies and the cloud computing what we see that the first this further development of cloud computing has to meet the 5G needs right. So, has to be in sync with the 5G. This is reflected by growing role of the age, mobile age, fog computing in the cloud computing paradigm as that is why in the last couple of weeks what we discussed about more of a cloud computing paradigm including your what we say age, fog, IOT taking them together right. So, second aspect is the 5G technologies are undergoing is the is are undergoing some sort of a what the term what is there is a cloudification through softwareization of the whole thing and network in NFVs that is network function virtualization, SDN software defined networks and etcetera like.

Even what we have seen that this your fast type of things right function as a service where the server is computing coming into play where this 5G deployment will be in a big way can be leverage upon to realize those things. So, both technologies types influence and development of each other right. So, 5G and cloud if you see both actually it is not like that they are conflicting I have things they are cooperative and goes hand in hand. So, 5G deployment brings up discussions about the convergence of computing, IOT, cloud, IOT that takes up the era of hyper connectivity. So, it is a what we see that we are rather somewhere on the run for a hyper era of hyper connectivity right.

Rather if you look at that the what we tried to discuss in the as in last phase of this course is more of this how they all these things work together right that IOT, edge, fog, cloud and along with aspects like mobility and things like that. So, edge in 5G is a one of the major aspect like edge computing or mobile edge computing more specifically like is the next generation cellular network that aspired to achieve substantial improvement on the QoS such as higher throughput, lower latency right. Edge computing is an emerging technology or rather it which is increasingly deployed that enables the evolution of 5G by bringing cloud capability near the end user right. So, the services or the capability near the things what we can say mini cloud to the things or user equipment sometimes called that is that equipment as user in order to overcome the intrinsic problems of the traditional cloud such as high latency etcetera. Edge computing is preferred to cater for wireless communication required for next generation applications such as augmented reality, virtual reality which are interactive in nature right.

So, these highly interactive applications are computationally intensive if you see that all these interactive applications whether it is a virtual reality for keming and other type of things and also even the things like medical or telehealth type of things. So, which have a high quality QoS requirement including low latency and high throughput right. So, also another aspect is that this type of application generates a huge volume of data which need to be moved processed at different hierarchy. So, 5G is expected to cater following needs of

today's network traffic right. One is handle massive amount of data which is generated by mobile devices and IOTs.

So, it is a huge a tremendous volume of data being generated by the mobile devices and IOT which need to be managed and cater. Stringent QoS requirement are imposed to support highly interactive applications right. So, this has to need to be handled right requiring ultra low latency and high throughput right. So, these are things which is there and heterogeneous environment right must be supported to allow interoperability of diverse range of end user requirement equipment like this things or this IOT devices are there can be different type of sensing things and like that. So, it need to support or in support or allow interoperability of this diverse range of devices QoS requirement network types etcetera.

So, this is important right. So, this is which are the things if we see which are needed in terms of making these cloud services more meaningful and that 5G gives you offers you a what we say the platform for that. So, if we look at the edge computing applications. So, there is a means big bunch of application health care, entertainment and multimedia applications, virtual reality, augmented reality, mixed reality right. So, tactile internet what we see that very very interactive things right where we look it for tactile you know or what it like with IOT and things like that. Internet of things something that what this all factories for the future which are hugely connected and things, emergency response specially in case of disaster and natural calamity and things like that and of course, ITS or Intelligent Transportation Systems which require a very not only edge level or mobile edge computing, but also it requires a huge means data load with very low latency and things like that.

And, finally, if we look at the 5G and mobile cloud computing. So, MCC is a cloud computing system include mobile devices and delivering application to the mobile devices which are which are becoming extremely popular. So, key features of MCC for 5G networks include sharing resources or mobile applications right, improve reliability as data is backed up and stored in the cloud right. So, that is there so that which reliability need to be ensured. As the data processing is uploaded by the MCC to the devices from the devices to the cloud right, fewer devices resources are consumed by applications and compute intensive processing of the mobile users request is offloaders from the mobile networks and cloud mobile devices are connected to the mobile via base station things like that right.

So, what we try to see if we look at the mobile cloud computing. So, what it has to do? It has to take care of this its backbone is the mobile network right. The whole computing paradigm is with the mobility aspect, but the backbone is the this your mobile network. So, with this 5G it supports this enhanced this data sharing capability like low latency and things like that. Rather this is a one popular picture if you see that mobile computing paradigm.

So, these are the different aspects of the things like there are different mobile networks and this your cloud data centers and things like that. And what you what we see that the overall

the success of this whole MCC depends on what is the backbone communication where 5G gives definitely a extra edge for better processing, better connectivity, low latency and huge data rate and things like that. So, what we see that if as a whole if we look at this cloud computing paradigm means when we talk about cloud computing we are talking about this cloud fog, edge, IOT devices also this mobile cloud computing and mobile edge computing.

So, if whole came it this one key feature is this your backbone network right which is dependent on the technology which is there available right. So, with this roll out of the 5G which is which promises a huge data rate, very low latency, better security we have not discussed those and high reliability with in terms of handling better this what we say QoS management and things like that.

So, what we see this Amal gametian of this cloud and the 5G will be a definitely a win-win situations right for the thing. So, or what we are looking for or what we are basically in for this cloud computing in this 5G domain ok. So, with this let us conclude our discussion. So, there is a bunch of references from where many of these materials are taken and you can I encourage you to go through these things they are that these are very very informative. And if you see that there are several organization companies which are which are now into this run for this deployment utilization of the 5G things right. So, let us conclude our discussion today. Thank you.

**Cloud Computing**  
**Prof. Soumya Kanti Ghosh**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**

**Lecture 57**

**CPS and Cloud Computing**

Hello, let us continue our discussion on different aspects of Cloud computing. So, if we today we will be taking up one concept or one paradigm which came up or coming up in a big way which is called CPS or Cyber Physical Systems. We will try to look at that how the cyber physical systems and this cloud computing work together or merge together to give a better service right. So, that those aspects we will touch upon give a overview of the things definitely it is it may not be possible that cover that all aspects of cyber physical system and what cloud computing coming up. Rather still it is a emerging area that at what point of this interaction will be there and specially with technologies changing in a big way specially this back backbone communication things like as we discussed on this 5G related issues. So, whole paradigm is going for a shift, but nevertheless this is possible to have this type of companion technology or what we say that which will be which are more real to that our actual working of the things right.

So, initially we will have a little overview of the cyber physical systems and then we will try to see that how this cloud and cyber CPS are act as a companion to companion technology.

So, we will discuss mostly on cyber physical system or CPS and cyber physical cloud CPS and cloud computing or more what we say sometimes we will say cyber physical cloud computing systems or cloud CPS and things like that right. So, there are different nomenclature which are being coined across the things right. So, mostly the keyword is this.

So, if you look at what first of all let us see what do you mean by the cyber physical systems rather if we look at intuitively rather we will see that many of the things we are already knowingly or unknowingly are that have different uses or sometimes involved in some of the things where the CPS is in place right. So, cyber physical system or sometimes called as cyber physical systems right. So, CPS is an orchestration of computing and physical system right. So, one is that cyber space and the physical space how this computing or how this loop goes on to fulfill a particular objective or business goal. So, that is one of the things.

So, embedded computer monitors control physical processes right usually with a feedback loop where physical processes affect computation and vice versa right. So, like one maybe one good example of this embedded systems and how this control and even there is a human in the loop right like one maybe one good example means modern day cars right. So, there are like sometimes it is people even say that these days the number of so called

embedded systems a modern day car is having maybe more than the mechanical system anyway that may not be true, but what it says that the good amount of embedded computing resources are being used to appropriately monitor or appropriately handle or manage this overall this automotive or a car or vehicle type of things right. Similarly with EV or electronic vehicles coming in a big way. So, this becomes more that using of the systems are there right.

So, it has a feedback loop like once like if you are driving accelerating then it should go on a loop that what should be the things if there is a sudden break or collision the air bag opens up right and there may be things that once you push a break automatically the this brake lights are on right and not only that these days those who have looked into aspects like vehicular communication or vehicular or vanit vehicular ad hoc network. So, there are different safety mechanisms which in come into play. So, once you once there is a sudden break or when once a car breaks. So, it sends a some sort of a call SVA type of alert slow vehicle alert type of things which other cars which are in that vicinity which are in this vanit purview they know that what there is a slow vehicle is slowing up. So, it may it may sometimes happen especially in high road that the vehicle need to react much earlier than the driver reacts right.

So, usually human reaction times as a and some require some lead time. So, but the system may want. So, what is going on everywhere is basically overall this monitoring and sensing actuating of the things are being in a somewhat automated or semi automated way where the human in the loop right. So, there is a physical system and what we say is a computational or a cyberspace which are working together right. Even I can put it in the cloud where it monitors the overall paradigm the cloud and request for a detouring of the vehicle if there is a congestion level is high.

As of now we may be many of time here what we are looking at that some of the mapping and see that if the congestion is high then we go for a detour and things like that, but that can be made in a some sort of a I can so called quote unquote I can have a CPS of doing this right. So, the terms psychophysical system emerge in somewhere around 2006-2006 when it was coined by Helen Gill at NSF in USA right National Science Foundation is in a thing. CBS, CPS is about interaction not exactly union of physical and cyberspace right. It combines engineering models methods from all branch practically all branch of engineering starting from mechanical, environmental, civil, electrical, biomedical, chemical, analytical, industrial and anything which are comes under things and method of computer science to act on it right. So, we there are different methods of computer science which act on it right like I can have different landing techniques, control systems and things like that and we can work on it.

So, application of CPS includes automotive systems as we see one of the one may be one of the best example of cyber physical systems may be anyway automatic manufacturing there are different other military system, assisted living and medical systems, traffic control,

safety and so and so forth right. Rather anything around us we can think of is where somewhere or other the some cyber physical systems is there. So, as I was telling one of the good example may be our whole this air traffic systems right air traffic in the sense this that airplanes taking off, moving and things like that and all those things there are thousands of codes which are being running there are human in the loop that ATAs and things which are controlling there are pilots which are taking call. But it works seamlessly right like some people say that if there is no human say intervention or human imparted error or there are no maintenance problem or extreme climate condition etcetera otherwise bringing a otherwise means a aircraft going to a problem may be something very very very difficult scenario very very low probability right it is hardly anything will go wrong. It is irrespective of so many aircrafts on the sky and things like that and what.

So, this is a unique example of this cyber physical system or even a traffic control or medic different type of medical or telehealth type of things also medical devices where you are sensing doing some things taking actuating and things like that. So, if we have a big picture taken from one of the references. So, we have on computational platform there may be which this there is a physical plant which gives input. So, it may be there are bunch of sensors etcetera which gives in to this computational platform and which calculate something along with the network things and if there may be a feedback loop which is going on and there is a there can be other computational platform which are looking into the things where it goes on a loop and execute the stabilize the physical system or basically manage the physical systems right physical plant in this case. Now, see this even the computational if there are more than one computational platform like that or computational unit.

So, it may happen that modeling one unit may be, but when you have multiple unit and do a overall modeling of the whole thing that is a big challenge making the CPS overall model is a big challenge right. Say even if I see that if the car type of example one component maybe like you put a brake the lights etcetera, but if it is related to other type of things like when to switch on this air this airbags or how to alert the adjoining vehicles and if there are other dynamics I am not a expert anywhere in the automotive, but there are if number of systems need to collaborate and to take a call then say modeling itself is a serious challenge right. There may be different cases where if this happens and etcetera and modeling all those cases is a serious big challenge. So, what we see CPS attempts to describe a broad range of complex multidisciplinary physically aware next generation engineering system that integrates embedded computing technology that is the cyber part into the physical part right. So, it is a really a complex multidisciplinary physically aware next generation engineering systems right.

So, in cyber physical system physical and software components are deeply intertwined right. So, it is not something detached it is something more means closely lived or what we say it is very strongly coupled right. Able to operate on different spatial and temporal scales right. So, both in spatial and temporal scale exhibit multiple and distinct behavioral

modalities and interact with each other in a way that changes with context right. So, things may change with context right.

So, even some activity in some context and some activity in some other context may be different right. Say maintaining a particular speed on a say particular kind of road, particular kind of weather may be different in the different type of context right. So, different weather condition and things like that right. So, it need to take care. So, CPS involve transdisciplinary approach.

So, across the discipline it merges theories of cybernetics, mechatronics, design and process science right. And over if we say that over and above there is a computation of the whole computation methodologies which into acting to things. So, what we see cyber plus physical plus computation plus dynamics plus communication plus security plus safety plus maybe the human in the loop right. So, that interaction of the thing. So, what we see that a bunch of field come into play to make this whole thing happen.

Now in order to design such system specially model such system is a very very serious problem right challenging problem. Now CPS definitely is a emerging discipline that involves engineer computing and communicating system interfacing. Ongoing advances in science and technology improve the time between computational physical means specially with increase of things like we have this cloud or multi tire like cloud fog age type of things. And to be things that if you look at the IOTs or the sensors which are at the sensing things of where which is there in the physical system in the CPS right. So, sometimes people try to say that this IOT driven ecosystem is the CPS right in some of the literature we find they try to they put this logic which is like it senses and actuating and the whole ecosystems had say age fog cloud into the play right.

So, potential application of cyber physical systems are several means nothing. So, includes intervention like collision avoidance as we are telling that if there is a if there is a sudden break in the traffic then how whether that it can send automated alert and takes a call on the things that how this collision avoidance process will be there. Precision robotic surgery nano level manufacturing so, these are the precision. Operation in say dangerous or inaccessible environment like search and rescue operation, fire fighting, deep sea exploration, coordination, air traffic control, war fighting, wartime situation these are the more of a it can acts as a coordination. Efficiency like zero net energy building if you want to increase the efficiency and actually lying away this you see that like if you walk through a corridor the lights will switch on during the evening times and once you are out going out the light gets switched off.

That means, if there is no there so, minimal lighting is there. So, major lights are off right. The day and night lighting those type of things are there like automatically once the sun that daylight is sufficient then the automatically switching off the lights things are there. So, there are there can be different control with respect to AC with respect to overall energy

management of the a building that is the net zero energy building type of things. Augmented human capability, healthcare monitoring, delivery and type of things like was sometime what we see that health band and we will discuss that that it can go to the fog etcetera that is basically emulating a cyber physical system thing right.

So, typical we have already discussed smart gate, autonomous, automobile systems uses of cyber physical system. So, it is interlinked of sensor actuator processing device create a vast network connected computing resources. So, integration of computational with the physical processes and use sensors actuators can be viewed as a sometimes called a computing at a physical as a physical act right. So, with the physical systems so, computing coming as a physical act where real world is monitored through sensors and transfer sensing data to the cyberspace where the cyber application and services use the data to be to affect the physical environment. So, I have physical systems they are sensing, sending to the cyber world and then the it is being with some methodology algorithms it takes a call and influence this physical system.

So, cloud computing services provide a flexible platform for realizing goals of CPS right. So, what we see this cloud computing paradigm may be a may be a good what we say may be a potential flexible platform for realizing this overall goals of this cyber physical systems right. So, there is a that what we can realize that as there are as we have seen this IoT sensing sending and all those things taking a call and the actuating is done. So, this can be realized through this cloud computing platform. So, if we look at the cyber physical system so, you underlining you have the physical system there are server cyberspace where there are data, processes, apps, web applications, services and different type of things are there and there are underlining this ubiquitous network right which connect it.

So, what we say that the sensing is data by the IOTs, IoT or different sensors which are transmitted there in a sense it that can be multilayer by edge, fog and things like that. So, that and it basically go and a feedback loop which is the. So, this is not only the cloud it also contains this cloud overall paradigm like underlining edge layer, fog layer and the cloud layer right based on the level of processing required and the type of services required things are there. There can be also possible there can be also possible rather these are being used these days like different services with the things can be there. Like if I know that for this automotive movement these are the services required at different places like at something at the edge level where the vehicle itself can be a has a the vehicle OBU or on board unit can acts as a edge layer or edge device and the fog can be the road side units and the cloud can be the back end cloud.

Then I can basically emulate different services which can be energized like one can be like something what safety related services like if there is a safety hazard it sends messages to the other things. There can be services related to say health type of things like if it is again there is something that if there is a requirement of medical things it coordinates with other hospital etcetera give intimations things. So, there can be different type of location driven

services which can be energized right. Starting from infotainment you crossing say Kharagpur it will automatically energize or load the Kharagpur related information starting from that what the place about what are the possible gas station or petrol pumps and also say food courts or medical centers etcetera right. So, those can be energized based on your source there are space, time, mobility what all everything comes into play right.

So, these are the things. So, if we assume that CPS and cloud. So, a so, also known as also sometimes referred as cyber physical cloud computing architectural framework can be defined as a system environment that can rapidly build modified provision cyber physical system composed of a set of cloud computing based sensors processes control and data services right. So, what we see these are scalable right can be provisioned as per need right.

I may not be one to always provision this when I am not needing this right. So, I want to provision when that there is a requirement for the things like specially when we have some of those application in that line.

So, that is nicely defined as a system environment that can rapidly build or provisioned modify a or provision a cyber physical system CPS composed of a set of cloud computing based sensors processes control and data services. So, the CPS itself now have that cloud computing as it is one of the component which will help in provisioning different type of services which is needs and also have the other cloud computing benefits into the play right. So, that like scalability issues of or in terms of efficiency or in terms of ubiquitous presence right. So, what we see that whether you have cyber physical computing and cloud computing and merge to a cyber physical cloud computing type of framework which is very much on the card right people are using it right. So, scalability and interoperability which this supports and on the other hand the interaction to the physical component.

So, what we see that is a some sort of a win-win situation we when we look at the cyber physical cloud computing. So, benefits it is already obvious like efficient use of resources, modular composition, rapid development and scalability, smart adaptation to environment at every scale right based on your that where which environment reliable and resilient architecture what we can provide right. And also additionally this sort of data can be used in the future for learning and developing better models of the things right. So, it is not only that supporting the systems which is at hand, but also in future this data can be analyzed to have improve the model overall CPS model right. So, if we look at the whole CPS and cloud computing in some other way.

So, what we see there are different type of domains right starting from school to hospitals, there are human user, traffic control, ambulance, medical devices, web service etcetera. And this cyber physical cloud have different type of services starting from sensor services, actuator services, processor services, data services and there are things right. There are different type of semantic engine which can work on those at the backend right within the cloud. And so, this several components can interact with this cyber physical cloud in a ubiquitous form right. So, what we see that this overall scalability in terms of not only that

number of user, number of devices coming into play, number of component, but also scalability in the terms of services right.

Scalability in type of storage and analysis at a high scale can be done, can be what we say can be realized using this cyber physical cloud system. So, what we see this overall this cloud computing paradigm becomes a candidate as a potential candidate to realize the cyber physical cloud computing system. So, if we look at the a high level the cyber physical cloud computing paradigm. So, what we have this at the underlining there are different sensors and there are different sensors, system user, there are other what we say mobile components or mobile in this case was this say ambulance or something or we say mobile objects right which contribute to this different type of data sets right. Which are which may have a direct interaction of the things like that if this sensor camera detects ambulance right which basically needs a priority lane or what we say priority movement permission or priority prioritization of its priority traffic what we say.

So, what we see that that we see this that this type sort of data can be there in this sort of a underlining the sensing actuating things right. So, some are some there may be direct interaction, some may be indirect interaction by the actuators right. So, so even actuating that based on that may be there if it detects a more traffic that it can send a signal to this type of priority vehicles like it may be ambulance, it may fabricate that you need a detour and things like that or in other sense you get a information things like that. So, there are there are different type of information being collected. On the other hand I can have a cyber physical system models right.

So, functional and non-functional requirement which can have a cyber physical system models where which are which basically tries to model the different scenarios based of the domain specific things right. So, there are system what we say modeler or system builder which are the model. In between what we see that a cyber physical cloud platform which basically ties this to the right. So, this model based on that different type of services are available or models and services are available in the service catalog which can be what we say provisioned by this cloud service API and there are run time services can be used right. So, what we see that this one side is that building of the cyber physical models, one side is that this overall interaction of what we say that physical scenario on the ground rather physical scenario on the other side and this cyber physical cloud computing platform bridges this things right provision the service takes the input from this physical scenario do some actuating based on that analysis or back end computation or that particular domain.

So, this is one thing again taken from of the references what we what that tries to see that it has the physical resources where CNC controller CNC machine tools etcetera there and there are local terminals which are sensing sensor using sensor system user interface and so and so forth that are the things which are more at the ground level or the factory level.

So, these are the physical resources on the other hand there are cloud resources like advanced signal processing and cognitive decision making type of algorithms are can be

run. So, which can feature extract and selection professional signal detection and there are different applications which gives a feedback to this physical systems right. In between there are local servers which takes different data do some pre processing and based on this diagnosis take some decision on the actions and actuate this CNC controller. What we see so, there is a physical system one side and there is a big resource of the cloud on the other side in between there can be local servers which accumulate data do some pre processing before pushing into the cloud and then some of the some of these diagnostic tools can directly send to this that missing monitoring status unit or it can shown to the reason on the accent to the local server which in turn control this controller.

So, what we see that intelligent monitoring of machining or machining processes I can have a realization of this cyber physical system. So, other another interesting work what you can what you see that what we see that this cloud edge computing framework for cyber physical system or they refer to as cyber physical social services right. So, there are there can be different type of instances like which are the cyber world, social world and physical world right. So, those are local and having the edge things there can be n number of such things and goes to the cloud for interaction which can take a global call. So, the local things are resolved at the local level, but something at the global level things can be done at the at a higher level.

So, in the next technique if I say that this car movement congestion etcetera that can be so long it is within the local domain it can the edge computing devices that are resolving it or that on board car things are resolving the things, but when it goes to a larger region then the now inputs from the number of cars etcetera are being processed at the cloud and helps in taking a call right. So, here also we see it is a multi layer cloud fog edge or cloud edge and then the sensory or the sensor actuator platform right. So, what we see that local CPS's or cloud computing a cyber physical social systems where human in the loop and other physical systems are there, there can be number of local cyber physical systems and then we have a cloud plane which takes inputs and from this and take a call and send the say processed information or decision to the this local CPS's right. And we can have different type of application to the things it can be application plane have a something a model for the smart city, it can be a smart home, it can be a smart factory or say smart health and can be say smart vehicular movement and anything right.

So, what we see smart grid and sort of things right. So, what we see that underlining there are physical systems which has local controlling and actuating the things right and then if you have a global view then it push it to the cloud plane or as we have seen that edge and fog and cloud plane right to take a more informed decision. So, what we see what we try to see in this whole thing that this cyber physical systems which are becoming a defacto platform for various of our systems as most of the physical systems are being have lot of sensing and actuating coming into play. So, this along with that cloud computing will provide a win-win situation to realize a better platform right. So, there are few references very interesting one there are several other references, but these are some of the things

which I referred in this thing. So, I encourage you to look at some of this interesting works see the things and this is a nice field for not only realization or what we say engineering or technology point of view also it is a nice field that is not only application, but also in terms of research also right.

So, with this let us conclude our discussion today on this cyber physical system and cloud computing paradigm. Thank you.

**Cloud Computing**  
**Prof. Soumya Kanti Ghosh**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**

**Lecture 58**

**Case Study I (Spatial Cloud Computing)**

Hello, so we will continue our discussion on this cloud computing course and in today we will take up one such case study where this cloud is where this application of cloud will be of great help or we will see that how this application of cloud basically support the this sort of applications right. This is basically on spatial cloud computing right. So, what we will do we will be discussing we will give a brief overview of spatial data if you recollect some time back means for few lectures we talk about spatial cloud, but now we will have some other perspective basically a application of some spatial analysis over cloud that we will try to see today right. And there are the spatial cloud and spatial cloud computing are the keywords right. So, when we talk about spatial analysis, so what we mean to say that what are the different type of analysis possible on spatial data set. So, what is a spatial data? So, spatial data is anything which is have something a feature along with a coordinate to the thing right or feature along with the location.

So, it is typically all our geospatial data or anything on the earth surface are spatial data right because there are objects feature or characteristics of the objects and there is a location in forming like if I say that the building where I am talking from or now this be this typical lecture is being recorded it is inside IIT Kharagpur. So, it is a studio of in Tokusla building of IIT Kharagpur. So, it has a coordinate and it has a feature like it is a say 3 store at building and there are this type of labs are there. So, there can be it has a name, it is a academic complex, it is a academic building not a residential or other type of things and there can be other features.

So, what we say there is a spatial object like the building, its location, its spread, its typically if it is a building it is a plan those are the spatial object and which are can be viewed from the sky right. I can fly a drone or put a camera and can picture and there are some non spatial or attribute data which we call non spatial things like name of the building. You cannot see the name of the building by flying a drone unless it is written somewhere at the top of the things right or which are the departments are there, how many people are there, whether it is a academic or residential building or any other type of thing.

So, that you cannot see right over the things. Similarly, if I if there is a road right.

So, we know that this is a road. So, some of the things which can see that this that is a what we say that low we can think of a series of lines or poly lines right. So, small line is a series of line or a series of coordinate systems right which are represent the road, but name of the road, type of the road and there may be width of the road and there may be other features

which are attribute the things. So, when you talk about spatial data more specifically we are talking about geospatial data, let me something related to the geo or the earth surface right.

It can be on the earth surface or near the earth surface right.

So, those type of data and as we see that in today's scenario any type of work or a rather any type of work involving say development planning or anything what government or federal structure wants to implement there is a spatial context into it right. So, there is a huge application of this data. So, it is a location along with a feature set rather it has an another dimension called time right. Over time how things are going on right suppose I want to see a forest patch or forest area right. So, over time how this forest coverage is changing or if even if we can say that a say a path of a river.

So, over time whether is a change of that particular path or not right. So, even say agricultural land over time how things are there even we can look at the other things like meteorological data right. So, there are there are lot of there is a temporal aspect. So, what we say when we talk about not only spatial rather spatio-temporal data analysis right, but spatial data can be other aspects also like say it can be a medical image right. So, like a say a scanned of a human body.

So, it is also have with some reference right with it if I put on a reference sheet like say graph paper. So, we can say that this organ is right of these or above these and like this right. So, but for our today's discussion today's context we will consider those data which are related to the earth surface or what we say geospatial in nature. So, what we see data is a it is a information that describes objects events or other features with a location on or near the earth surface right. It can be on the earth surface something or near earth surface may be something like if we say that temperature of the things or moisture content in the area, pollution of a particular region there is everywhere there is a location is there.

So, geospatial data typically combines location information usually coordinates on the earth right. The most popular coordinate we are used to is that lat long and attribute information characteristics of the objects events or phenomena concerned right. So, these are the things with temporal information and time or lifespan at the location. So, what we see that location some feature set and temporal aspects like over time how it is behaving right. So, what we see there is a enormous variety of applications which comes into play right that we will see some of the things we will discuss and things.

And similarly a vast type of analysis tools or analysis which is possible over the things right. So, like we are used to some mapping services right going from one location to another you want to find the what is the shortest route, what is the congestion level etcetera. Even if you see that in case of some of the cases like national natural disasters like cyclone movement etcetera finding the path, detouring vehicles. So, it is a whole dynamics is there. Another one study shows that more than 90 percent of the data set across the world has some special context right some location information.

Even if we look at the student database. So, it contains like which college or which department see or he is studying, where see or he is staying and type of things everything has a coordinate information inherently in there right. So, it makes sense to work with this type of data set. Anyway our context is not discussing on the spatial data per say, but try to see that where cloud come into play and how it help in having better analysis of this type of data right. So, whenever we look into a map we inherently start turning the map into informations right.

So, whether it is any map where you are seeing. So, what we are trying it is a not only it is a we are talking about two dimensional, three dimensional we put another dimension into it. So, what we try to do we try to look into the map and try to extract information out of the things like contextual information etcetera. So, we try to find patterns, access, whether there is a trend and try to make decisions right. Like there are things like crime study, like crime hotspot in the in a particular region right like major cities they work on crime hotspots mapping and try to because the amount of police force or amount of this patrolling of the city police is it is limited right.

You cannot go on a enormous things, but they want to find out that which are the targeted things and whether it varies over time and over time and events driven right. When there is a some event like some festivals or something that crime is in some dimension or normal time something, day time something, evening something, night something. So, based on that there is a thing called that I want to find out that what are the crime hotspot or crime things and based on that want to do some routing of this police vehicles etcetera type of things right. Similarly, want to look at some of the things like flood or drought analysis right that what sort of things whether I can have a predictive model that what is going to aspects taking a multimodal data analysis right. Finding optimal path which are we are used to that for a region, predicting lot of things like starting from say like like calamities like drought or flood, looking predicting something other events like whether this event whether there is a congestion or traffic jam is expected right.

So, where the spatial data come into play in some there are other data inputs definitely. Now, if you see all these things this is a huge volume of data right, enormous volume of data those who have worked or if you can search in the internet you will see the volume of data is so large doing any meaningful any meaningful work within a tangible time sometimes become a serious challenge right. So, whether we can leverage on this cloud or some sort of a cloud fog age type of things to work on this type of things. And of course, when we talk about analysis of this data like attempt to solve location oriented problem for better understanding of where and what is occurring around the particular region right, location oriented is challenges right. I want to see that if this is this happens so, what are the different effects is there like if there is a congestion in a region say a particular region location in a city.

So, what will be the effect in the surrounding area or whether some other region will be affected for that and type of things right. So, there is a there is lot of if there are heavy rainfall in some area or whether there is a chance of getting flooded in other area due to say more overflow water over the rivers and etcetera right. Some of the things intuitively we can know some of the things we need to study lot of other things like lot of data analytics come into play even there may be some science or physics theory we also will be there right. Nevertheless it is a huge volume of data need to be churned for a meaningful reason within a within a say tangible time or near real time if it is required then. So, it is something beyond mapping only just looking at the map is not the thing it is beyond mapping we want to find out that characteristics of the region and places and relationship between this region and places right.

So, these are the things right. So, spatial analysis tends new perspective to any to any decision making it lets you to pose different question and derive answer on the data right you can say that if this happen what will happen and type of things over the things and helps to derive new information and make informed decision right. So, in the spatial contact we can have lot of decisions specially in development back planning activity of the government where want to do some development planning in terms of different areas you have seen the different government plans and things like that. So, where it comes into a big way and organization that use spatial data analysis in their work are wide ranging starting from state means local government, state government, national agencies or national government, central government, businesses practically of all kind, utility companies like telephone or we see water supply, electricity everywhere some spatial context is there and of course, things like that e-commerce sites etcetera. They also need to look around that if I want to this how what will be my delivery time and things like that right and of course, researchers etcetera in colleges, universities, NGOs and it is a endless type of things.

So, what we want to do we need to identify the data sets right which we require want to put on the same page right same page in the sense in the same base map right I want to see that two data from two sources if I want to analyze then need to sit one over another right.

So, they should be on the same platform in this case the location need to be mapped appropriately. So, that is one thing. So, identify the things analyze the data based on those data sets all may not be spatial there can be non spatial attribute data sets also and then want to draw inference out of it. So, these are the typical three blocks we can see there can be we can expand this into different dimension, but one is that identifying the data in the which needed then analyze and things and finally, drawing inference out of it right.

Now so, what we see that this spatial data analysis are are have lot of applications and there is a need to look into this analysis in a bigger way and as the data is pretty large any movement of the data from one to one place to another is a serious constraint right. So, if terabytes of data need to be moved from one to another it is a challenge it is one of the major challenges right. So, if you see that what are the challenges this first of all it is a very voluminous and it is not only spatial, spatio temporal in nature integration of

heterogeneous data sources. So, different sources like suppose you want to do something that estimation of or crop production and things. So, it is not only the what is there on the crop right the whether there are expected rainfall right or whether there is a drought condition, whether what is the different temperature variations right and there can be lot of other factors will come into play even some of the economic factors whether there is a demand or things like that right.

So, it is not only the demand for this crop the demand for fertilizer etcetera water supply all those things into come into play. So, it is a heterogeneous data source multimodal data source which will come into play you need to analyze those data. So, we need to have something which will better support discovery access and utilization of the data and data processing as to relieve scientists and etcetera. So, whether we have some of the way that we can do a quicker way of which that like different sources of data where I will get at where is the latest data format etcetera. So, provide something what real time or near real time resources to enable some of the things which demand some sort of a very quick or real time applications like specially for emergency response if there is a fire break out or cyclonic storm etcetera then we may have to do in a very quick fashion.

So, huge data need to be churned and worked into there may be we need to deal with access spikes right suddenly the data access things will be there analytical things will be more for things right otherwise the servers are not so busy, but suddenly there is a lot of demand on the things. And provide more reliable and scalable service for massive numbers number of concurrent users to advance public knowledge etcetera. So, there can be lot of demand there can be lot of data put into it for the things. So, I need to a scalable service like so, when the demand is scale down when the demand is more it will scale up and things like that. So, what we see looking at all these aspects this cloud computing paradigm right may be a good fit right.

So, cloud computing what we see it is a model of enabling ubiquitous as you if you remember your early slides convenient on demand network access to a shared pool of configurable computing resources that is network, server, storage, application, services that can be rapidly provisioned and released with minimal management effort or service provider interaction. So, this is the things which we have already know. So, this if you see is fitting into the this context. So, what we see this spatial analytics right or spatio temporal analytics plus cloud computing is a good match right. So, emergence of cloud computing provides a potential solution for elastic on demand computing platform to integrate several things like observation system, parameter extracting and all those different aspects right.

And provide some of the things which have a social impact and of course, giving users a better sense of thing right. So, search access and utilize geospatial data in a better way there may be a chance. Configuring compute infrastructure to enable computability of intensive simulation model and things like that where you require a huge simulation for doing some of the climatological studies and other type of things. And adopt spatio temporal principle

to support spatio temporal intensive applications in the cloud right. So, these are the things which are possible with our knowledge of cloud computing what we have seen in this context right.

So, spatio temporal cloud computing if we try to now put both together. So, refers to the cloud computing paradigm that is driven by geospatial sciences and optimized by optimized by spatio temporal principles by enabling geospatial science discoveries and cloud computing within a distributed computing environment right. So, lot of things has been told in this slide. So, first of all whatever the cloud features are there it is there, but along with that that is the thing what we require when we put into the domain right. We are now trying to see that how the cloud computing will fit or how things will help in the spatio temporal domain.

Had it been in a banking domain that will be another way of looking at it, had it been in say some other domain of say study of climatological specifically or something that will be some. This is if we try geospatial data sources. So, first of all geospatial sciences should come into play right. They are way of handling the say knowledge build up and type of things right. Like I will tell that one basic theory what they assume in this say what they say that Tobler's first law of geography.

It says that the near all objects all this objects on earth are somewhere related to other objects right. The more near the objects they are more related than the distance object right. Like what we say like temperature of Kharagpur more likely outside definitely more likely will be near to this temperature of say Midnapur which is around 10-15 kilometers away. It is less likely something like temperature of say something like Darjeeling or Simla right. So, one context is the far is the object, less is the your influence right that is one that and there are other context right.

Like if we say that what elevation or at height we are there etcetera. So, what we try to see that there is some principles or science of that particular domain right. So, the methodology or knowledge of the domain that need to be embedded into the things for better analysis. So, that is why it says that is driven by geospatial science and optimized by spatio temporal principles for enabling geospatial science discoveries and cloud computing within a distributed computing environment.

So, this is there. So, if that if the domain changes the things will have they are science etcetera coming into play. So, this is what we see that in this spatio temporal spatial cloud computing context what it goes on right. So, spatial cloud what we then if we try to set it supports shared resource pooling which is useful for participating organization for both common and shared goal right. This is true for any type of cloud, but in the spatial context choice of various deployment service business model to best of suit the organization models and manage service prevent data loss from frequent outage minimizing financial risk etcetera. So, all those things are really also suits for spatial cloud context right.

And of course, these are some of the advantages already we are some of the things already we are knowing easy to use scalability cost reliability risk and things like that. And there may be other things which are typical features of this cloud things which comes when you look at the spatial cloud computing context also ok. And typical architecture you may recollect that in our means we discussed the spatial context in some other means before also then also this came into play like if we sorry if we look at that thing. So, what we see that is a this is the cloud where we have software as a service or storage as a service or it can be a platform as a service and infrastructure as a service these are the different service.

And if we try to look at the different type of spatial things like it is a data provisioning where we have different type of spatial web services web feature service catalog service and things like that those are the things processing services map services and there are catalog specific catalog and work means type of services these are the different services which are there which can be provisioned out in here in spatial cloud right.

So, these are the services which basically give provided to the users it can be organizational system or individual users right. So, private and public organization wants to share their spatial data that is that is there they are willing to share different requirements for geospatial data space and network bandwidth will come into play easy access of spatial services and GIS decision making the geographic information system decision making made easier like integrate latest database merge the separate system exchange information internal and external things like that. And we have a typical architecture what we see we do not have to break our hand with all these terminologies, but what we see there is a registry where we can find there is a OGC client service OGC stands for Open Geospatial Consortium which makes standards for the spatial data. So, those are things so, any OGC client instance can access this services there are different feature services like feature service, map service, processing service which helps and it can be done that some of the things which accesses different databases it can be post grace oracle and different type of spatial databases and things. So, it over around that more breakdown architecture if we try to see that there are different features.

So, what we try to impress upon here is that it is possible to map this with appropriate cloud instances or cloud services right or we can basically instantiate those services to support spatial things right those are these one point which is there. Now we will try to look at one typical example of mobility analytics right like we will see few slides on that how it is how the spatial cloud or cloud for IoT age IoT infrastructure helps this. So, this will discuss so, any this is a work of my ex scholar Dr. Seyago she worked extensively on this area how this mobility analytics there. What do you mean by mobility analytics like every when we move specially any human movement it may be working or moving through other transportation system with moves with a intent right right it moves with a intent right for that matter if you look at whole animal world there is a when there is a movement usually there is a intent in it right.

Now if we will see our cities these days right. So, there are lot of there is a lot on the travel time right. So, that is the first thing we see like if you are in a place like Kharagpur you do not realize that what is the importance of travel time, but once you go to Kolkata, Adil, Bombay, Bangalore or any city for that matter you find that there is a travel time is a major thing to look into right. So, it makes sense if we look at the urban context for example, so it makes sense to see that what is the overall mobility patterns. If we know this mobility pattern whether it is possible to have a better prediction of the things like how this overall mobility dynamics can be improved. So, that the congestion can be reduced travel time can be improved right and or what are the things I am going to expect right.

So, whether it is even for individual whether it is good for start at 4 pm or 3 pm or 5 pm to have the minimum traffic if I have flexibility of starting time right. So, nevertheless looking at this mobility dynamics one is that is important from the urban context. Even if you look at mobility of animal natural or natural phenomena like hurricane tornado migratory birds etcetera that is to study all those things I again require a mobility dynamics like how they are moving. So, human movement like daily GPS footprint, travel logs, check ins, mobile phone, activity logs this also helps for a particular human what is the movement patterns and accordingly we may facilitate different type of things into the things right. If lot of people are going then we can have say even a more number of buses in that line or more number of resources in terms of say vehicles or more resources on the ground like maybe waiting places or something drinking water facility and washroom facility and things like that right based on your overall movement patterns human movement patterns.

Or it may be allow me to planned overall traffic or road signaling mechanisms right. So, movement of vehicle and transportation data those are the things and smart home sensors these days sensor data IoT home monitoring systems etcetera try to see that whether starting from uses better utilization of the utility services right and things like that. So, what we see this mobility analytics may help in different aspects starting from transportation sector to say looking at animal movement to say predicting or forecasting of some phenomena like cyclone etcetera where the next location will be there individual people individual citizen or smart home things where within the home what are the movement pattern etcetera specially for old age and etcetera we can look into the things right. So, in the in our PhD works they are developed one a generic framework for trajectory trace mining for smart city applications or it can be looked into the things we are not going to all details of the things. So, it is intelligent framework to model city dynamics effective storage and computational method learning techniques to extract mobility patterns and facility in recommendation system things like that for all these what we require may be a cloud based systems will be a good thing to look at because a huge amount of data have lot of applications which come into play.

Another there is a concept called SDI which is based on web services and this is called a spatial data infrastructures right how the spatial data infrastructure can work. So, what we look at some of the things like a trajectory cloud for enabling efficient mobility services

what we named it as a trash cloud. So, what is a which helps in trip planning services travel time prediction and ride sharing services. So, these are different type of services which can be enabled.

So, we are not going into nitty-gritty of the thing. So, if you see that that spatial trajectory the different component. So, a spatial trajectory a stress generated by the moving object in the geospatial phase using different type of things. So, it is series of x y semantic trajectory like human moved with a intent. So, whether there is a semantics into the things right like if we say that student during the evening when they are coming to the hostels they will likely to stop in that food joints and coffee shops etcetera.

So, there is a intent of doing this. So, there is a semantics into the things right which is not there when they are going from hostel to the classroom anyway that is the thing when it is not there. So, that can be mapped into different type of things right. So, what we worked on is trash cloud named as a trash cloud for analyzing the urban dynamics right. So, mobility trace analysis has a significant role in mapping urban dynamics. So, these analysis help in location based service provisioning right and facilitate an effective transportation mechanism.

So, key aspect of the intelligent transportation network. So, ITS is one of this looking at the dynamics. Using mobility traces and providing location aware services is a very very challenging task. So, an end to end cloud based framework may facilitate efficient location based service provisioning right what we have tried to seen and shown experimentally also. So, it helps to minimize service waiting time and service provisioning time to location based services such as food delivery, medical emergency etcetera right where the time is a major critical things right. So, a trash cloud which analyze this trajectory analysis over the cloud and give this type of services on a much better way.

So, if we look at the big picture of the trash cloud what say I worked on for the one part of our PhD is that one is that trash cloud which has one side that urban sensing through data acquisition like that can be GPS, road network, traffic information that is may be different sources of data which are coming to this spatial trash cloud. There are trajectory data pre-processing which need to be analyzed or pre-process some sort of analytic tool which worked on the things. There are trajectory data modeling right using map matching and trajectory stress geo tagging and etcetera. So, trajectory data modeling the trajectory data and trajectory trace processing right trajectory information retrieval, spatio-temporal range queries. So, answer different query, trajectory based, query processing and different type of things.

So, these are the different component and if we look at the different type of services one is the trajectory trace indexing type of services, trajectory map matching services and trajectory query processing services these are the different services which this particular trash cloud provides right. So, we are not going again to the details. So, which has a if you

see that for this type of services that what has been used is this Google cloud platform for this experimentation which interfaced with this trash cloud in the GCP and different type of like trash data indexing services it uses Google BigQuery and cloud SQL storage Google cloud SQL storage, trajectory map matching, Google compute engine, trajectory query services, Google compute engine and cloud SQL. So, these things have been exploited. So, what we see with this example a scenario a use case or case study which has been experimented with which has been implemented over Google cloud rather we have tried with our own we have a that Megamala cloud which is there in our in house cloud we also tried with that, but what we show in the Google cloud you can have a feel that you can also have some sort of a use case of your thing and work on it.

So, what we try to see in this particular today's discussion that how I can have a motivating example or motivating case study of a spatial cloud and how I can implement over the things right ok. So, in the we will continue our discussion we will see what are the other type of application etcetera. So, this is the reference things you if you are interested you can go through this work. Thank you.

**Cloud Computing**  
**Prof. Soumya Kanti Ghosh**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**

**Lecture 59**

**Lecture 59 Case Study II (Internet of Health Things) (Part-A)**

Hello. So, we will be discussing today on one case study which will try to show you that how this cloud to things paradigm right, cloud, fog, age and internet of things how this whole paradigm works in a integrated fashion right. So, it is primarily of internet of health things that is a application related to the health, health application though the application per say may not be totally medically proved right. That means, whether or I should say that may in order to deploy you may require lot of other things, but nevertheless it shows that how this sort of application will be there. So, primarily we will be looking at cloud, fog, age, IOT framework as a whole or what we what these days we say that cloud to things continuum that means, it is not separated, but rather it is a integrated things right. So, you cannot say that the only I manage this I the IOT or age or fog or cloud.

Now, in present day situation you need to look into the whole thing in a more holistic fashion more integrated fashion. And as we will be looking at a case study on internet of health things. So, that is that is the one work which we did in our lab. So, that that will show you that how things can be done.

So, this is the thing. So, what as we are discussing. So, it is primarily a case study on cloud, fog, age computing for internet of health things right ok. So, we will we what we have seen in this course or what overall that this there is a service that cloud computing which provides different type of service on a subscription base right. That means, you subscribe for that service right.

So, that like it can be infrastructure as a service, it can be platform as a service, it can be software as a service. So, you pay as you go type of model right. And it is theoretically infinite level infinitely scalable that means, if you go on demanding it go on increasing the resources and in both side that means, you can go on acquiring resources, you go on releasing resources and also it is ubiquitous right like anywhere everywhere it is available. The major challenges what we faced in this cloud thing what we can perceive is that you require a good connectivity right quote unquote good connectivity that means, you should be always connected when you are there. And that is one part another part is that anything it has to go to this cloud and come back right.

So, this there are several service providers these days we have many service provider to name a few of them like Amazon, Google, IBM, Microsoft and many others right. And you can have your own organization cloud or community cloud also. So, in other sense so, how this

service provider maintaining those cloud they are infrastructure they need to have some infrastructure right. So, they are maintaining across different geographical spread or what we say that more the this cloud data center or CDCs right. So, there are data centers across the thing.

So, now, your this traffic say I am sensing the temperature of this room and want to see that the whether the temperature within this limit then that I need to collect by some IOT devices right or some sensors which collect this thing, push it to this back end cloud wherever it is and it takes a call based on that what is the temperature required for a say so called studio like this. So, it takes a call and then revert back whether temperature normal or AC needs to be switched on or needs to be switched off or temperature needs to be increased and type of things right. So, what we see that this add to this latency and some of the things like as I mentioned earlier also this could have done in a much lower level right or not need not have to be traveled so much there could have been a device or the network device in this which is connected with the sensor node could have taken a call because this is may be a trivial thing that this temperature is within the thing or in other sense or at least it can aggregate this thing like instead of sending all this if there are 10, 20 sensors not all the sensors, but say average or some sort of a average along with the mean along with the variance and type of things which can send that this is the situation of this temperature in this room and the first of all the data load is less and it also do some downline the calculation. So, something at the age could have taken a call right or something what we say that may be at a intermediate level may be the IIT Kharagpur router which is not in this particular department somewhere else in this institute could have taken the taken some decision on that or that whether the some sort of actuating is required itself. So, at the fog level and then if something is required or this aggregated thing could have gone into the cloud level.

So, we can have this multi tier things right. So, that I am not going into the again the full details. So, fog as we have seen takes the cloud closer to the data producing sensor devices, devices such as routers, servers, switches etcetera which act as a fog node if the processing power is employed for the data processing and result analysis. So, intermediate things can act as a fog node. So, what so use of fog technology for real time or near real time.

So, that the your overall latency to travelling to the cloud data centre and coming back may be reduced. Aim to develop a fog based health care model in this particular case study what we try to do. So, collect data at the edge devices to reduce latency, network usage over all things right. So, there are lot of advantage once you bring the things to the edge right. So, that means, you are less latency.

So, your turnaround time is less and you can you better uses of the network menu there is no unnecessary wastage of network menu is an overall cost incurred in the cloud may be less at doing the things. The other side is that that your edge devices should be smart enough in doing this right that may be a resource requirement at the edge. Had it been

cloud you are send it and forget it right not forget it per say, but it is the responsibility on the things right. So, the whereas, even you bring it to the edge those edge devices should be smart enough to handle those things right. There are as these days the edge devices are becoming smarter even whatever the switches etcetera we buy these days at least at the organization level are much more resourceful doing lot of processing.

So, there may be a issue or there may be a chance of having a better performance using this edge and intermediate fog and the cloud. So, this some literature or from the internet we have taken this picture the references are at the end, but what we see the overall. So, at the cloud at the top level right then we have fog nodes then we have that edge layer and then we have this IoT's or the that at the bottom layer right. So, what we see that mostly this cloud is mostly involved in more of a so called business analytics and intelligence. That means, they are doing at the much higher level.

So, doing something at a much higher level of business analytics and inclusion at the cloud level and then we have this fog layer which is down below which is more which leveraging on local network data analysis and reduction of the control response virtualization standard and externalization this can be instantiated in this fog layer. So, it comes down. So, you can have some sort of a low level data analysis and things like that rather these days what we are looking at not only the fog node or more of a fog network which can communicate between themselves to take account right. And once we come into the edge because of this huge volume of IoT related data or being sensed and pushed to the cloud. So, the edge may be a smarter way of that some sort of a real time.

So, large volume of real time data processing at source more or less. So, it is at the edge on premise data visualization some industrial PC's embedded system, gateway, micro data storage all those things can act as a thing. And at the down the line we have the sensors and controller where are the data is being originated from right. So, we have all these four layers and if you see these are not isolated things right these are not isolated in the sense not that they are definitely connected, but isolated in the sense they are processing say what say at the edge level processing is going on that is related to that what processing the fog layer will do that is related to the cloud and other way around right. So, this over overall system should work in appropriate what we say proper orchestration right otherwise the overall gaining this overall efficiency in terms of say latency, in terms of energy, in terms of say some sort of a real time application visualization and other things that may not be feasible unless they work in a integrated fashion.

So, what these days are more what we look for is more of a cloud what the term basically you find in a cloud to things continuum right. So, it is a integrated framework rather than a piece wise layers right. So, rather if some of the other things are also becomes important see if the information is being percolated. So, at the ground level when the information is going up there is a issue of what we say compromising on the privacy of the data or rather if I take this data to the end then I know that at different layer I know that where the data is

generated what type even what type of sensor is at the different level, but if I do something at the age of my own network. So, I basically I am forming some aggregation.

So, that my automatically if there are some sensitive information those can be protected right. So, in going out of the network in number of cases that may be a issue though as of now we want to look it more of a in terms of that how good in this in terms of energy in terms of say what we say network delay and so and so forth right. So, those are those are the things which are more things. So, if I if we see in a this sort of a integrated fashion this may help me in a better management. So, the same thing if we now see this hierarchy at the end of course, the same things are there it may so happen this your at the this IoT devices and these edge devices means IoT devices itself can have some computing resources and it can be able to push those thing at the up in the line.

So, as we are discussing so the limitation or I should say the challenges in cloud is that the latency in several cases is pretty high, large volume of data being generated at the cloud level means at the sensor level and which are need to be transmitted to the thing and it in turn requires more network bandwidth right. So, not only the network delay the network bandwidth requirement is also high in order to have appropriate response time right. So, these are the serious challenges though the though the cloud do have enormous computing ability with definitely a huge support on the storage side. So, that is not there may not be major issue, but these are the serious challenges especially for applications which are which demands real time or near real time response time or near real time responses right. So, if we look at the other side at the edge of the thing is the IoT devices.

So, their challenges are definitely the processing power they have a limited processing power storage power requirement and at times they do not have a global view of the thing or at least a more larger view of the thing right. So, I for example, that say if I consider a large room with 10, 20, 10, 20 temperature sensors and say one sensor saying that the temperature or say 2, 3 sensors one corner of the things are saying temperature is normal whereas, at the other end there may be some challenges due to some there may be some due to some say heating effect due to maybe for the server or some other things and the temperature is not normal right. Now, individual sensors are looking as they are own small region. So, it may not have a more so called global view of this room that high overall things are there, but if I could integrate this whole set right this 10, 20 sensor and go little higher level that the switch which is connected with this things or the server which is connected with the switch which is connected which is basically accumulating the sensors of this room that may take a better call of the things right. It may see not only the individual sensors, but also see that what is the different mean median or mean and variation of the temperature within the things and if that is beyond the model which to be there it can do.

So, this is a very maybe may not be a very what we say complex example, but there we can do in we can see that there are need of a little higher level of abstraction of the things right. If I go little more higher level like all the labs of this building all the labs of the institute then

I may go little more higher level maybe at the IIT Kharagpur router end or things like that and then we can take a better call of the things right. So, what we see is the same this hierarchical layer helps me in reduced latency support for real time applications right. So, if I can make this type of fog and edge computing in between. So, it can support less network congestion because if more traffic are going at the more towards the cloud.

So, it is a tapering and more network congestion and may reduce the cost of execution at cloud right. So, cloud you need to pay it is a subscription model as we as we know. So, you have to pay more and if it is down the line then it is already a resource which is surplus in my premise and we are exploiting that and better handling of some closer data and generated by the sensors and this means cleaning and other things more of data location aware things right more of location aware application and things like which are more location aware stuff right. And of course, there are issues which are related to the data security and information security which can be. So, this hierarchy level really helps and if we look at this particular say what we say this demonstration or this case study what we are going to discuss in more details is a fog edge cloud fog edge IOHT that internet of health things overall framework.

The overall objective of the thing was to design a fog edge computing based health model to reduce latency network usage and cost incurred to the at the cloud right. So, whether we can have this for health related applications to test and design the fog model using iFog. So, we use the iFog sim simulator. So, this simulator to have a simulation. So, what we usually try to do that once you try to realize things before going towards the hardware you want to see that how is the feasibility and the simulator really comes handy not only here in different applications.

And then we try to develop a customized wearable device for collection of the health parameter and to implement the proposed model over hardware and test its efficacy right. So, that whether we can have a hardware to test to study the due based computing and study its efficacy in the proposed health scenario right. So, whether we can have a due based computing framework and propose study the efficacy in the health things right. So, couple of things which come up one is that we want to see that whether this type of this integrated thing to implement and see that how whether they are efficient for say cloud only type of scenario type of things right. And then have a simulation environment and then go for a hardware implementation and then to have another that due computing concept we try to look at those type of things.

So, if we look at the overall workflow of this over this particular case study. So, conceptualization and the modeling designing and implementing the FOG model followed by performance evaluation using iFOG sim. So, just to by a simulator that is iFOG sim simulator to just see that what is the how the whether the model is workable or not. So, results are in infernace from the simulator then we try to make a hardware implementation of the simulated model including fabrication of a customized body sensor and activity using using

accelerometer and a what we say heart attack alert what we say, but to make it very clear that it is this the type of algorithms or type of things what we want to do it is has no medical and clinical implication and has been used only for a demonstration purpose right. So, it is only for a case study we did.

So, actual things is required lot of other things like domain knowledge from the from the doctors and things like that all those things will be required and then activity detection and using accelerator this that activity detection that means, whether we can have some of the things like accelerator gives you that the some notion of the movement and there are other health parameter give some other notion from there whether we can converge to a things that where what is what is the going on and then results and inference and we also proposal for a new framework what you will share and then inference from the whole thing right. So, primarily want to look at a motivating case study and see that whether by using this sort of arrangement whether we can have a right whether we can have a efficiency in terms of say delay, overall costing, bandwidth utilization, energy and so and so forth. So, this is the simulation using this iFoXim simulator. So, though if you are interested you may download the simulator it is a open source rather iFoXim was initially the base version was developed in a collaborative team with IIT Kharagpur and University of Melbourne. So, and later on now is being maintained by University of Melbourne for this FOG simulator and being widely used by different this particular this IAM or cloud FOG edge community.

So, here you have so, I am not going into these details of these because this is more related to the simulator, but definitely you have different component like module placement, mapping, module placement and then FOG device, this edge app and application controller circuit and things like that. So, you can have a simulator which can work and we also looked into this type of hierarchical topology where at the down level you have that IoT devices for our case study is IOHT. So, internet of health things then at the other end you have the cloud or the what we call level 0 intermediate there are FOG devices it may be the area gateway or the ISP area these are the area gateway and then we have the edges. So, it is layer wise. So, at the top layer 0 is the cloud then we have the FOG layer edge and then the IoT.

So, this way it can be realized. So, that what we look into is the network topology in this fashion and we use some typical configuration cloud FOG edge IOHT. So, typical configuration. So, if you see that if we have this cloud ISP area gateway and mobile. So, different type of things and then what we see that the cloud is definitely more resourceful like in terms of say MIPS or RAM or even the this overall power and type of things it is much resourceful whereas, if I see this mobile edge devices these are less resourceful, but nevertheless what we see that if we look at an integrated fashion then we can have a better response. So, source or body sensors mobile area gateway operator ISP mobile device for the display when in the it returns back to the things like I say alert and then I return back to the device for the things right.

So, these are the latency sources. So, where it will be delayed and things like that and we taken some references like if it is a cloud the latency is the maximum because you need to travel a much further distance and whereas, if it is a edge in the mobile or this body sensor the minimum latency. So, it is very near and it may be the minimum latency is the right. So, in this case what we use is the mobile as your as collecting the data from the sensors and then pushing it to the rest of the things like fog and things will come to it. So, and also mobile is used as a as you are telling that it is giving some sort of a alert.

So, the alert is displayed on the mobile device that health related error that it is normal or you need to visit a doctor and things like that all those alerts are provided in this using this mobile device. So, if we look at the overall workflow. So, one side is the body sensors or the which is collecting information from the body for this particular application because it is a what we show internet of health things with the body sensor. Then the client module, then the data filtering module, then we have data processing module, event data event handler module and it comes back to the cloud module and the display of the things right. So, this is the path and there may be a conformity module that if you are predicting a event then who will confirm it may be confirmed with the with some other knowledge base or type of things which may reside in the cloud and things like that right.

So, or rather if there are things it is reporting that logging the data there for further use. So, if you see this is the overall flow of the things like how the process flows. So, it collects data and put the client module which may be in the mobile, then put the data filtering module with the next step in the fog nodes and data processing also can be done in the fog nodes or it is pushed to the cloud and then come back into the thing. There are several other things like you need to data cleaning and things like that it is not that the data are all directly fed into the processing unit. So, if you look at the application placement the client module is placed in case of a fog based model in the mobile and in case of a cloud based model also in the mobile devices right.

So, the client module that what we here we are considering as the age whereas, this is the data filtering module that is area gate it is placed in the area gateway whereas, in case of cloud only scenario it has to be pushed to the cloud right. And data processing module also in the area gateway can be at the area gateway or in the cloud only it will go to the cloud in case of a even handling module it can be in the area gateway. So, that is much lower or much nearer to the source and we have in the other sense other than the this age or the where it is sensing all are the all in the cloud based model all are pushing to the cloud right and the confirmity module is both the cases cloud. So, what we see that what we see that in case of a this age fog cloud type of things that there is distribution of the work is there whereas, in the cloud only module it push to the cloud. There are one way it sounds pretty efficient in the terms that I bring it down, but definitely you require resources you require proper orchestration of the things how much to be calculated here and a double neighbor and hierarchy type of things who will respond to whom and how things will be generated where it is if it is a like virtually some central place like cloud then it takes the call overall that is

there is a challenge that in the terms of a resource and things like that.

And then there are also involvement of processing etcetera. So, though we do not bear the cost of the cloud, but there are some of the costing out here also, but nevertheless what to our what our say proposal is that as if you are having resources which are surplus can be utilized. So, this can be there and of course, this configuring this age fog cloud type of infrastructure is much more complicated than a cloud in cloud only scenario right because that is already there here you have lot of loading on your on the part of the individual or the organization to set up this type of things right. So, let us stop the discussion here and rather in the subsequent talk or the subsequent discussion or session what we will do we will continue with this and see that how keeping these premises how we achieve the this IOHT for age for cloud type of paradigm which allows which allows us to have more efficiency in over overall which allows us to have better efficiency in terms of say latency in terms of bandwidth utilization or network utilization in terms of overall costing other things right ok. So, let us stop our discussion here. Thank you. .

**Cloud Computing**  
**Prof. Soumya Kanti Ghosh**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**

**Lecture 60**

**Case Study II (Internet of Health Things) (Part-B)**

Hello. So, we will continue our discussion on cloud computing more specifically on the case study we are discussing on internet of real things. We discussed last session last class and we will continue that. So, what we try to look at is that a health or medical application using IOT devices and other like fog, age, fog and cloud right. So, it is more of looking at the thing like cloud to things or cloud to IOT paradigm or the continuum how it works that we try to look at right. So, this work was primarily done by one of our one in one of our M.

Tech project under my supervision and there are involvement of other research scholar B.Tech students and type of things right. But we implemented this work in our lab here in computer science department. So, as the name suggests is cloud, age, fog computing for IOT application on medical or IO internet of health things.

We have seen little something on internet of health things what do you mean by this thing. So, we will have couple of slides which may be repetition from the previous session just to have a continuity. So, what we what is the overall objective of this particular case study or the particular this work is that to design a fog age definitely cloud one side and IOT other thing is there based health model to reduce latency network uses and cost incurred in the overall cost incurred in the system or cost incurred in the cloud. So, what as we discussed if you remember in the previous session that once if some of the computation if it is possible to bring down from the cloud to the lower level then it may help me in better in terms of latency or reduction of latency better network uses and the overall cost incurred if you are trying to run something on the cloud based system right. So, we tested this overall system in a with a simulator which is iFogSim as I mentioned previously.

So, this is a simulator which was developed initially jointly by IIT Kharagpur and University of Melbourne, but now it is being maintained by the clouds lab of University of Melbourne and that you can get the it is a free simulator to work on right and it has a back to back say sinking or linkage with this cloud sim which is a cloud simulator. So, to develop a customized wearable device. So, the objective one other objective was to whether we can device wearable device right for collection of health parameter from our body like we have designed a health band which can collect different sensor information and can transmit to the to the cloud via taking that mobile as your initial say age then going to the other fog like server or raspberry pi sort of hardware and then it can push to the cloud right. We use both private cloud which is in campus, but we are having that Megumala that IIT Kharagpur own cloud and also with a public cloud with that AWS that is Amazon cloud

to show that this overall working of the things right it is to demonstrate the things. Now, if you see that there may be may not be very great per say research challenge right, but there is there are implementation challenge definitely and also it tries to showcase that using this type of hierarchical structure may help me in reducing latency, better network utilization and also overall cost incurred using this cloud based systems.

So, that was our that was our motivation of doing this and to also at the end there is a study of DEW based computing like how we can have this structure along with the DEW servers. So, that overall efficacy can be to give a better efficiency of the overall things right in terms of we will see that in terms of storage and things like that, but we will show this primarily this IoT to cloud type of things under the health domain. So, as I believe I discussed in the last class also. So, that the how we approach the problem you can have any problem actually the overall approach may be the more or less same like conceptualization and modeling of the this problem like in our case the problem is that to collect health related information from using body area network or sensor body sensors I should not say overall band, but body sensors and then to study that health parameters and give some predictive or alarm type of things right. So, here we considered a called heart attack alarm or heart attack prediction algorithm which has been there it is something which is basically look you look at as a algorithm right no or look at as a method or procedure nor as a on the health point of view.

So, it has no medical or clinical implication neither tested and has not used for and this is being used primarily for demonstration purpose right this can happen actually in real world also we can real life also we can have this type of things. So, we have this conceptualization and modeling of the problem design implementing the FOG model followed by the performance evaluation using iFOCSING. So, initially did a simulation. So, usually many of the things we try to what we try to do a simulation study and find out that what are the different parameters and then actually go for actual implementation or hardware design right. So, results and inference from the simulation and based on that we implemented the hardware implement hardware we design a hardware implementation of the simulated model including fabrication of customized body sensor network right.

So, customized body not body sensor network customized body sensor right. This is just different health bands are available which have lot of parameters, but extracting data from them is for some other purpose is always a challenge. And also our basic goal was that whether I can have the sensors based on the my requirement my requirement in the sense that particular medical requirement and collect the data along with coordinate and things like that and whether I can have a end to end so called internet of health things to the cloud this overall the realization of the overall system right. So, what it does activity detection using accelerometer. So, there is a accelerometer sensor by the activity detection and also heart attack alarm generation or prediction module type of things right.

So, these are the two primary things and again drawing the results and things and then we

proposed a new architecture based fog model to enhance the overall robustness of the system and then finally, looking at the inference drawing in terms of things. So, we did lot of case studies try to rather lot of experimentation try to see that how the overall performance of the systems or how the overall system behaves and that may be a good thing to look at and you can always try with other type of sensing and actuating things right. So, this is overall simulation using iFoxy if you have used iFoxy you know that what are the different things it requires there are module placement mapping module mapping controller and so and so forth we are not going into the nitty-gritty of the things, but it shows that overall how it works that component of iFoxy. And this is the network topology what we looking for. So, there are multi layer at the bottom layer is the IoT or IoHT for health application what we are telling IOHT in our case that in this case and then at the top we are having cloud.

So, it is if we consider as a level 2 then we have level 1 as a FOC then level 2 as edge and IOHT or level 3 as the bottom line. So, this is the overall structure overall hierarchy of the things and if we look at this application or the analysis. So, some of the things if the IoT's are having their own resources for doing any processing then it can be done as a at a much lower level right. So, where we can have this at this level we can process, but what is happening it is only looking at its own means own sensors and things like that. If something to be done at that ends fine otherwise we can have this at the edge level where the mobile devices are acting as a edge right.

So, at the edge level where it can takes more than input from more than one such sensors or body sensors and take a call and then we can have at the FOG level which may be the area gateway or something in between this edge and the cloud and then finally, pushing to the cloud taking a things global decision or over a region on a larger geographical space. So, what we see that some of the computation is can be brought down to up to this edge level at least. So, it will have less round time less latency in this round trip time it is not crowding the network at the upstream like near the cloud and so forth right. And overall if your cloud and other things are having posting like you are subscribing the cloud then you have a better way of handling the things right you do not have so much crowding on the things right. So, this can be one stuff one thing right.

So, there can be cloud FOG edge this what we are talking about. So, if we look at the for the experimentation. So, these are the consideration. So, this device is cloud and MI MIPS and the bandwidth that how much upstream bandwidth, how much downstream bandwidth and what are the different at what level what we designated that and this cost plus MIPS and so and so forth right. So, what we see these are the some of the things taken from the literature which we try to make that it make sense that considering this all this consider all these parameters.

And also we consider that the if the source is body sensors and destination is the mobile the latency is 1 that is one hop sort of a things mobile to area this is 2 area to ISP

gateway area gateway to ISP gateway it is again latency is 2 whereas, ISP gateway to the cloud it is more thing there is 100 and displaying the mobile from the things it may be some latency of 1 and so. So, there is also taken from different considering different literatures, but it make sense that if we consider this type of things you can have other parameters as well. So, if we look at the overall flow the so, it body sensors from the body sensors this client module takes the data it is data filtering module which filters the data and sort of a cleaning the data and then data processing module event handler model that in case of a event what it should respond and then it come back to the client module and display on the display unit of the mobile device of the user. So, it is showing that like if you are you need to visit you are normal or wish to go to visit to doctor or and so and so forth right. Some of the things some of the alerts we also see in our present day commercial that health bands right so, called health bands where it is give some alert that based on your body parameters right, but these are all more customized and sometimes costly and it is the data you cannot extract very easily for other processing right whatever they are showing on the display it is the things.

There may be other things like privacy issues that extracting the data in the raw format right ok. So, that those are. So, if you look at the application placement in the client module in our case if we look at this processing the client module is mostly is in the mobile or the edge device in case of fog means edge fog cloud infrastructure right whereas, in the cloud it is also in the mobile device that the cloud module should run at that time whereas, the data filtering is maybe in the fog level in case of this type of fog edge based system whereas, in case of cloud based is in the cloud. Similarly, data processing maybe at the fog level which is in the cloud data event handler can be in the fog whereas, in this cloud and of course, this confirmatory module that like you confirmatory or taking a what we say metascale or at a more area scale processing those are at the cloud or placement of the cloud based model. So, this can be the things which are there.

Simulation configuration is something like this like you have this cloud and ISP and thing. So, what if you see like the blue is the what we see as the this cloud fog edge or cloud fog edge IoT type of configuration where the red is the cloud only scenario and these are the things what different simulation scenario with 4 users, 8 users and things like that. And if you look at that different type of results which we came up with through this type of data configuration or what we say that data sets what that in the simulation environment. So, what we see that for different configuration like these are the different configurations what we see that in configuration 1, 2, 3, 4 and 5 the number of user varies and number of that number of area gateway. So, what we see that based on this average latency once this configuration is on the on the higher configuration we are going 4, 5 things number of gateways etcetera increase.

So, what we there are there is a cost in the cloud increased drastically right substantially in terms of the delay whereas, it is this fog is more or less in the constant level right near 0, but it is a fog by fog based system or in the sense fog edge based system it is more the things

right. So, latency is fixed as the application module which from the part of the control loop are located at the area gateway itself and the modules are located here the modules are in the CDC or that cloud data center right. Performance evaluation network uses so, here also we see that in case of a network uses the costing is much much performance in terms of network uses in terms of k by means kilobytes are much higher than the when we have fog edge based systems right. So, that is also you can see and if we see the cost of execution. So, what is there that the only resources on cloud incur cost other resources owned by the organizations right.

So, the here the hunch is that the organization is only subscribing the cloud rest of the resources are with the organization. So, in other sense this the cost primarily is dictated by the subscription cost of the cloud. So, that is why we see that the cost of execution this cloud only system is much much higher than the when you have this fog edge type of structure right. So, more processing at the cloud leads to higher cost in case the cloud based in custom cloud based architecture. So, it goes to the cloud for all type of processing and then it gives a cost of execution and you need to pay more subscription charges for this cloud.

Now, if you look at the in case terms of energy consumption. So, energy consumption also if you see. So, that is the cloud energy mobile energy and when we have some sort of a cloud fog type of thing. So, energy consumption at the mobile device remains same in fog as well as cloud as the load does not change. So, it is more or less same whereas, the energy required the fog devices and the data centers so that means, that is red is the your mobile energy which is more or less same right when we are using fog and things.

But whereas, in the mobile required this energy requirement of fog devices and the data center changes as the configuration changes from the fog based to cloud based architecture owing to the shifting of the application module right. So, that what we see that the energy consumption of the mobile devices. So, there may be there is some little typo I believe that is the blue will be the mobile devices which is more or less constant whereas, this cloud energy when it is at the fog level cloud related things are much less whereas, where the configuration 4 is the cloud only situation or when you go to the higher configuration is more towards the cloud. So, it the requirement is much higher right. So, you read it that this more or less the blue is the mobile energy which is mobile devices which remains more or less same because they are doing the same processing type of things where the cloud goes on increases is expected right.

So, what we try to see here if you try to see all these results. So, what we try to see that more or less it looks like that the cloud may be always in the means down part, but that there are some other consideration per say like in case if you are doing this you are maintaining edge fog and type of layers the challenge is that you need to maintain all those things at your at your premise right at your what we say it is all are in your control right or you need to manage them right. So, but wherever you subscribe for the cloud the everything is the responsibility of the cloud per say right. So, your management part is if it is organized

then much less on the thing. So, that is there is another side of the story, but nevertheless if we have surpassed resources the basic hunch is there then there is a possibility that we can use those intermediate things like edge devices and fog devices to overall reduce the latency, the overall cost of cloud or subscription cost and the your round trip delay and this network congestion and things like that right.

So, we planned for a also tried for a hardware implementation where simulated model hardware implementing using customized body sensors right. So, some of the body sensors which are available in the market are made a customized band and type of things. So, customized body sensors simulated sensor data and used as very pi as fog devices and AWS as the external cloud also we tried with our internal mega mala, but as we are telling that this your own cloud this if you go for the public cloud there are lot of services which are being enabled. So, that is a that is a you do not have to write all service level procedures and things like that anyway that is different. So, what we see that we use something like a some sort of a customized BP and pulse meter rather we used a customized box to have this using different sensors which use different body sensor network and also send the data to the external this sensor data through this Wi-Fi and which are which are kept Wi-Fi or Bluetooth both we implemented and it transfer it to the mobile device right.

So, we have this customized BP and pulse meter and node MCU module board, accelerometer and raspberry pi devices using this we simulated this which basically takes the body sensor a body using body sensor the different parameter of the head parameter and transmit to this intermediate fog and the cloud devices to and the in the cloud devices to take different processing challenges ok. So, if we look at the activity of the accelerometer the data extraction some sort of a some smoothening algorithm like 5 point smoothening algorithm feature extraction and then the classifier to classify the data right. So, data extraction the collected data has 3 component x y and z z axis right 3 axis and the computation that is that trivial things we can have this this  $x^2 + y^2 + z^2$  method we have used to reduce any induced noise each signal is obtained as an average of 5 signals 2 preceding signals and a signal itself and 2 succeeding signal right. So, this is using this 5 point smoothening to reduce that any induced noise right. Then we have feature extraction following feature were extracted from the filtered signal that means, what we smoothen out signal that is minimum amplitude minimum amplitude mean amplitude standard deviation energy in time domain energy in function domain these are the things which are being obtained and then we have used k nearest neighbor classifier to classify this data right.

So, k nearest neighbor. So, what again what we try to do that these are the different standard method we have used, but never delays in a in this site of a continuum things, but the you can use other type of filtering techniques other type of classifier if it is supported by the if the hardware supports that type of operations and things like that right. So, no issue on that this is again I am telling this for demonstration purpose to show that this type of things are possible. So, we did one thing called cardiac attack prediction or what we say

some sort of a heart attack alarm. So, have a different features like BP and this diastolic, systolic and things. Then we say we have a simple if then l clause like where if P is greater than 170, S is greater than equal to 180 and this then it then the alarm is true otherwise false and it return the alarm to the things.

The overall functionality is not that great neither we claim so and I also specifically till the nothing no as such medical or clinical implication because this is as to for demonstration proper things, but what we see that I can have any algorithms which can be implemented in that hardware right. If my raspberry pi is able to run that I could have I can do it and it is getting the data from different sensors may be from even different objects or different human being and then different person and can do a different type of call of in this context right. So, this is this is important to look at this sort of a model right. So, this is again I am telling this for the demonstration purpose only. So, then we could not demonstrate it, but what we planned of a or we can see is a due computing as you as we discussed that due computing is an on premise computing software hardware organization paradigm in the cloud computing environment where the on premises computer provides functionality that is independent of the cloud services and also collaborative with the cloud services.

The whole of the due computing is to fully realize the potential of the on premise cloud sensors and the services. So, one of the major challenges what happens that this type of if I have this cloud edge cloud fog type of things specially the connectivity which is out of my own network that then the reliability of the connection matters much right. So, if the if there is a failure of the network or there is a drop in the overall performance of the things then the overall efficiency of the this system reduces. So, due gives a on premise some sort of a provides like on premise computers provides functionality that is independent of the cloud service and also collaborative with the cloud. That means, it can work whenever if even the connection with the cloud is not there and it can basically sync with the cloud right whenever you are doing.

Like as if you are working with something which are in the drop box at the data and then you are or something some drive like Google drive or other type of a shareable things then you can go on working even your system if you are your if it is installed in your PC itself PC or laptop even the connectivity is not there and whenever there is a connectivity it gets synced with that data set right. So, if it is a very mutable data set then we are having challenges definitely, but what happened most of the cases the data set are local to the things right in practice. So, we can leverage on this type of scenario. This is more of a bigger view of the due like where we have that due script for say the one side the due server and this is one sensors and display and then we can have this FOC devices other end is AWS then with the due server accessed when the internet breaks down right that time the due server is accessed and the database syncing is done when the connection is there right. So, it acts as a what we can say your some sort of a local service provider for the whole things right, which is on premise and you can work, but what you are expecting it is not it the connection will be restored within a quick within less time.

That means, you can within a less time due less time the connection will be restored by that time this due can take up the load on the things right. So, this is important things though we cannot show you the any results and demo we this on this, but this type of infrastructure make sense like it is you are always on the processing board. So, if you look at the overall workflow of the things. So, this looks like this that where the due is there. So, it is additionally the due connectivity available or not it looks into the things and that processing and syncing with the cloud come into play ok.

So, if we do a comparative study of the cloud, cloud, FOG, edge and the same thing with due. So, we see that on premise resource utilization is low in cloud, cloud FOG edge is suboptimal and with due is optimal connectivity required, uptime required, bandwidth required and latency is high low similarly infrastructure and then data storage type of things right in cloud the it is ideally infinitely scalable. So, that processing power and the data storage power are pretty high right. So, but what happened for the cloud FOG edge as there are FOG and edge devices.

So, there is limited right. So, you do not have the much liberty to play, but if with cloud with due you have some moderate thing because that you have may be having some more infrastructure support in terms of processing and in terms of storage and processing ok. So, what we see that what we try to impress upon that if we look at this overall architecture like what we see that this cloud to things continuum. So, I can we can expect a overall better performance in terms of latency, less latency, better bandwidth utilization and also some types of a costing, but as I was talking in the beginning it cannot be thought of a disjoint set right. So, it is more of a connected things right it is not like that only my cloud is not required or edge is not required or FOG is not required need to be seen in a hierarchical fashion and try to see that where the processing can be done at a much lower level and also we need to we are trying to see that things like that whether you can have like other features like better may be in terms of security information security point of view like the data you are not exposing the raw data to the external world or the to the cloud and things like that ok. So, this is the case study what we try to do and this is some of the references which you may find it interesting things like this right.

So, if you look at this if you look at this whole course what we tried we started with that basic notion of why this cloud computing as a cloud computing as a subscription based services and things like that and we looked into cloud architecture different deployment model consideration like SLA, economics point of view and also things like how this processing over the cloud like map reduce type of things what we do and also seen some aspects of FOG, IOT and how they and some of these I should say use cases where this cloud comes into play right. So, which are becoming slowly at with different level of different level of means what we say deployment like SAS, PAS, IAS type of things it is now becoming a feature computing feature which is more or less becoming a defacto things which is coming in a big which already came in a big way and we are using. So, rather in the last part of this

course what we looked into that more cloud FOG, IOT type of paradigm what we where we try to see that with respect to resource management or this container based service, sustainable computing and also this looking at the continuum cloud to things continuum how things behaves right. So, overall we at what we tried what we tried in this course to show you that this paradigm of cloud or cloud FOG, H, IOT paradigm which plays much bigger role in overall computing things right or computing era what we are what we are looking at right. Hope this will help you this basic level understanding of the things will help you in by taking more research related problem or looking at more utilization of this type of service in much better way ok. With this let us conclude our session. Thank you.

**THIS BOOK  
IS NOT FOR  
SALE  
NOR COMMERCIAL USE**



(044) 2257 5905/08



nptel.ac.in



swayam.gov.in