



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

CLOUD COMPUTING

Cloud Computing - Overview

PROF. SOUMYA K. GHOSH

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
IIT KHARAGPUR

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."

Cloud Computing Course - Overview

- I. Introduction to Cloud Computing
 - i. Overview of Computing
 - ii. Cloud Computing (NIST Model)
 - iii. Properties, Characteristics & Disadvantages
 - iv. Role of Open Standards
- II. Cloud Computing Architecture
 - i. Cloud computing stack
 - ii. Service Models (XaaS)
 - a. Infrastructure as a Service(IaaS)
 - b. Platform as a Service(PaaS)
 - c. Software as a Service(SaaS)
 - iii. Deployment Models
- III. Service Management in Cloud Computing
 - i. Service Level Agreements(SLAs)
 - ii. Cloud Economics
- IV. Resource Management in Cloud Computing



IIT KHARAGPUR



NPTEL
ONLINE
CERTIFICATION COURSES

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



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

Trends in Computing

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



IIT KHARAGPUR



NPTEL
ONLINE
CERTIFICATION COURSES

Distributed Computing

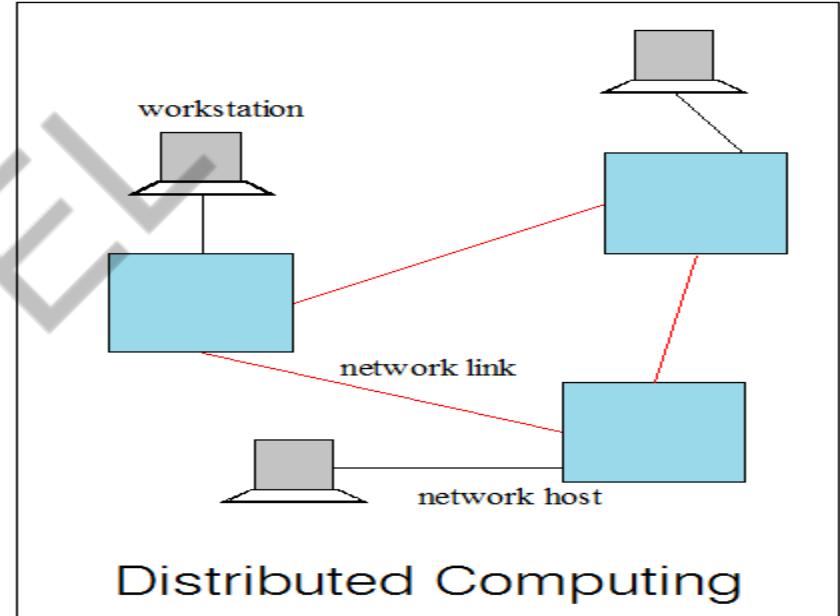
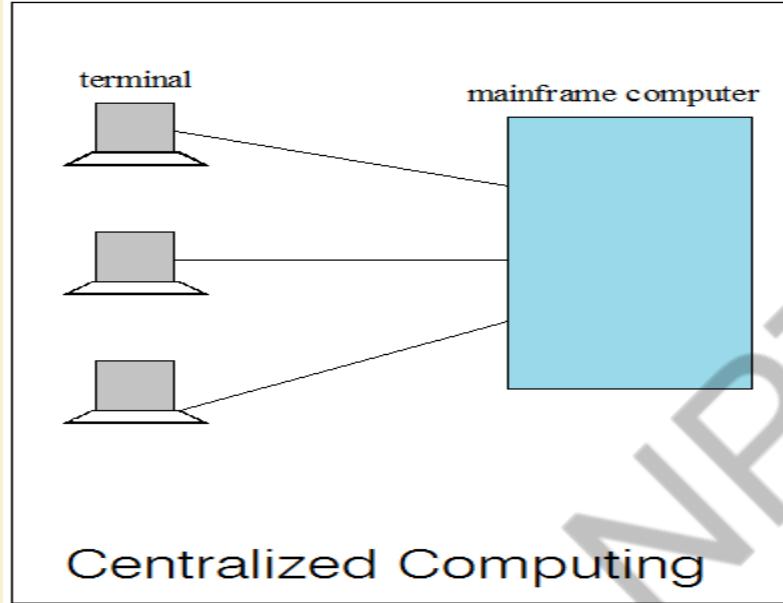


IIT KHARAGPUR



NPTEL
ONLINE
CERTIFICATION COURSES

Centralized vs. Distributed Computing



Early computing was performed on a single processor. Uni-processor computing can be called *centralized computing*.



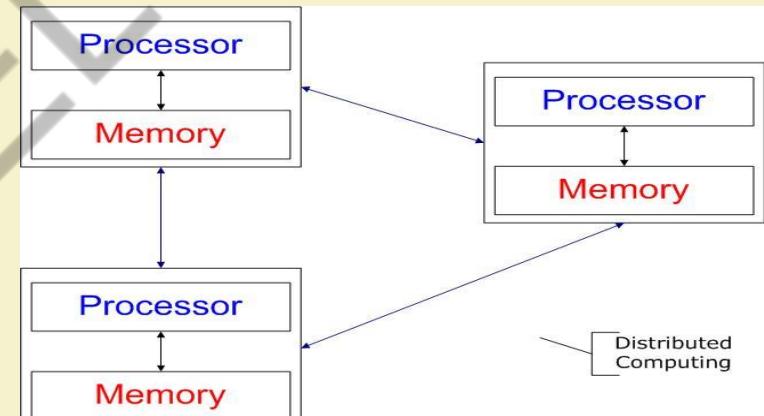
IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

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.



Example Distributed Systems

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



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

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.



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

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



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

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.

Thank You !!

NPTEL



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

CLOUD COMPUTING

CLOUD COMPUTING OVERVIEW (contd..)

PROF. SOUMYA K. GHOSH

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
IIT KHARAGPUR

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.

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

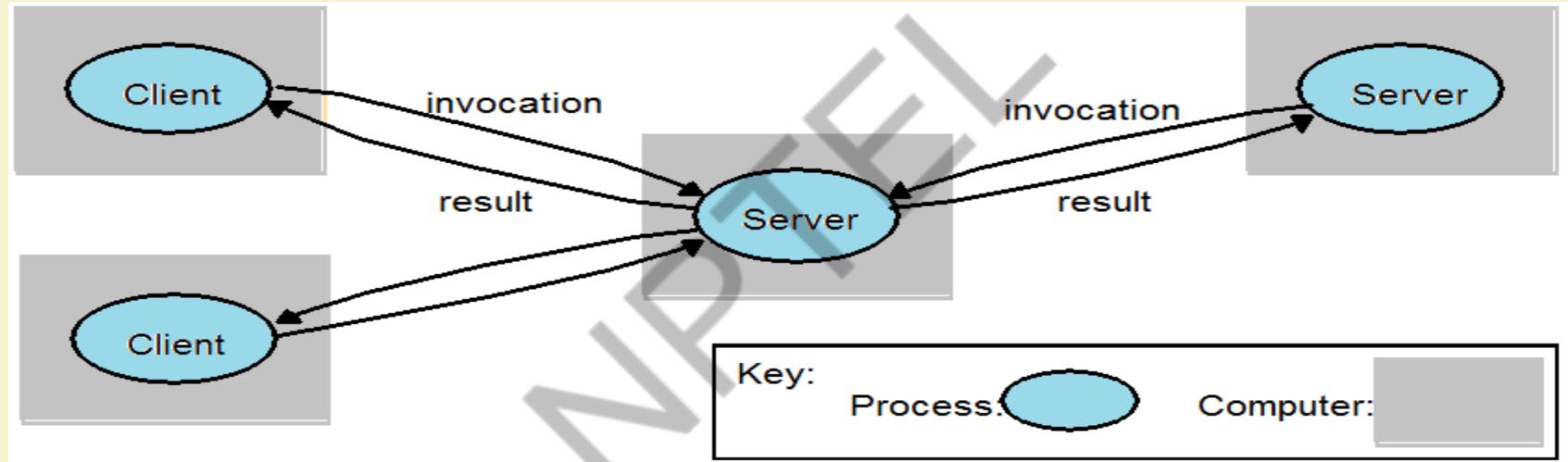


IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

Clients invoke individual servers

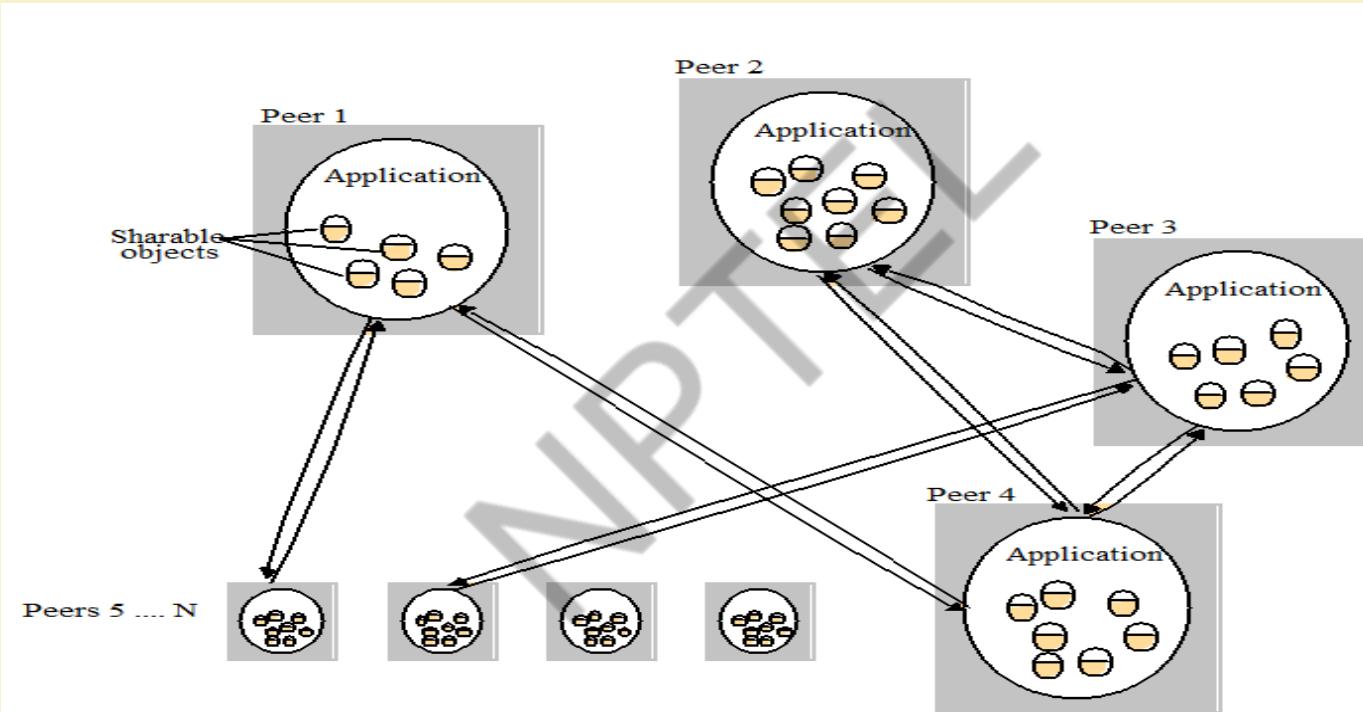


IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

A typical distributed application based on peer processes



Grid Computing



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

Grid Computing?

- Pcwebopedia.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

Electrical Power Grid Analogy

Electrical Power Grid

- Users (or electrical appliances) get access to electricity through wall sockets with no care or consideration for where or how the electricity is actually generated.
- “**The power grid**” links together power plants of many different kinds

Grid

- Users (or client applications) gain access to computing resources (processors, storage, data, applications, and so on) as needed with little or no knowledge of where those resources are located or what the underlying technologies, hardware, operating system, and so on are
- “**The Grid**” links together computing resources (PCs, workstations, servers, storage elements) and provides the mechanism needed to access them.



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Grid Computing

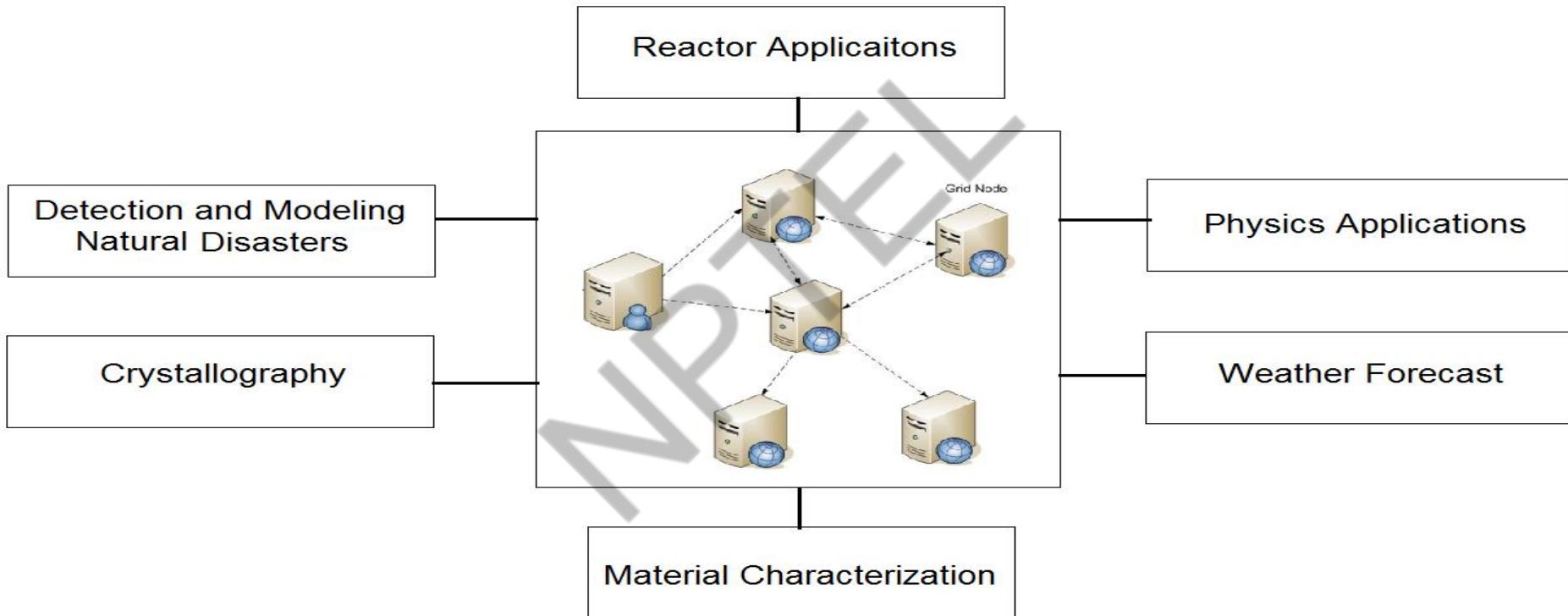
When v use

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 seamless to the user.

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 Mathematical modeling of systems
- 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

Who uses Grid Computing ?



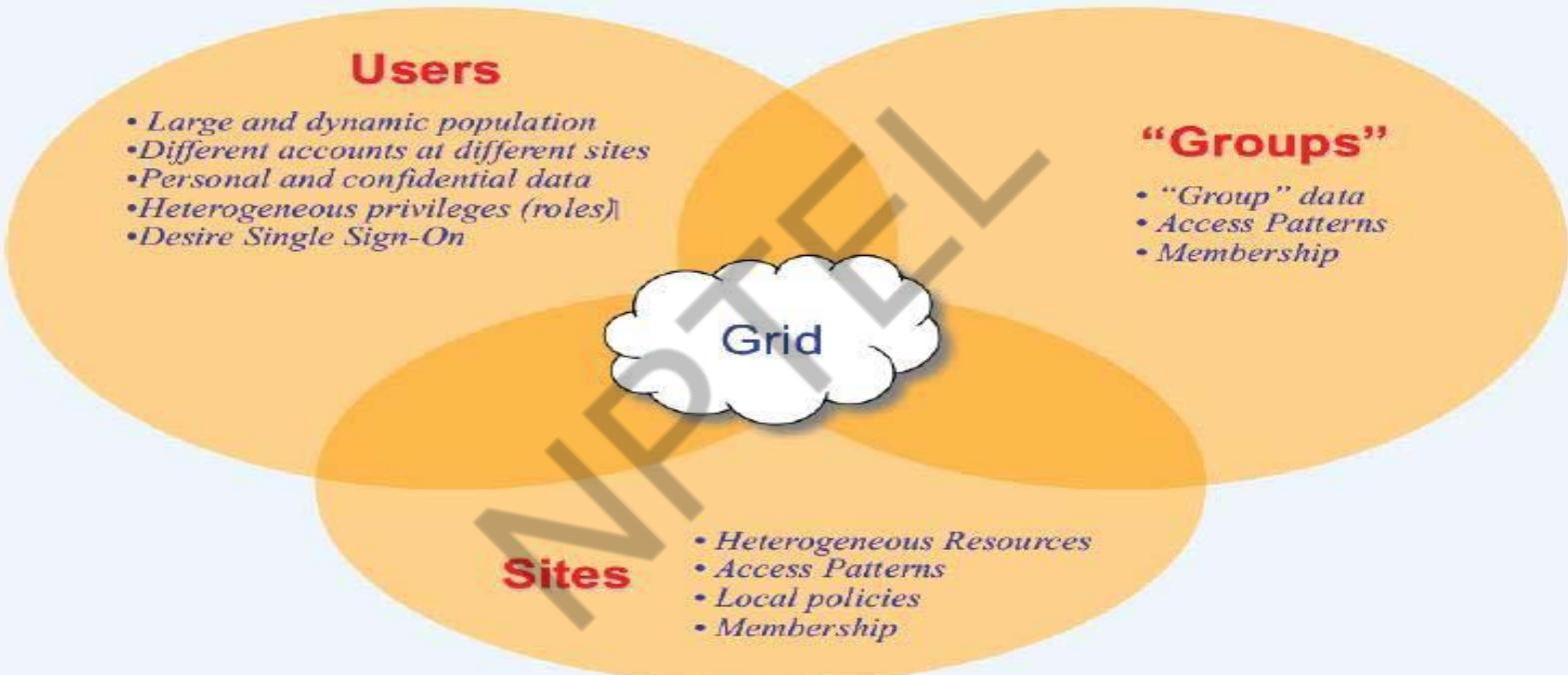
Type of Grids

- **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.

Type of Grids

- **Network Grid:** A Network Grid provides fault-tolerant and high-performance communication services. Each grid node works as a data router between two communication points, providing data-caching and other facilities to speed up the communications between such points.
- **Utility Grid:** This is the ultimate form of the Grid, in which not only data and computation cycles are shared but software or just about any resource is shared. The main services provided through utility grids are software and special equipment. For instance, the applications can be run on one machine and all the users can send their data to be processed to that machine and receive the result back.

Grid Components



IIT KHARAGPUR

Source: Kajari Mazumdar “GRID: Computing Without Borders” Department of High Energy Physics TIFR, Mumbai.



NPTEL
ONLINE
CERTIFICATION COURSES

Cluster Computing



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

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.

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

Types of Cluster

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



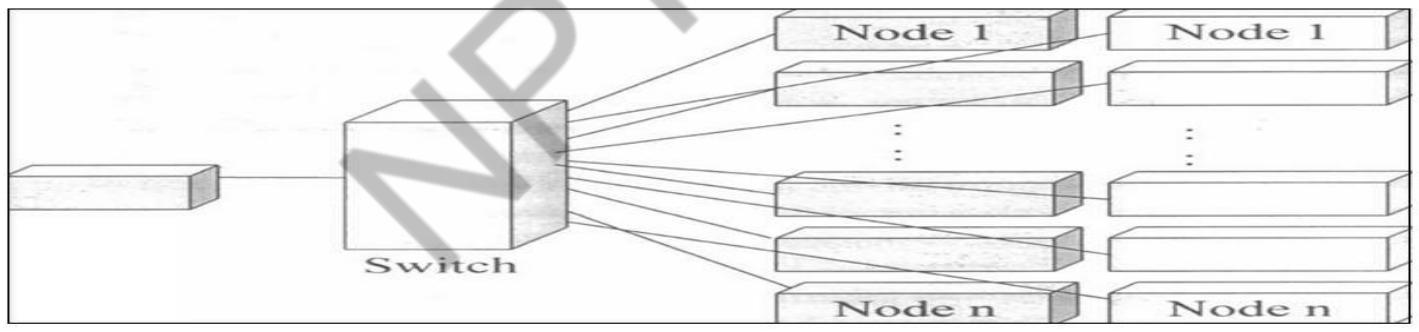
IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Cluster Components

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



IIT KHARAGPUR



NPTEL
ONLINE
CERTIFICATION COURSES

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)



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

Utility Computing



IIT KHARAGPUR



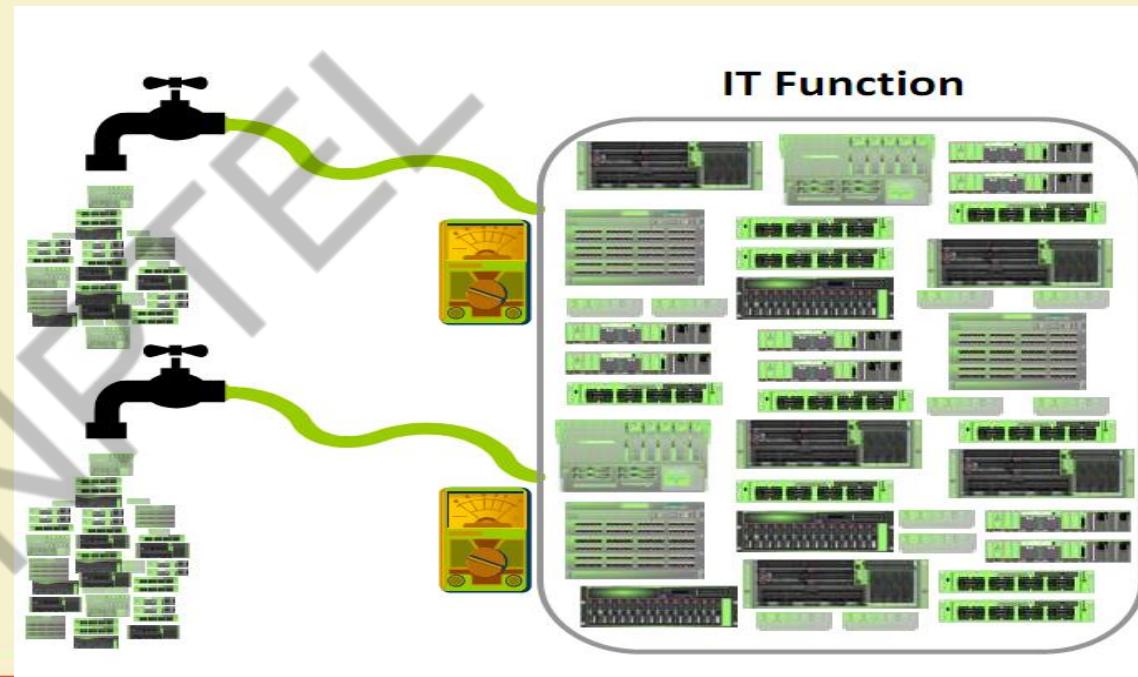
NPTEL ONLINE
CERTIFICATION COURSES

“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*

“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.



“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



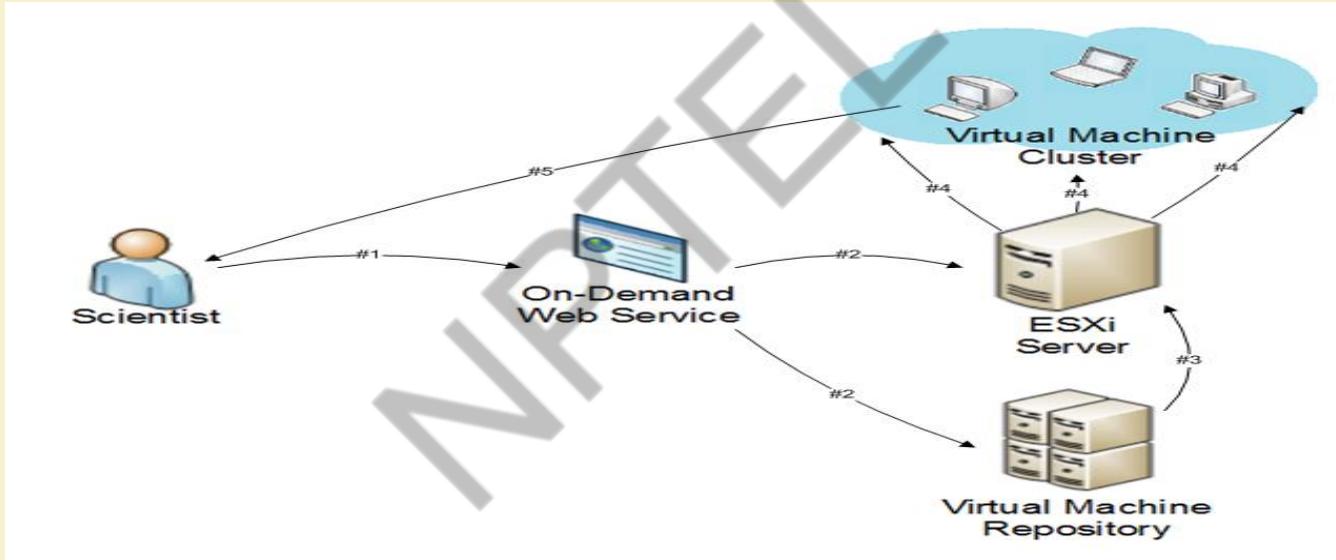
IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Utility Computing Example

On-Demand Cyber Infrastructure



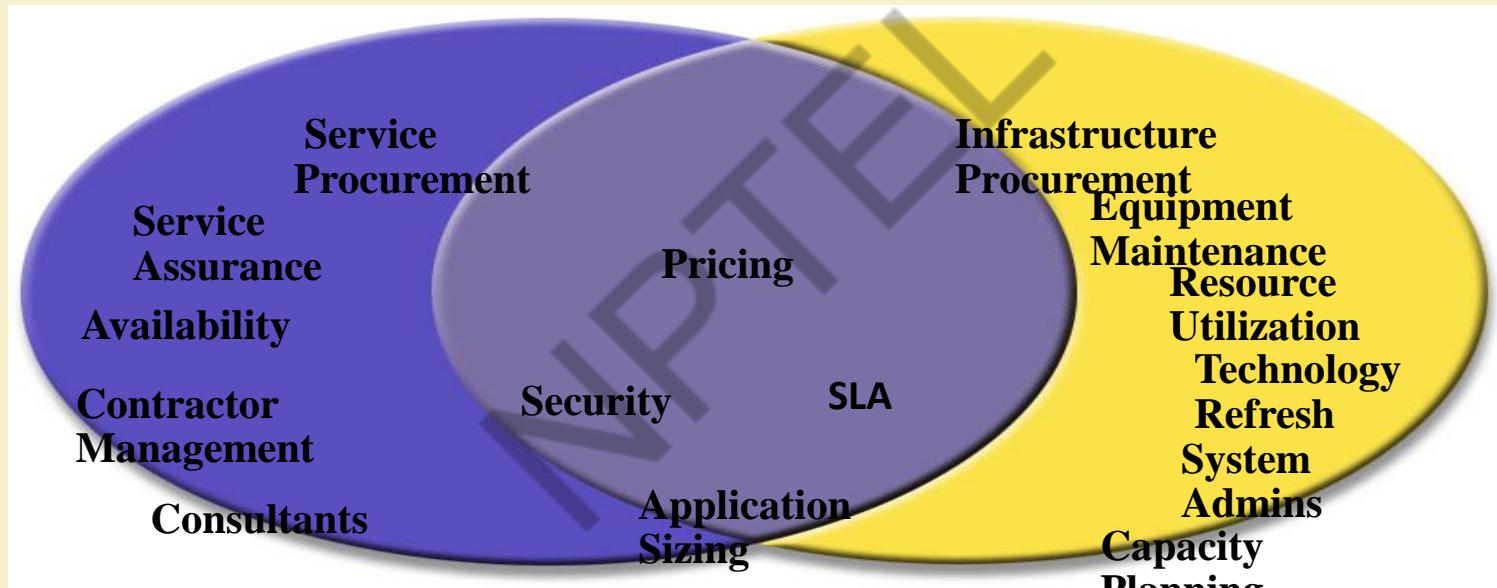
IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Utility Solution – Your Perspective

Consumer vs Provider



Source: Perry Boster, "Utility Computing for Shared Services",
Massachusetts Digital Government Summit, September 23rd, 2004 –
Boston, MA

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

Risks in a UC World

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



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Cloud Computing



IIT KHARAGPUR

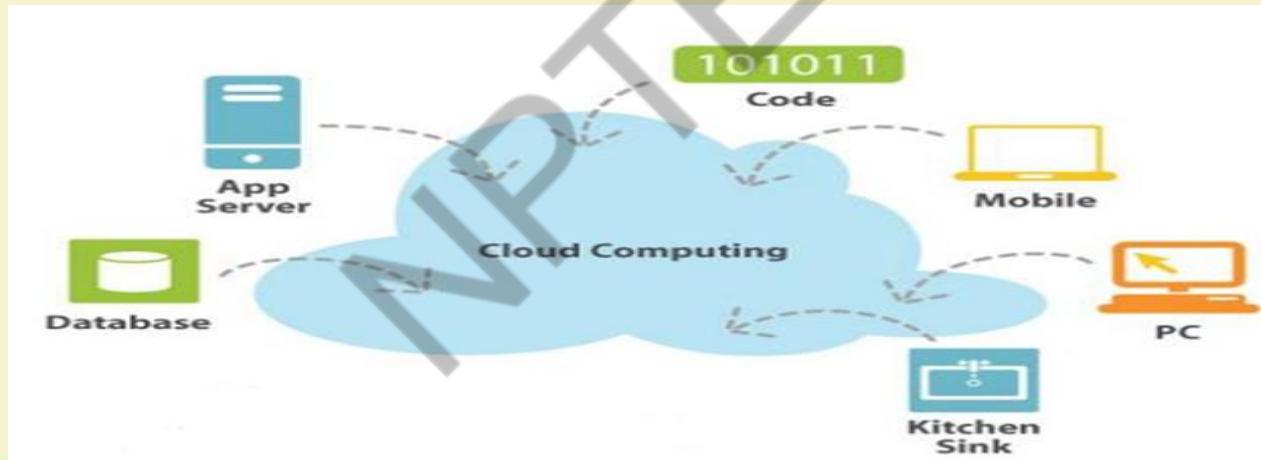


NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

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. ”



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

Thank You !!

NPTEL



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES



IIT KHARAGPUR



NPTEL

NPTEL ONLINE
CERTIFICATION COURSES

Cloud Computing - Overview

Prof. Soumya K Ghosh

Department of Computer Science and Engineering
IIT KHARAGPUR

Cloud Computing



IIT KHARAGPUR

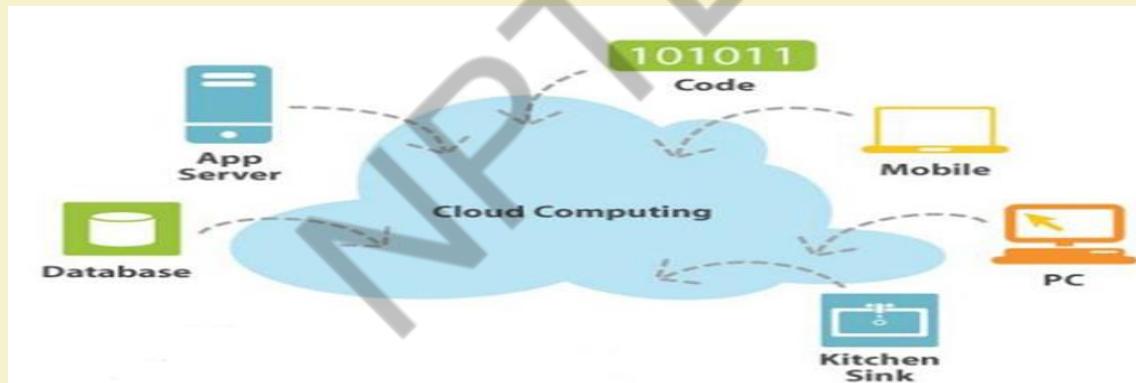


NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Cloud Computing

US National Institute of Standards and Technology (NIST) 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. ”



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

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.

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.

Common Characteristics

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

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)**

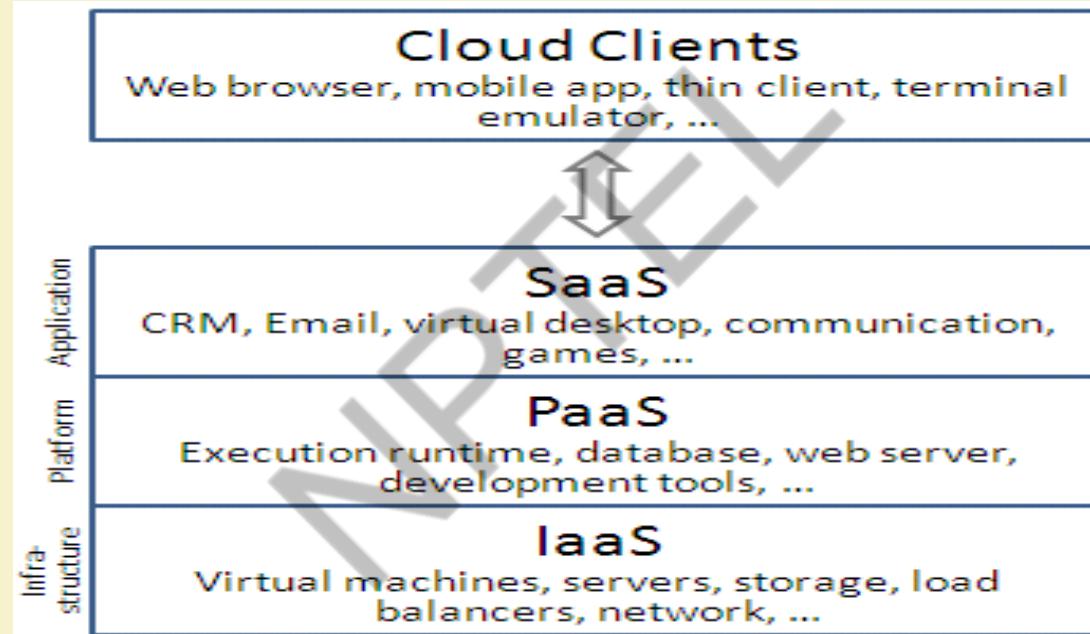
- The capability provided to provision processing, storage, networks, and other fundamental computing resources
- Consumer can deploy and run arbitrary software
- e.g: *Amazon Web Services and Flexi scale.*

Cloud Services Models

Platform as a Service (PaaS)

- The capability provided to the consumer is to deploy onto the cloud infrastructure consumer-created or acquired applications created using programming languages, libraries, services, and tools supported by the provider.
- The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, or storage, but has control over the deployed applications and possibly configuration settings for the application-hosting environment.

Cloud Services Models



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.



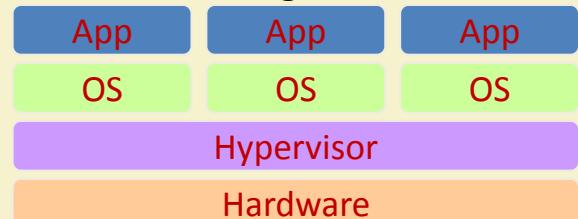
IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

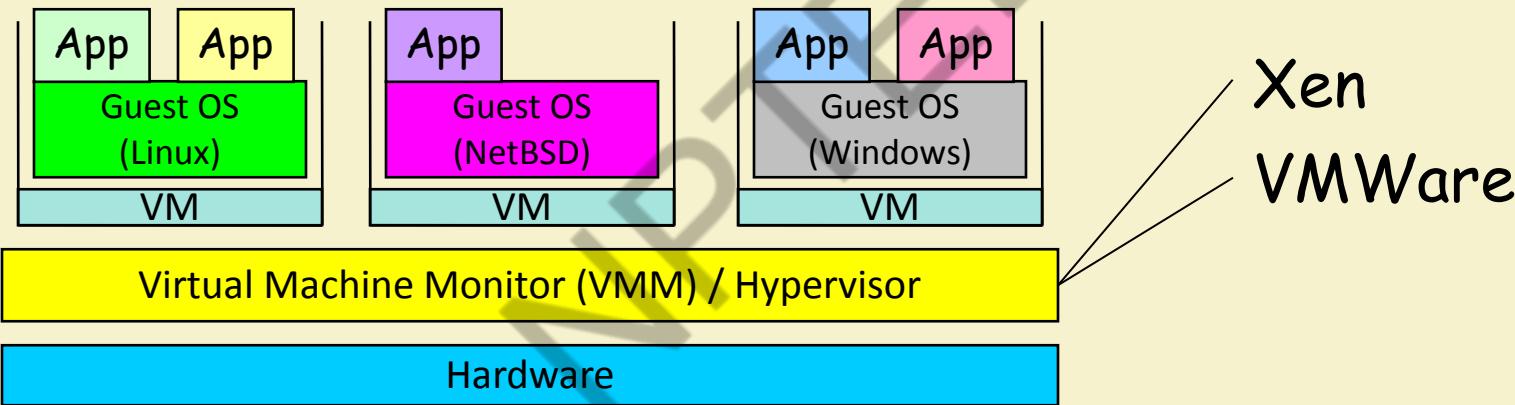
Cloud and Virtualization

- **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



Virtual Machines

- VM technology allows multiple virtual machines to run on a single physical machine.



- Performance: Para-virtualization (e.g. Xen) is very close to raw physical performance!

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



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

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

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
NPTEL ONLINE
CERTIFICATION COURSES

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.

Advantages of Cloud Computing

- **Improved performance:**
 - With few large programs hogging your computer's memory, you will see better performance from your PC.
 - Computers in a cloud computing system boot and run faster because they have fewer programs and processes loaded into memory.
- **Reduced software costs:**
 - Instead of purchasing expensive software applications, you can get most of what you need for free.
 - most cloud computing applications today, such as the Google Docs suite.
 - better than paying for similar commercial software
 - which alone may be justification for switching to cloud applications.

Advantages of Cloud Computing

- **Instant software updates**
 - Another advantage to cloud computing is that you are no longer faced with choosing between obsolete software and high upgrade costs.
 - When the application is web-based, updates happen automatically available the next time you log into the cloud.
 - When you access a web-based application, you get the latest version without needing to pay for or download an upgrade.
- **Improved document format compatibility.**
 - You do not have to worry about the documents you create on your machine being compatible with other users' applications or OS.
 - There are less format incompatibilities when everyone is sharing documents and applications in the cloud.

Advantages of Cloud Computing

- **Unlimited storage capacity**
 - Cloud computing offers virtually limitless storage.
 - Your computer's current 1 Tera Bytes hard drive is small compared to the hundreds of Peta Bytes available in the cloud.
- **Increased data reliability**
 - Unlike desktop computing, in which if a hard disk crashes and destroy all your valuable data, a computer crashing in the cloud should not affect the storage of your data.
 - if your personal computer crashes, all your data is still out there in the cloud, still accessible
 - In a world where few individual desktop PC users back up their data on a regular basis, cloud computing is a data-safe computing platform. For e.g. Dropbox, Skydrive

Advantages of Cloud Computing

- **Universal information access**
 - That is not a problem with cloud computing, because you do not take your documents with you.
 - Instead, they stay in the cloud, and you can access them whenever you have a computer and an Internet connection
 - Documents are instantly available from wherever you are.
- **Latest version availability**
 - When you edit a document at home, that edited version is what you see when you access the document at work.
 - The cloud always hosts the latest version of your documents as long as you are connected, you are not in danger of having an outdated version.

Advantages of Cloud Computing

- **Easier group collaboration**
 - Sharing documents leads directly to better collaboration.
 - Many users do this as it is an important advantages of cloud computing multiple users can collaborate easily on documents and projects
- **Device independence**
 - You are no longer tethered to a single computer or network.
 - Changes to computers, applications and documents follow you through the cloud.
 - Move to a portable device, and your applications and documents are still available.

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.

Disadvantages of Cloud Computing

- **Features might be limited**
 - This situation is bound to change, but today many web-based applications simply are not as full-featured as their desktop-based applications.
 - For example, you can do a lot more with Microsoft PowerPoint than with Google Presentation's web-based offering
- **Can be slow**
 - Even with a fast connection, web-based applications can sometimes be slower than accessing a similar software program on your desktop PC.
 - Everything about the program, from the interface to the current document, has to be sent back and forth from your computer to the computers in the cloud.
 - If the cloud servers happen to be backed up at that moment, or if the Internet is having a slow day, you would not get the instantaneous access you might expect from desktop applications.

Disadvantages of Cloud Computing

- **Stored data might not be secured**
 - With cloud computing, all your data is stored on the cloud.
 - The question is How secure is the cloud?
 - Can unauthorized users gain access to your confidential data ?
- **Stored data can be lost!**
 - Theoretically, data stored in the cloud is safe, replicated across multiple machines.
 - But on the off chance that your data goes missing, you have no physical or local backup.
 - Put simply, relying on the cloud puts you at risk if the cloud lets you down.

Disadvantages of Cloud Computing

- **HPC Systems** **High performance system**
 - Not clear that you can run compute-intensive HPC applications that use MPI/OpenMP!
 - Scheduling is important with this type of application
 - as you want all the VM to be co-located to minimize communication latency!
- **General Concerns**
 - Each cloud systems uses different protocols and different APIs
 - may not be possible to run applications between cloud based systems
 - Amazon has created its own DB system (not SQL 92), and workflow system (many popular workflow systems out there)
 - so your normal applications will have to be adapted to execute on these platforms.

Evolution of Cloud Computing

Business drivers for adopting cloud computing



IIT KHARAGPUR



NPTEL
ONLINE
CERTIFICATION COURSES

Reasons

- The main reason for interest in cloud computing is due to the fact that public clouds can significantly reduce IT costs.
- From an end user perspective cloud computing gives the illusion of potentially infinite capacity with ability to scale rapidly and pay only for the consumed resource.
- In contrast, provisioning for peak capacity is a necessity within private data centers, leading to a low average utilization of 5-20 percent.



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

IaaS Economics

	In house server	Cloud server
Purchase Cost	\$9600 (x86,3QuadCore,12GB RAM, 300GB HD)	0
Cost/hr (over 3 years)	\$0.36	\$0.68
Cost ratio: Cloud/In house	1.88	
Efficiency	40%	80%
Cost/Effective hr	\$0.90	\$0.85
Power and cooling	\$0.36	0
Management Cost	\$0.10	\$0.01
Total cost/effective hr	\$1.36	\$0.86
Cost ratio: In house/Cloud	1.58	

Benefits for the end user while using public cloud

- High utilization
- High scalability
- No separate hardware procurement
- No separate power cost
- No separate IT infrastructure administration/maintenance required
- Public clouds offer user friendly SLA by offering high availability (~99%) and also provide compensation in case of SLA miss.
- Users can rent the cloud to develop and test prototypes before making major investments in technology

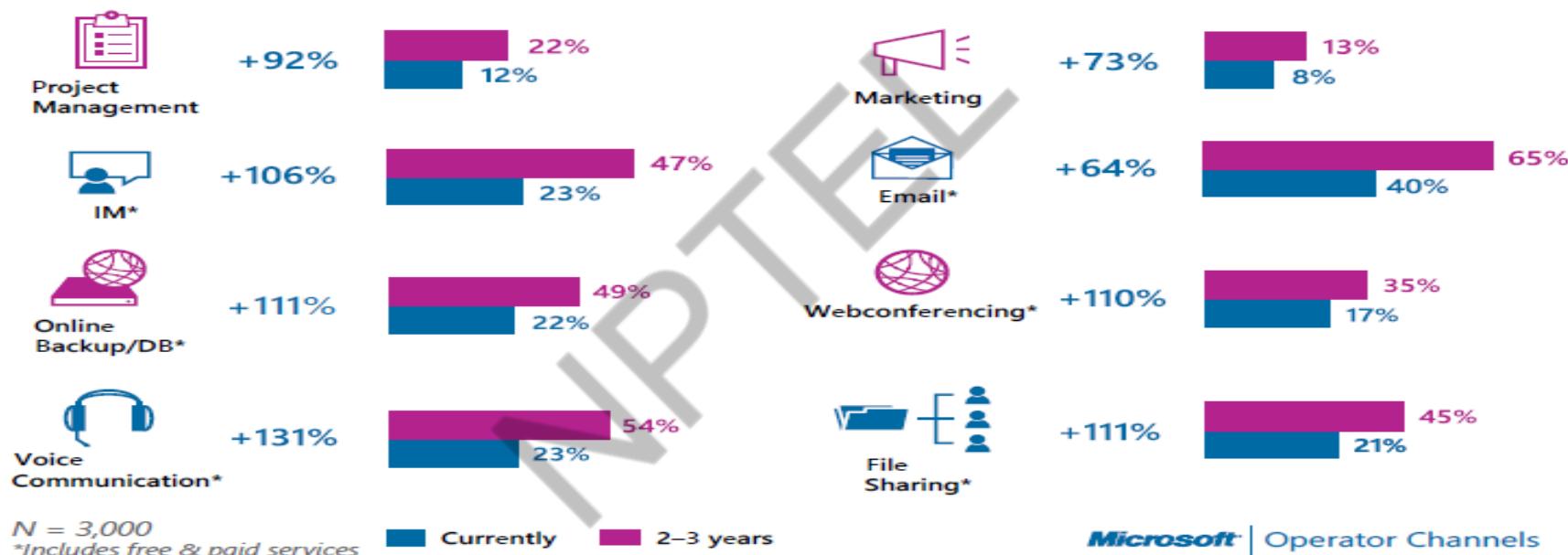
Benefits for the end user while using public cloud

- In order to enhance portability from one public cloud to another, several organizations such as Cloud Computing Interoperability Forum and Open Cloud Consortium are coming up with standards for portability.
- For e.g. Amazon EC2 and Eucalyptus share the same API interface.
- Software startups benefit tremendously by renting computing and storage infrastructure on the cloud instead of buying them as they are uncertain about their own future.

Benefits for Small and Medium Businesses (<250 employees)

SMBs & Cloud Services

Tasks in cloud services currently and in 2-3 years



Source: <http://www.microsoft.com/en-us/news/presskits/telecom/docs/SMBCloud.pdf>

Benefits of private cloud

- Cost of 1 server with 12 cores and 12 GB RAM is far lower than the cost of 12 servers having 1 core and 1 GB RAM.
- Confidentiality of data is preserved
- Virtual machines are cheaper than actual machines
- Virtual machines are faster to provision than actual machines

Economics of PaaS vs IaaS

- Consider a web application that needs to be available 24X7, but where the transaction volume is unpredictable and can vary rapidly
- Using an IaaS cloud, a minimal number of servers would need to be provisioned at all times to ensure availability
- In contrast, merely deploying the application on PaaS cloud costs nothing. Depending upon the usage, costs are incurred.
- The PaaS cloud scales automatically to successfully handle increased requests to the web application.

Source: Enterprise Cloud Computing by Gautam Shroff

PaaS benefits

- No need for the user to handle scaling and load balancing of requests among virtual machines
- PaaS clouds also provide web based Integrated Development Environment for development and deployment of application on the PaaS cloud.
- Easier to migrate code from development environment to the actual production environment.
- Hence developers can directly write applications on the cloud and don't have to buy separate licenses of IDE.

SaaS benefits

- Users subscribe to web services and web applications instead of buying and licensing software instances.
- For e.g. Google Docs can be used for free, instead of buying document reading softwares such as Microsoft Word.
- Enterprises can use web based SaaS Content Relationship Management applications, instead of buying servers and installing CRM softwares and associated databases on them.

Customer relationship management



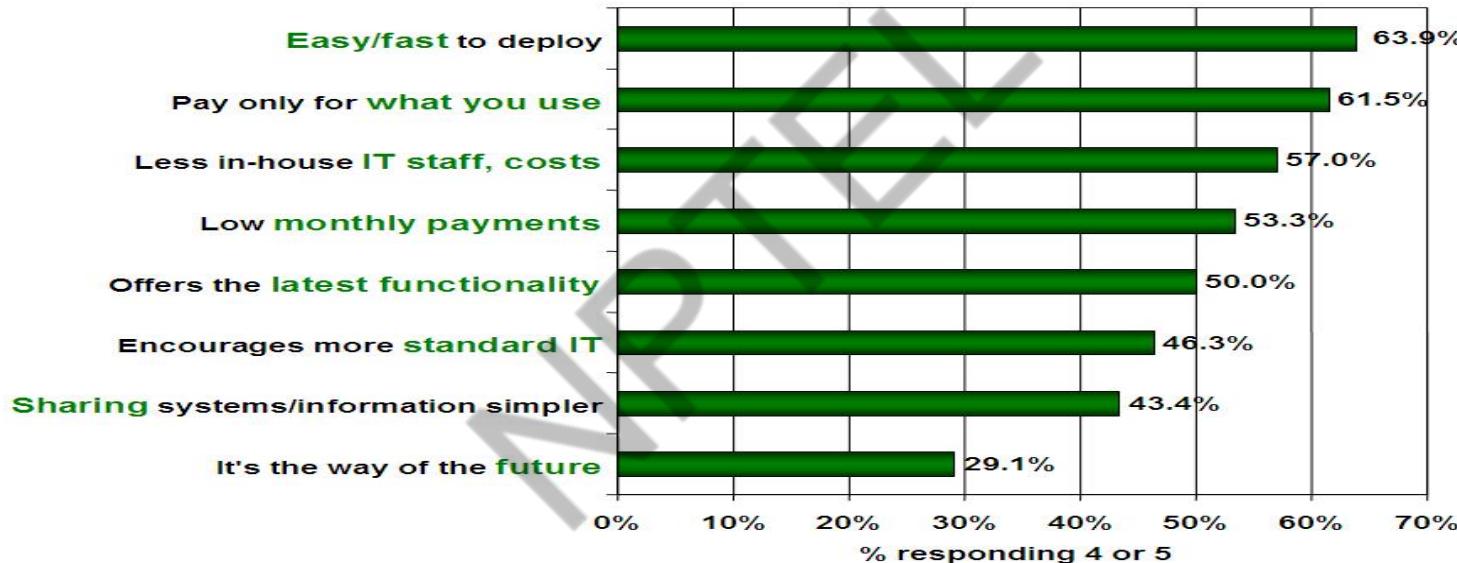
IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Benefits, as perceived by the IT industry

Q: Rate the benefits commonly ascribed to the 'cloud'/on-demand model
(1=not important, 5=very important)



Source: IDC Enterprise Panel, August 2008 n=244

Factors driving investment in cloud

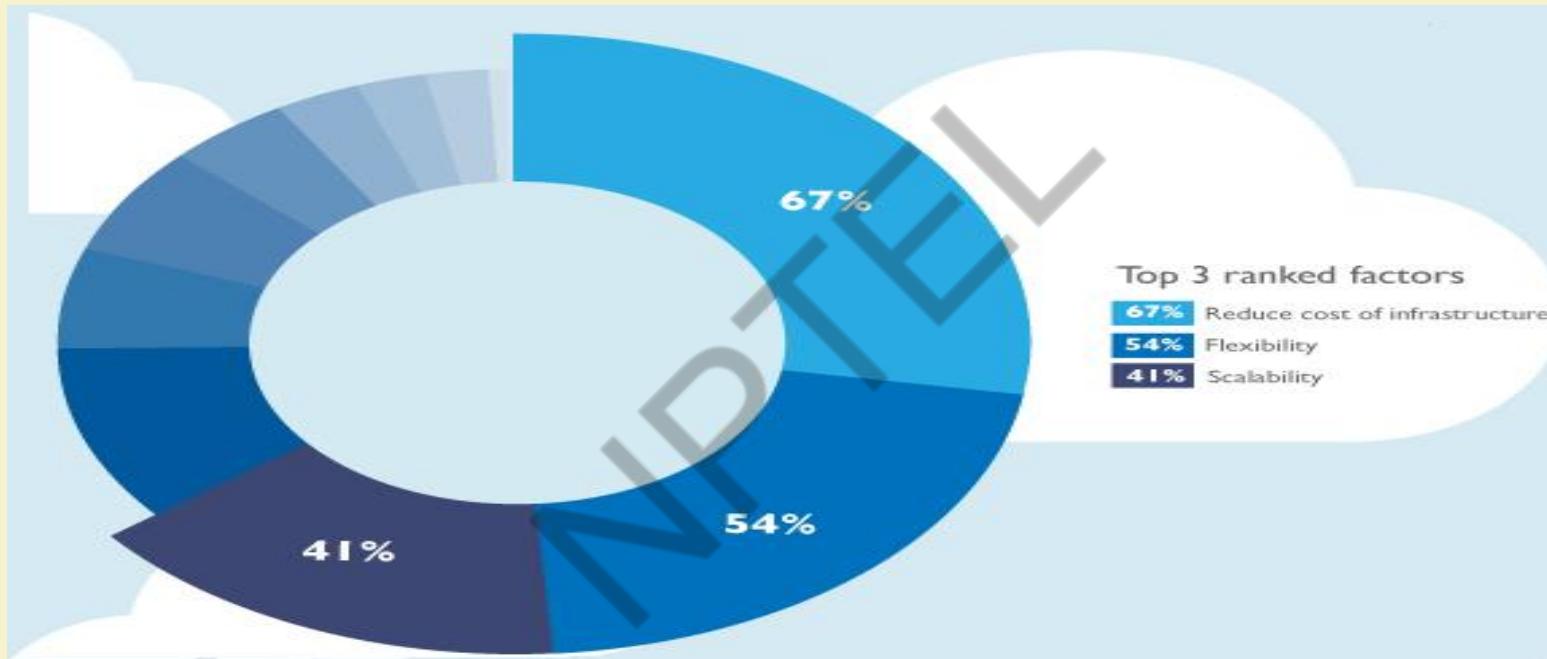
Factors driving investments in cloud per business size

- Large companies
- Medium companies
- Small companies



Source: <http://www.cloudtweaks.com/2012/01/infographic-whats-driving-investment-in-cloud-computing/>

Factors driving investment in cloud



Source: <http://www.cloudtweaks.com/2012/01/infographic-whats-driving-investment-in-cloud-computing/>

Purpose of cloud computing in organizations

- Providing an IT platform for business processes involving multiple organizations
- Backing up data **Enterprise resource planning**
- Running CRM, ERP, or supply chain management applications
- Providing personal productivity and collaboration tools to employees
- Developing and testing software
- Storing and archiving large files (e.g., video or audio)
- Analyzing customer or operations data
- Running e-business or e-government web sites

Source: <http://askvisory.com/research/key-drivers-of-cloud-computing-activity/>

Purpose of cloud computing in organizations

- Analyzing data for research and development Put an end
- Meeting spikes in demand on our web site or internal systems
- Processing and storing applications or other forms
- Running data-intensive batch applications (e.g., data conversion, risk modeling, graphics rendering)
- Sharing information with the government or regulators
- Providing consumer entertainment, information and communication (e.g., music, video, photos, social networks)

Source: <http://askvisory.com/research/key-drivers-of-cloud-computing-activity/>

Top cloud applications that are driving cloud adaptation

- Mail and Messaging
- Archiving
- Backup
- Storage
- Security
- Virtual Servers
- CRM (Customer Relationship Management)
- Collaboration across enterprises
- Hosted PBX (Private Branch Exchange)
- Video Conferencing

Source: <http://www.itnewsafrica.com/2012/09/ten-drivers-of-cloud-computing-for-south-african-businesses/>

Thank You!

NPTEL



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

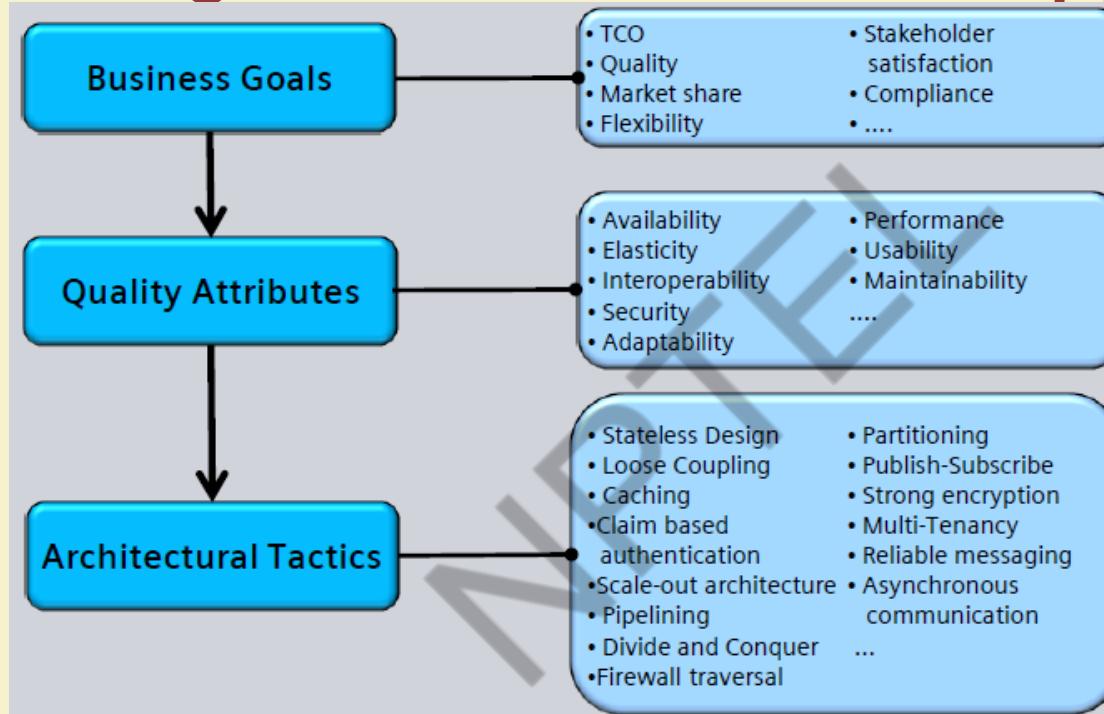
CLOUD COMPUTING

CLOUD COMPUTING ARCHITECTURE

PROF. SOUMYA K. GHOSH

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
IIT KHARAGPUR

Context: High Level Architectural Approach



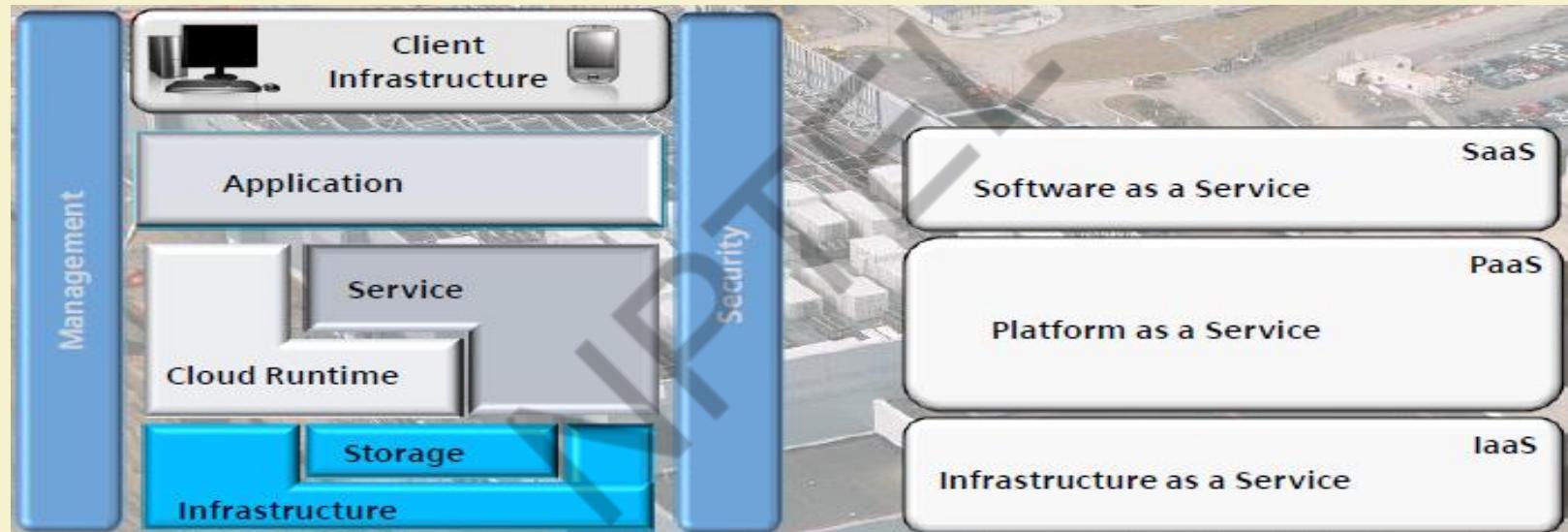
Source: <http://www.sei.cmu.edu/library/assets/presentations/Cloud%20Computing%20Architecture%20-%20Gerald%20Kaefer.pdf>

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

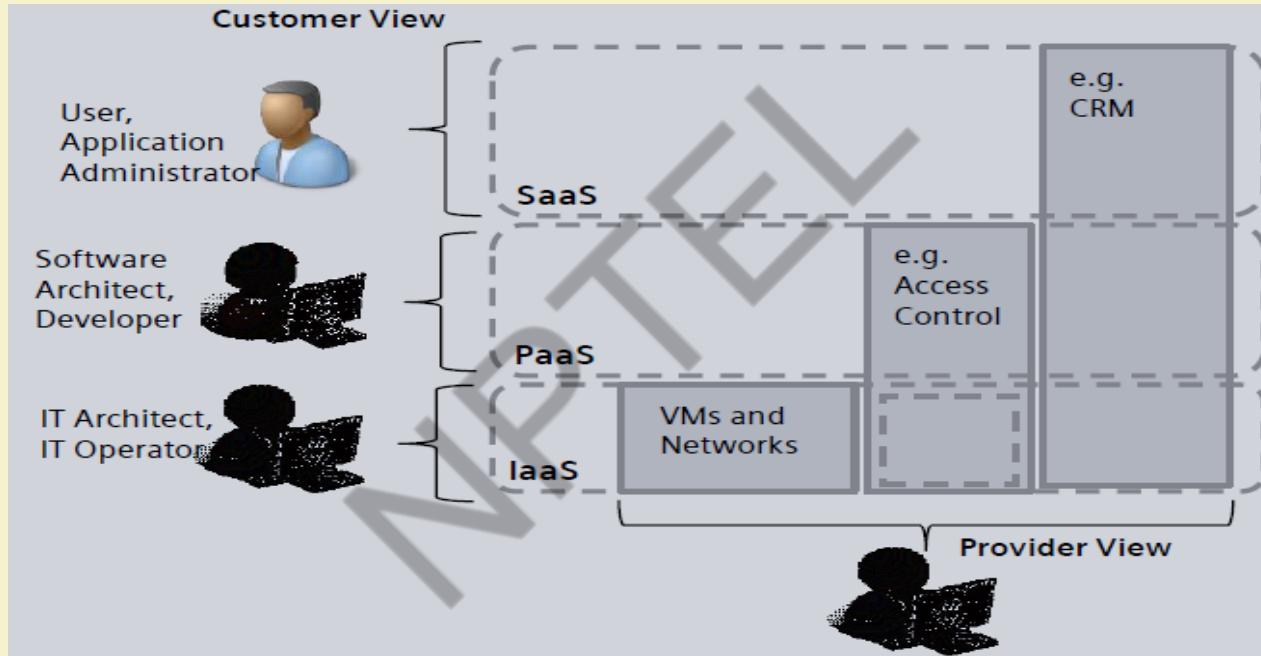
Ref: <http://www.sei.cmu.edu/library/assets/presentations/Cloud%20Computing%20Architecture%20-%20Gerald%20Kaefer.pdf>

Cloud Computing Architecture - XaaS



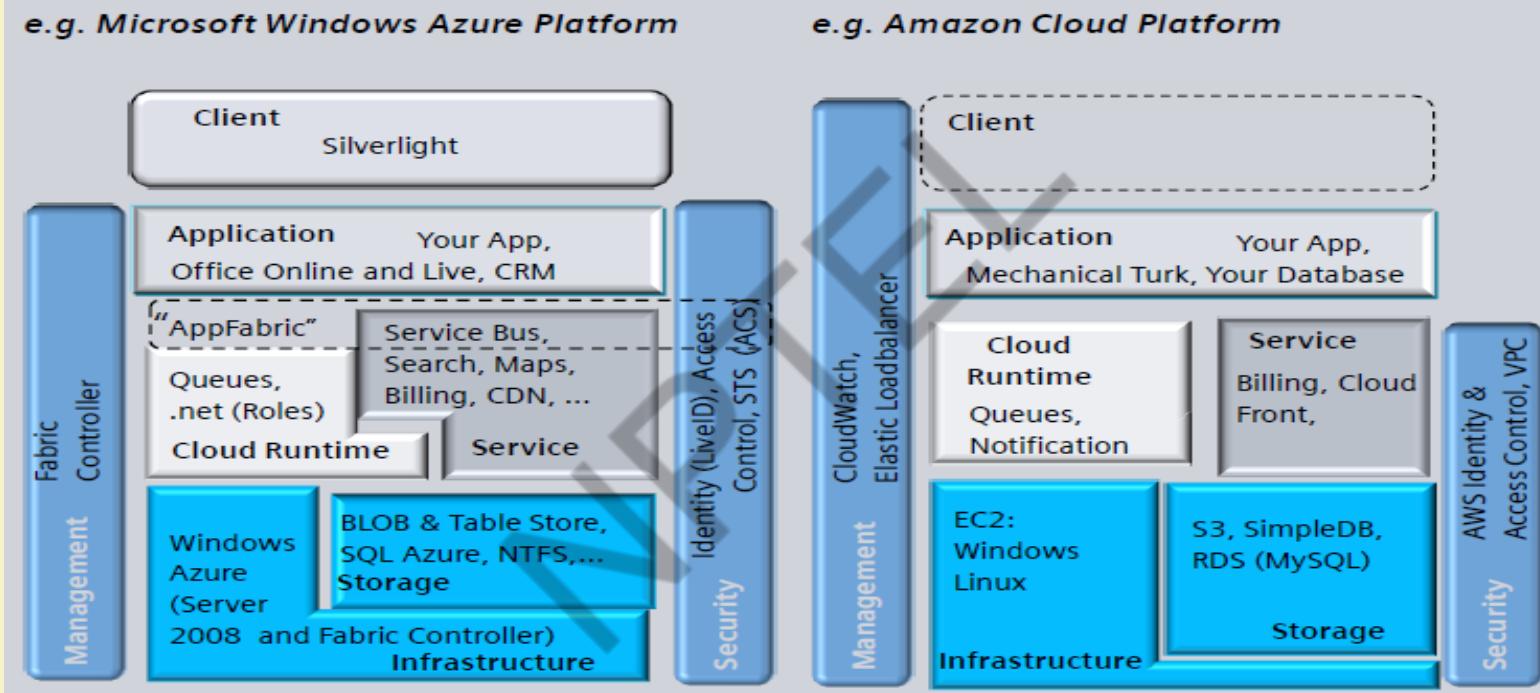
Source: <http://www.sei.cmu.edu/library/assets/presentations/Cloud%20Computing%20Architecture%20-%20Gerald%20Kaefer.pdf>

XaaS Stack views: Customer view vs Provider view



Source: <http://www.sei.cmu.edu/library/assets/presentations/Cloud%20Computing%20Architecture%20-%20Gerald%20Kaefer.pdf>

Microsoft Azure vs Amazon EC2



Source: <http://www.sei.cmu.edu/library/assets/presentations/Cloud%20Computing%20Architecture%20-%20Gerald%20Kaefer.pdf>

Architecture for elasticity

Vertical Scale Up

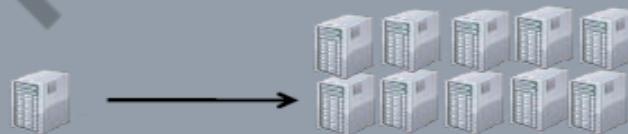
- Add more resources to a single computation unit i.e. Buy a bigger box
- Move a workload to a computation unit with more resources



For small scenarios scale up is probably cheaper - code "just works"

Horizontal Scale Out

- Adding additional computation units and having them act in concert
- Splitting workload across multiple computation units
- Database partitioning



For larger scenarios scale out is the only solution
1x64 Way Server much more expensive than
64x1 Way Servers

Source: <http://www.sei.cmu.edu/library/assets/presentations/Cloud%20Computing%20Architecture%20-%20Gerald%20Kaefer.pdf>



IIT KHARAGPUR

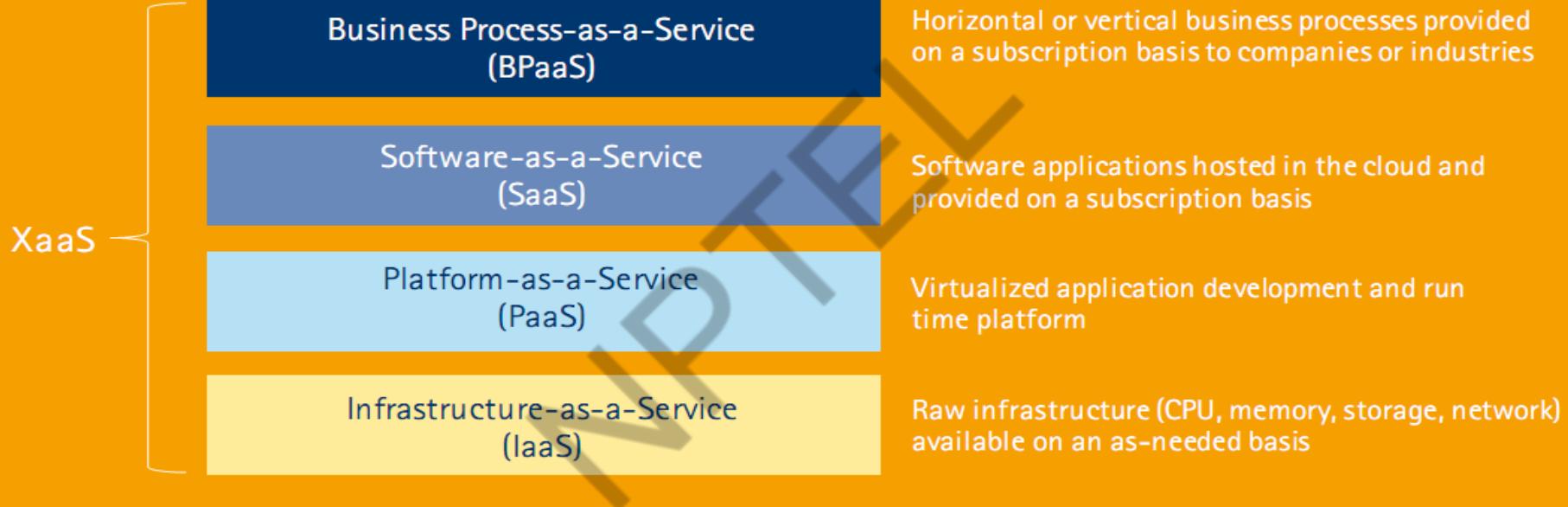


NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

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.

Service Models (XaaS)



Service Models (XaaS)



Source: Cloud Security and Privacy: An Enterprise Perspective on Risks and Compliance by Tim Mather and Subra Kumaraswamy

Service Models (XaaS)

- **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)

Requirements of CSP (Cloud Service Provider)

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

Service Models (XaaS)

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

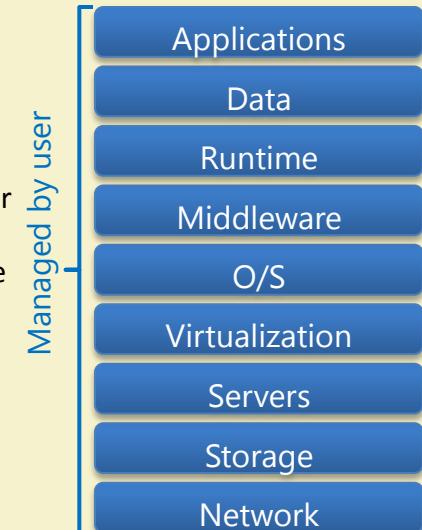


Source: Understanding the Cloud Computing Stack: PaaS, SaaS, IaaS © Diversity Limited, 2011

Classical Service Model

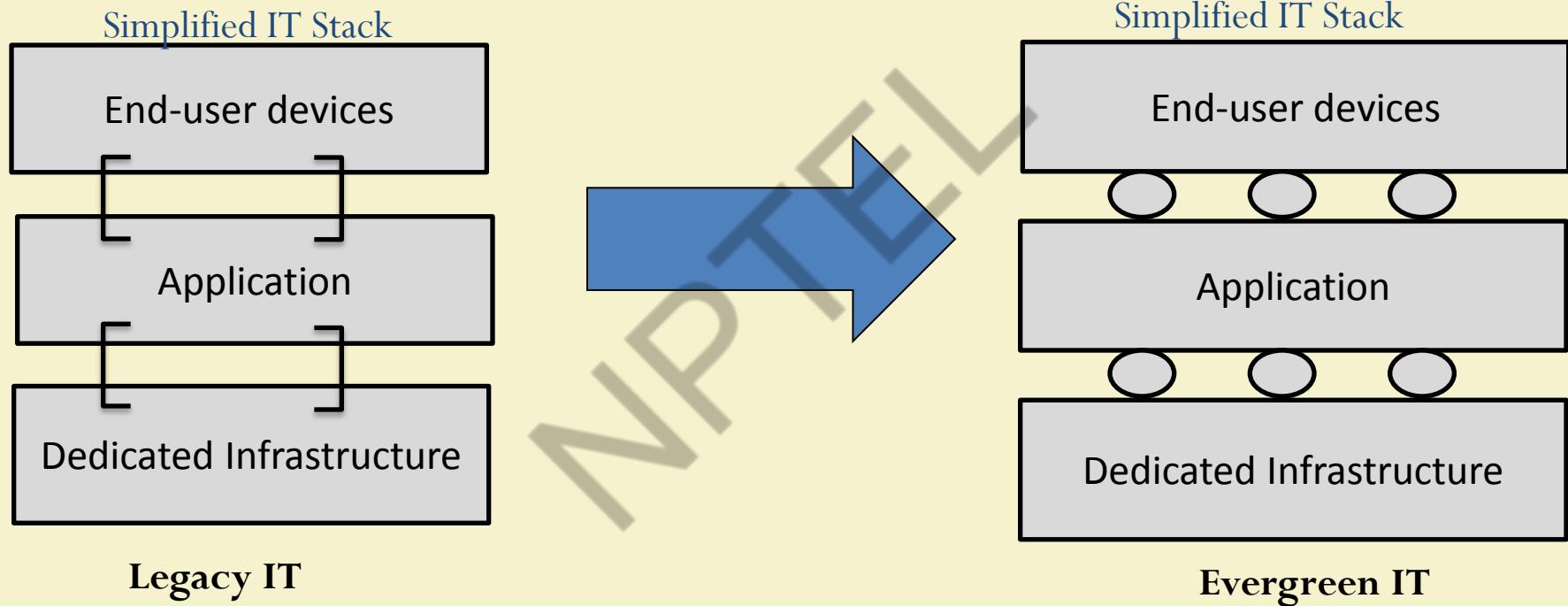
- All the Layers(H/W, Operating System, Development Tools, Applications) Managed by the Users
- Initial IT budget and resources.
- Users bears the costs of the hardware, maintenance and technology.
- Each system is designed and funded for a specific business activity: custom build-to-order
- Systems are deployed as a vertical stack of “layers” which are tightly coupled, so no single part can be easily replaced or changed
- Prevalent of manual operations for provisioning, management
- Result: Legacy IT

ADR MOV SSN



Source: Dragan , “XaaS as a Modern Infrastructure for eGoverment Business Model in the Republic of Croatia”

Key impact of cloud computing for IT function: From Legacy IT to Evergreen IT



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

Client Server Architecture



Source: Wikipedia



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

Thank You!

NPTEL



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

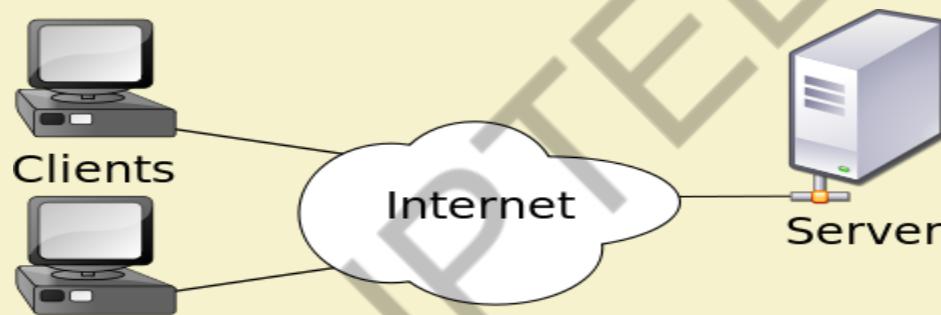
CLOUD COMPUTING

CLOUD COMPUTING ARCHITECTURE

PROF. SOUMYA K. GHOSH

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
IIT KHARAGPUR

Client Server Architecture



Source: Wikipedia



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

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
NPTEL ONLINE
CERTIFICATION COURSES

Three Tier Client-Server Architecture

Presentation tier

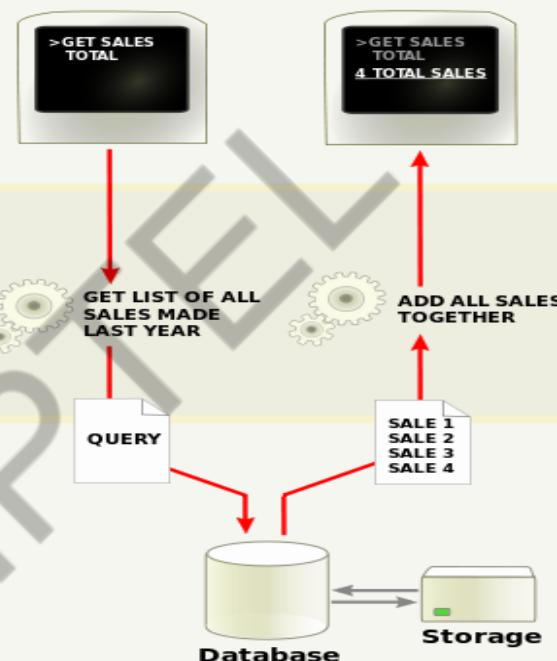
The top-most level of the application is the user interface. The main function of the interface is to translate tasks and results to something the user can understand.

Logic tier

This layer coordinates the application, processes commands, makes logical decisions and evaluations, and performs calculations. It also moves and processes data between the two surrounding layers.

Data tier

Here information is stored and retrieved from a database or file system. The information is then passed back to the logic tier for processing, and then eventually back to the user.



Source: Wikipedia



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Client Server model vs. Cloud model

Client server model

- Simple service model where server services client requests
- May/may not be load balanced
- Scalable to some extent in a cluster environment.
- No concept of virtualization

Cloud computing model

- Variety of complex service models, such as, IaaS, PaaS, SaaS can be provided
- Load balanced
- Theoretically infinitely scalable
- Virtualization is the core concept

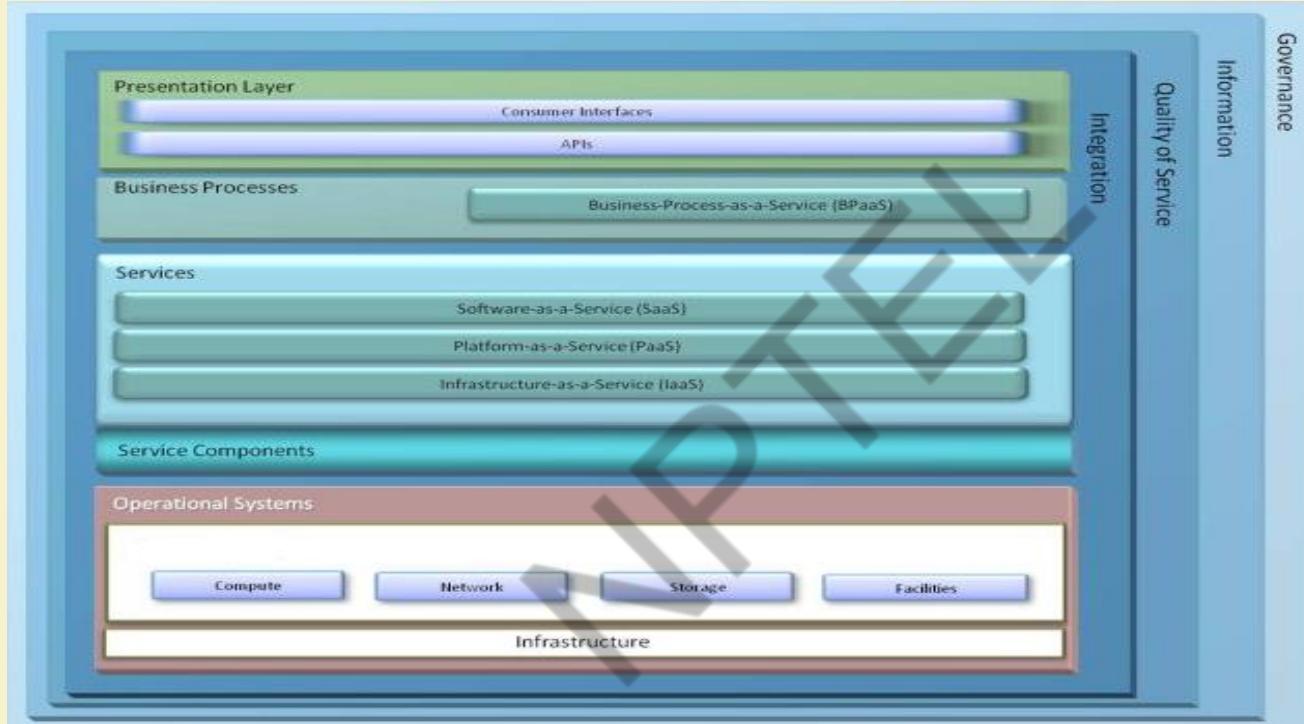


IIT KHARAGPUR



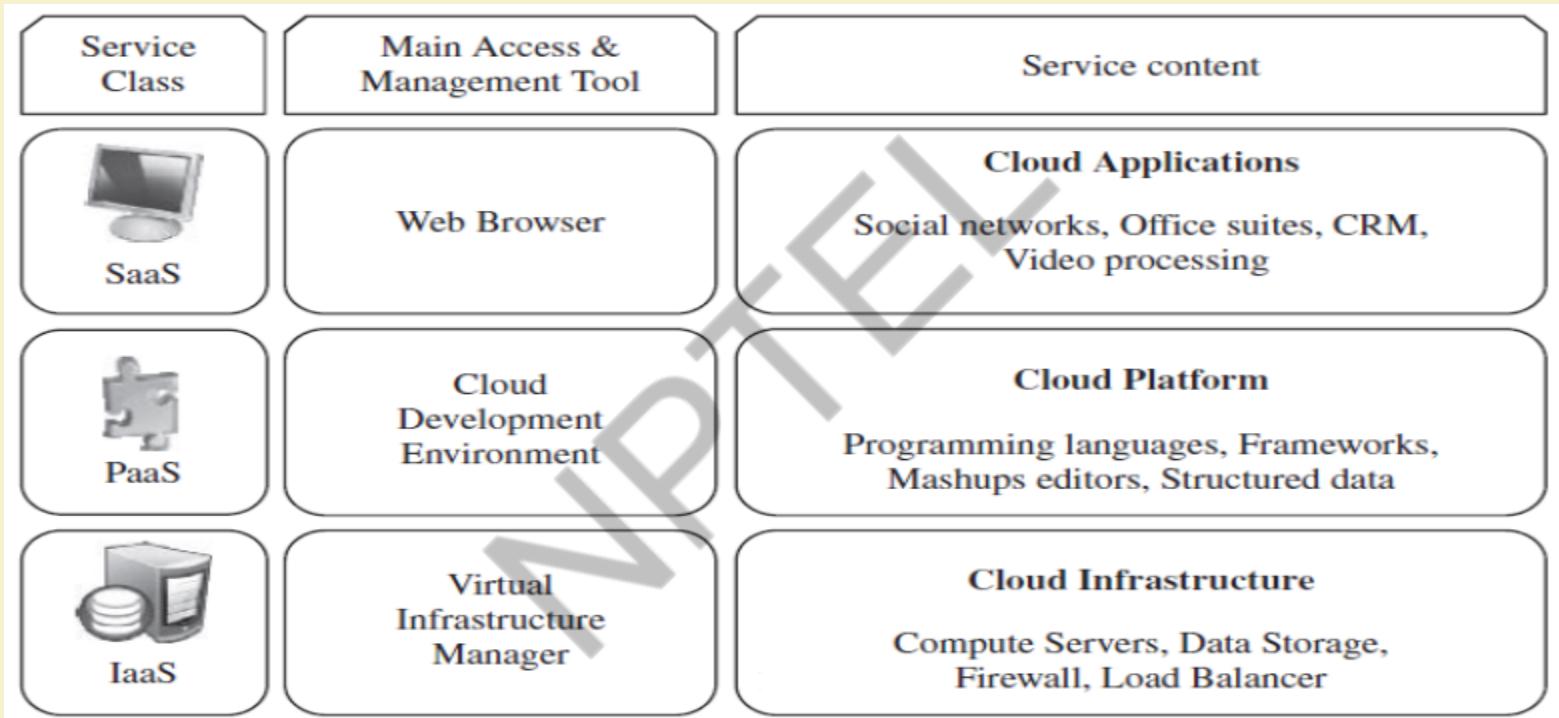
NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Cloud Services



Source : <http://www.opengroup.org/soa/source-book/socci/extend.htm#figure2>

Cloud service models



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



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

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

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

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.

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

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
NPTEL ONLINE
CERTIFICATION COURSES

Applications where SaaS is used

- 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 need.
- Software where demand spikes significantly. E.g. Tax/Billing softwares. **Put an end**
- E.g. of SaaS: Sales Force Customer Relationship Management (CRM) software

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

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

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.

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

Characteristics of PaaS

- Services to develop, test, deploy, host and maintain applications in the same integrated development environment. All the varying services needed to fulfill the application development process.
- 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

Scenarios where PaaS is used

- 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.
- The 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.rackspace.com/hosting_knowledge/whitepapers/Understanding-the-Cloud-Computing-Stack.pdf

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

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

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

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

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

Scenarios where IaaS may not be the best option

- 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

SaaS providers

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/epsavola/seminari/Cloud%20Service%20Models.pdf>

Feature comparison of PaaS providers

Provider	Target to Use	Programming language, Frameworks	Programming Models	Persistence options
Aneka	.NET enterprise applications, Web applications	.NET	Threads, Task, MapReduce	Flat files, RDBMS
AppEngine	Web applications	Python, Java	Request-based Web programming	BigTable
Force.com	Enterprise applications	Apex	Workflow, Request-based Web programming, Excel-like formula language	Own object database
Azure	Enterprise applications, Web applications	.NET	Unrestricted	Table/BLOB/queue storage, SQL Services
Heroku	Web applications	Ruby on Rails	Request-based Web programming	PostgreSQL, Amazon RDS
Amazon Elastic MapReduce	Data processing	Hive and Pig, Cascading, Java, Ruby, Perl, Python, PHP, C++	MapReduce	Amazon S3

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

Feature comparison of IaaS providers

Provider	Geographic distribution of data centers	User interfaces and APIs	Hardware capacity	Guest operating systems	Smallest billing unit
Amazon E2C	US Europe	CLI, WS, Portal	CPU: 1-20 EC2 compute units Memory: 1.7-15 GB Storage: 160-1690 GB, 1 GB – 1 TB (per ESB units)	Linux Windows	Hour
Flexiscale	UK	Web console	CPU: 1-4 Memory: 0.5-16 GB Storage: 20-270 GB	Linux, Windows	Hour
GoGrid		REST, Java, PHP, Python, Ruby	CPU: 1-6 Memory: 0.5-8 GB Storage: 30-480 GB	Linux, Windows	Hour
Joyent	US		CPU: 1/16-8 Memory: 0.25-32.5 GB Storage: 5-100GB	OpenSolaris	Month
RackSpace	US	Portal, REST, Python, PHP, Java, .NET	CPU: Quad-core Memory: 0.25-16 GB Storage: 10-620 GB	Linux	Hour

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

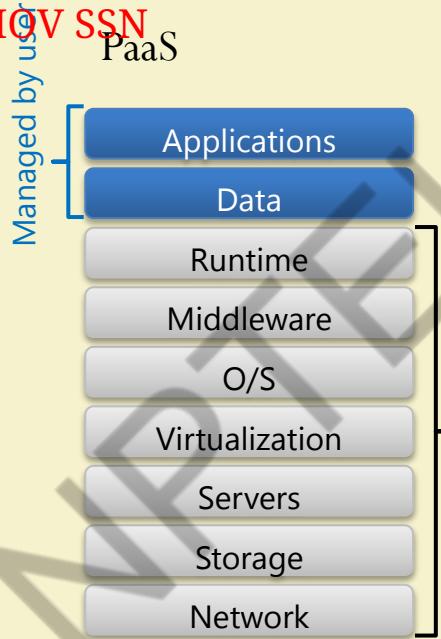
XaaS

SaaS

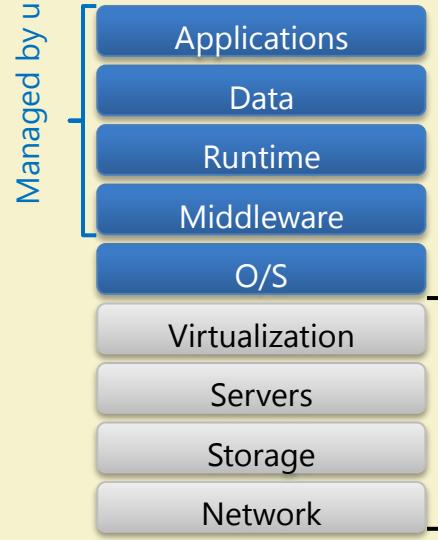


ADR MOV SSN

PaaS



IaaS



Managed by service provider



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

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>

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>

Network Function Virtualization

Definition: “Network Functions Virtualisation aims to transform the way that network operators architect networks by evolving standard IT virtualisation 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, as illustrated in Figure 1. 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.”

Source: https://portal.etsi.org/nfv/nfv_white_paper.pdf

Network Function Virtualization

Classical Network Appliance Approach



- Fragmented non-commodity hardware.
- Physical install per appliance per site.
- Hardware development large barrier to entry for new vendors, constraining innovation & competition.



Source: https://portal.etsi.org/nfv/nfv_white_paper.pdf



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

Thank You !!

NPTEL



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

CLOUD COMPUTING

ARCHITECTURE - Deployment Models

PROF. SOUMYA K. GHOSH

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
IIT KHARAGPUR

Deployment Models

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



IIT KHARAGPUR

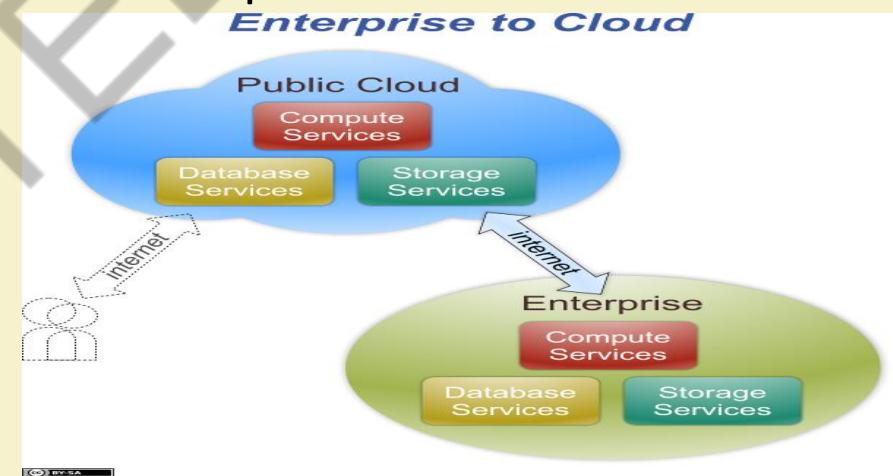


NPTEL
ONLINE
CERTIFICATION COURSES

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.

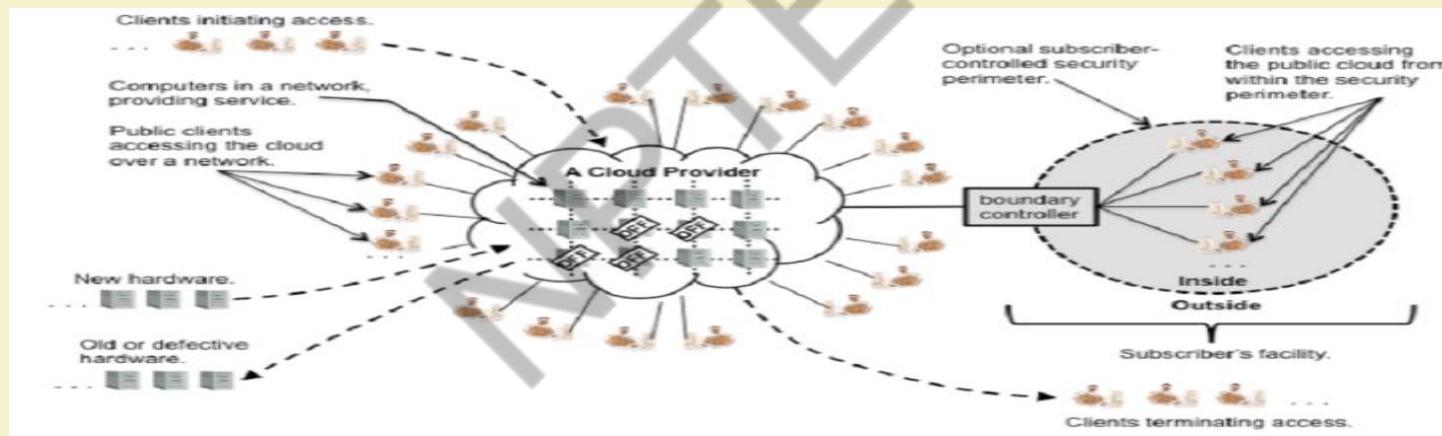
- Examples of Public Cloud:
- Google App Engine
- Microsoft Windows Azure
- IBM Smart Cloud
- Amazon EC2



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

Public Cloud

- In Public setting, the provider's computing and storage resources are potentially large; the communication links can be assumed to be implemented over the public Internet; and the cloud serves a diverse pool of clients (and possibly attackers).



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

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
NPTEL ONLINE
CERTIFICATION COURSES

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

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

Public Cloud

- 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



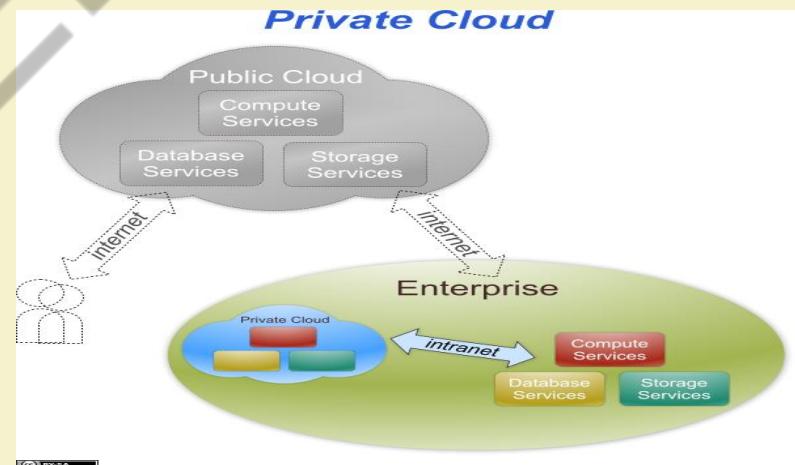
IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

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.



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.



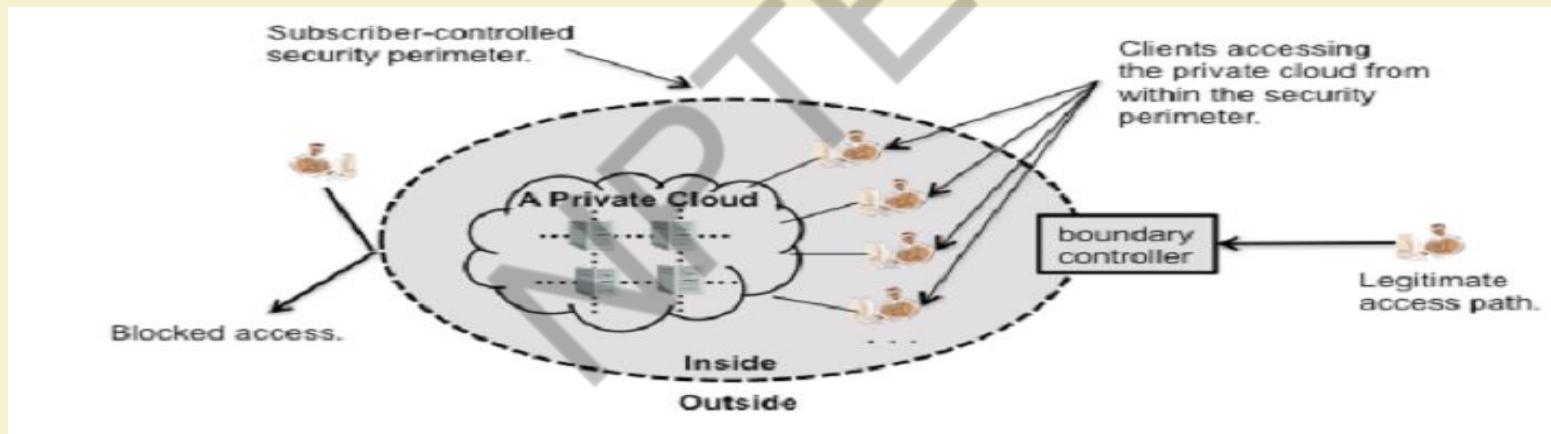
IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

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

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

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.

On-site Private Cloud

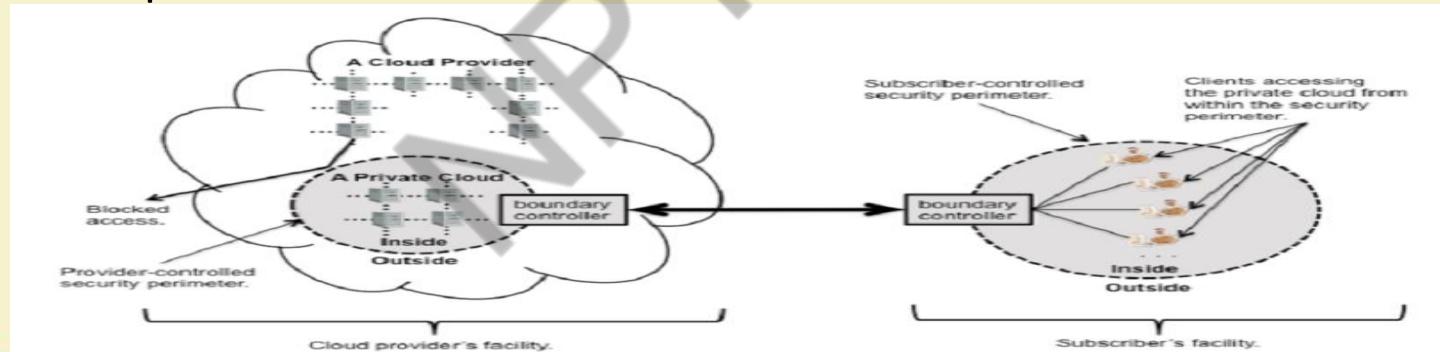
- **Potentially strong security from external threats (on-site-private):**
 - In an on-site private cloud, a subscriber has the option of implementing an appropriately strong security perimeter to protect private cloud resources against external threats to the same level of security as can be achieved 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.

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.

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
NPTEL ONLINE
CERTIFICATION COURSES

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.

Outsourced Private Cloud

- **Data import/export, and performance limitations (outsourced-private):**
 - On-demand bulk data import/export is limited by the network capacity between a provider and subscriber, and real-time or critical processing may be problematic because of networking limitations. In the outsourced private cloud scenario, however, these limits may be adjusted, although not eliminated, by provisioning high-performance and/or high-reliability networking between the provider and subscriber.
- **Potentially strong security from external threats (outsourced-private):**
 - As with the on-site private cloud scenario, a variety of techniques exist to harden a security perimeter. The main difference with the outsourced private cloud is that the techniques need to be applied both to a subscriber's perimeter and provider's perimeter, and that the communications link needs to be protected.

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
NPTEL ONLINE
CERTIFICATION COURSES

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.



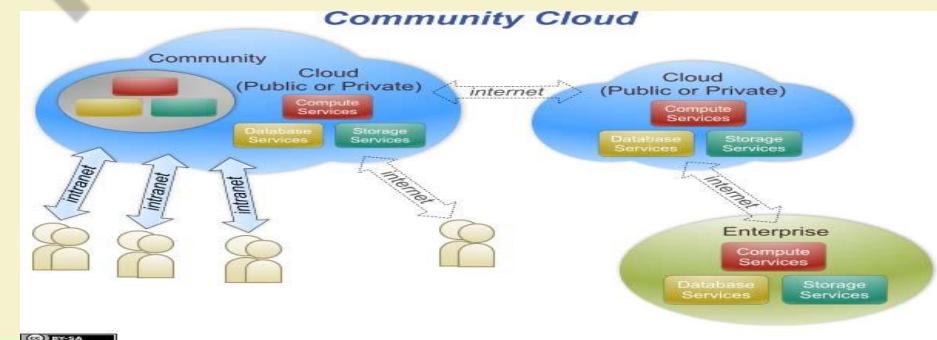
IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

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



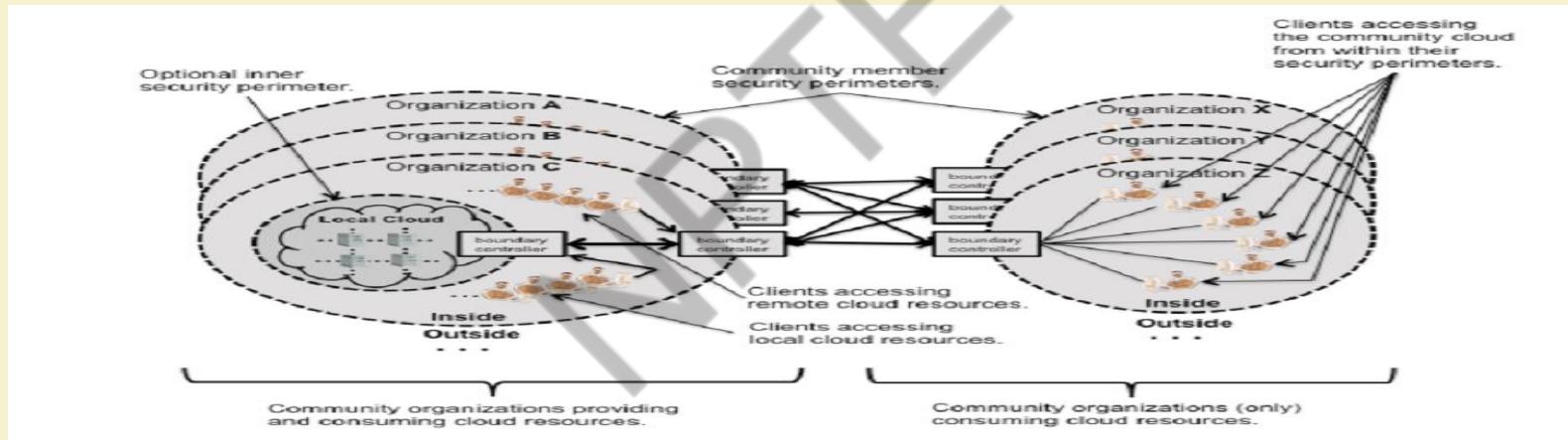
IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

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 "

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.

On-site Community Cloud

- Organizations considering the use of an on-site community cloud should consider:
 - **Network Dependency (on-site community):**
 - The subscribers in an on-site community cloud need to either provision controlled inter-site communication links or use cryptography over a less controlled communications media (such as the public Internet).
 - The reliability and security of the community cloud depends on the reliability and security of the communication links.

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.

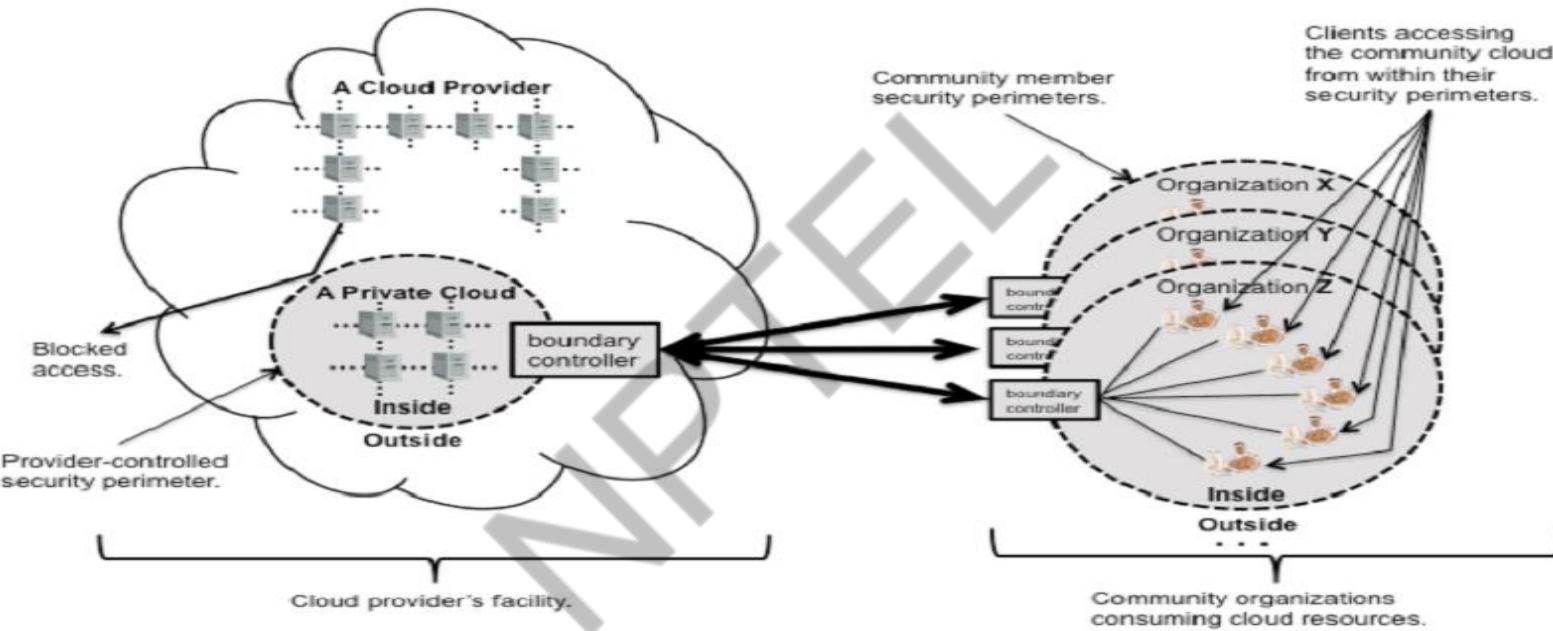
On-site Community Cloud

- **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.

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).

Outsourced Community Cloud



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

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

Outsourced Community Cloud

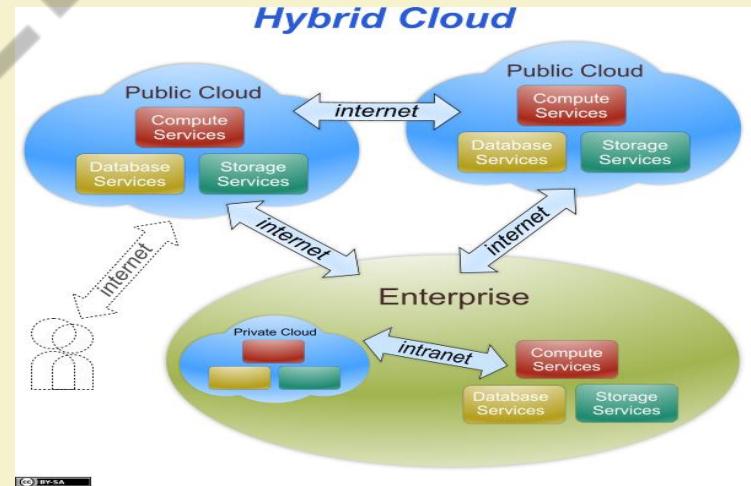
- **Risks from multi-tenancy (outsourced-community):**
 - Same as the on-site community cloud
- **Data import/export, and performance limitations (outsourced-community):**
 - Same as outsourced private cloud
- **Potentially strong security from external threats (outsourced-community):**
 - Same as the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

Outsourced Community Cloud

- Extensive resources available (outsourced-community).
 - Same as outsourced private cloud

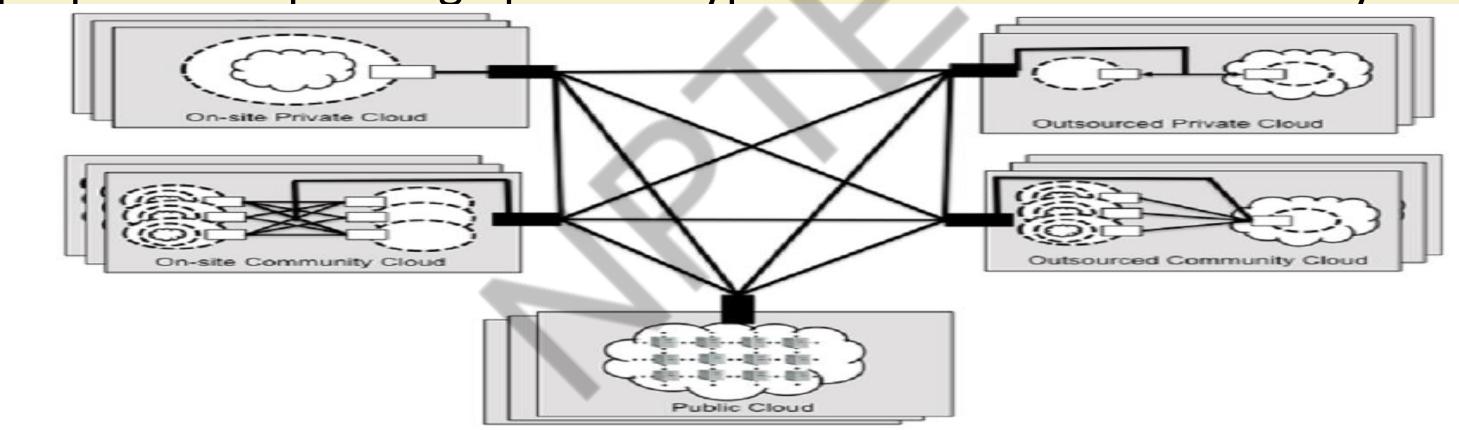
Hybrid Cloud

- 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)



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 "

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
NPTEL ONLINE
CERTIFICATION COURSES

Thank You!

NPTEL



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES



IIT KHARAGPUR



NPTEL

NPTEL ONLINE
CERTIFICATION COURSES

CLOUD COMPUTING

Virtualization

PROF. SOUMYA K. GHOSH

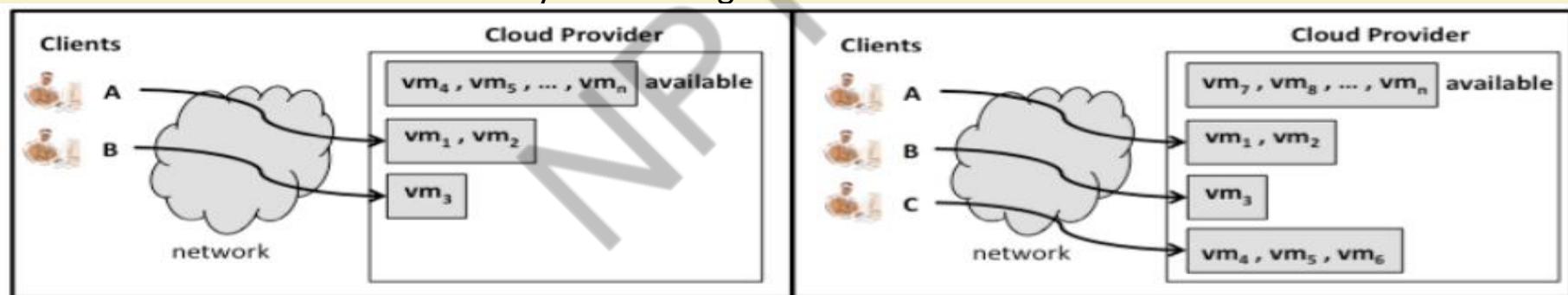
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
IIT KHARAGPUR

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)

IaaS Provider/Subscriber Interaction Dynamics

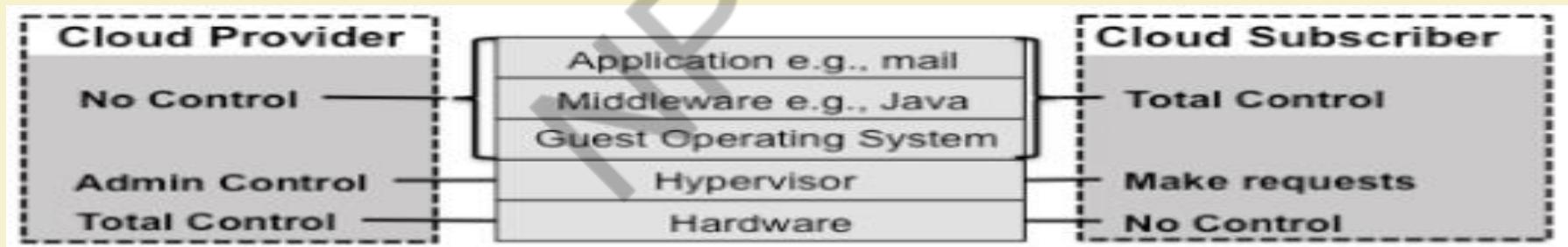
- The provider has a number of available virtual machines (vm's) that it can allocate to clients.
 - Client A has access to vm1 and vm2, Client B has access to vm3 and Client C has access to vm4, vm5 and vm6
 - Provider retains only vm7 through vmN



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

IaaS Component Stack and Scope of Control

- IaaS component stack comprises of hardware, operating system, middleware, and applications layers.
- Operating system layer is split into two layers.
 - Lower (and more privileged) layer is occupied by the Virtual Machine Monitor (VMM), which is also called the Hypervisor
 - Higher layer is occupied by an operating system running within a VM called a guest operating system



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

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

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.



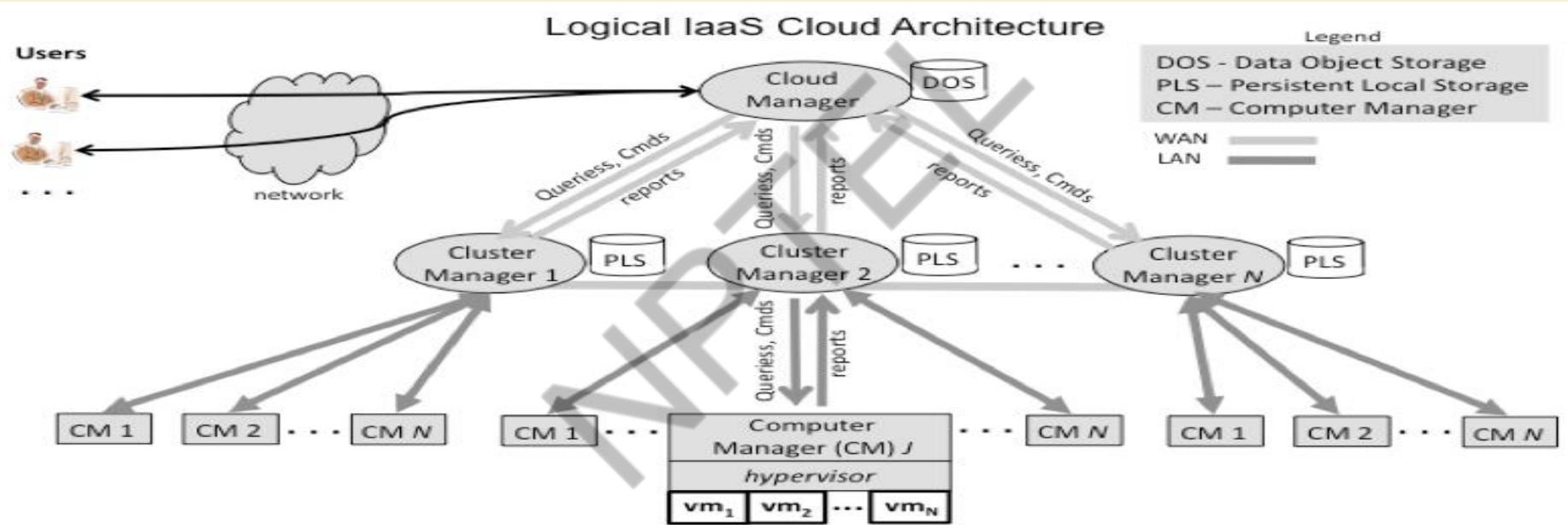
IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

IaaS Cloud Architecture

- Logical view of IaaS cloud structure and operation



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

IaaS Cloud Architecture

- Three-level hierarchy of components in IaaS cloud systems
 - *Top level* is responsible for *central control*
 - *Middle level* is responsible for *management of possibly large computer clusters* that may be *geographically distant* from one another
 - *Bottom level* is responsible for *running the host computer systems* on which virtual machines are created.
- Subscriber queries and commands generally flow into the system at the top and are forwarded down through the layers that either answer the queries or execute the commands



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

IaaS Cloud Architecture

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

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



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Data Object Storage (DOS)

- DOS generally stores the subscriber's metadata like user credentials, operating system images.
- DOS service is (usually) single for a cloud.

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
NPTEL ONLINE
CERTIFICATION COURSES

Operation of the Cluster Managers

- Directed by the Cloud Manager, a Cluster Manager then instructs the Computer Managers to perform resource allocation, and reconfigures the virtual network infrastructure to give the subscriber uniform access.
- Each Cluster Manager is connected to Persistent Local Storage (PLS)
- PLS provide persistent disk-like storage to Virtual Machine



IIT KHARAGPUR



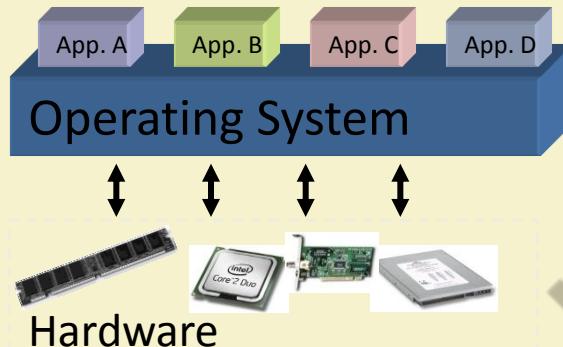
NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

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

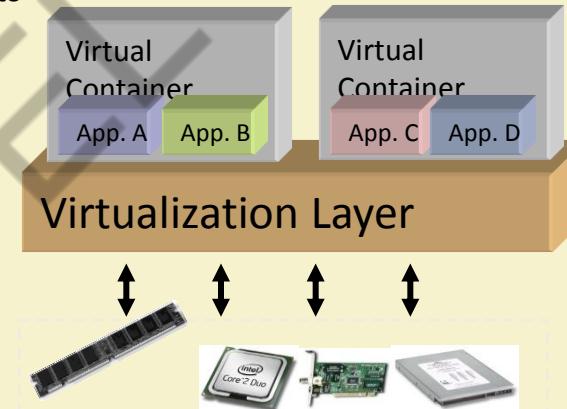
Virtualization

- Virtualization is a broad term (virtual memory, storage, network, etc)
- Focus: **Platform virtualization**
- Virtualization basically allows one computer to do the job of multiple computers, by sharing the resources of a single hardware across multiple environments



'Non-virtualized' system

A single OS controls all hardware platform resources

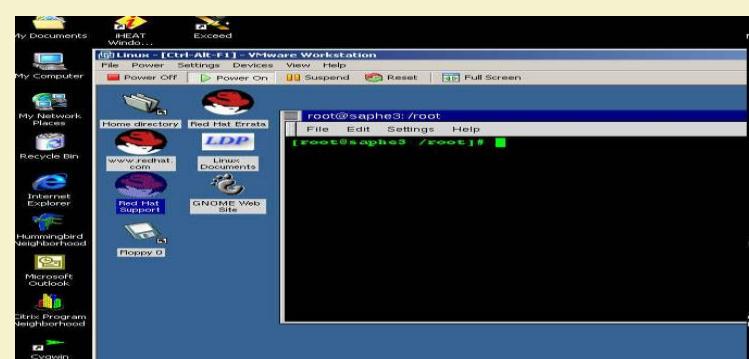


Hardware Virtualized system
It makes it possible to run multiple Virtual Containers on a single physical platform

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

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



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.uba.ar/events/eci/2008/courses/n2/Virtualization-Introduction.ppt

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

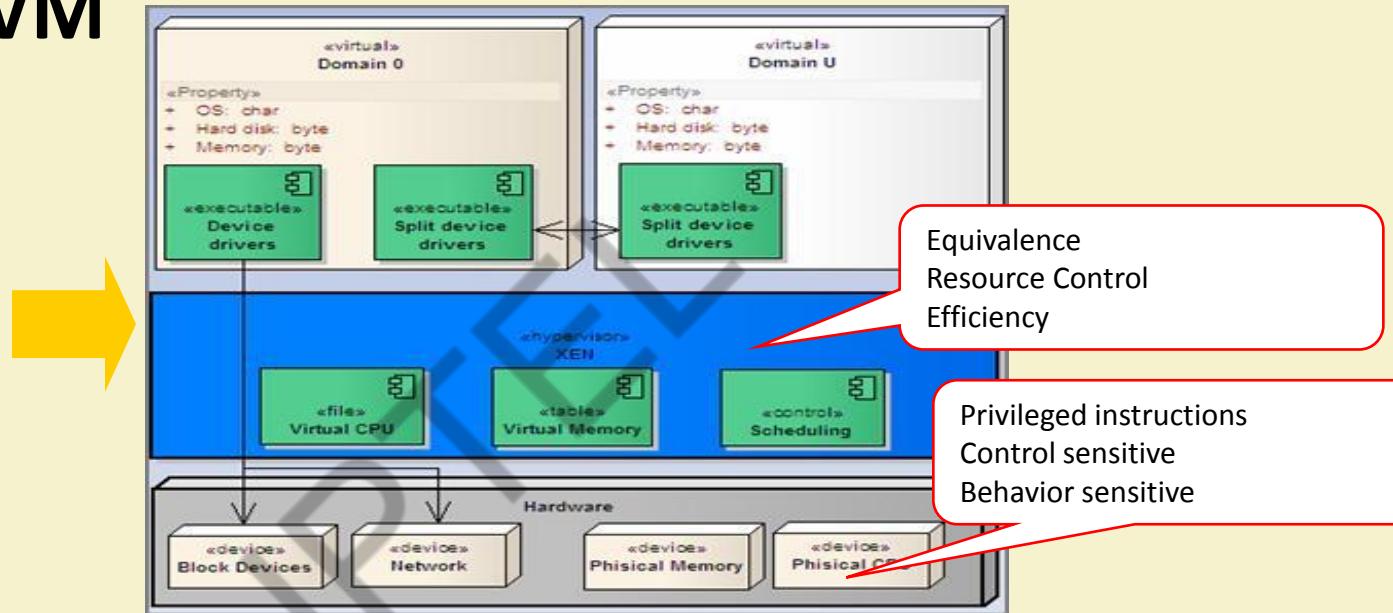
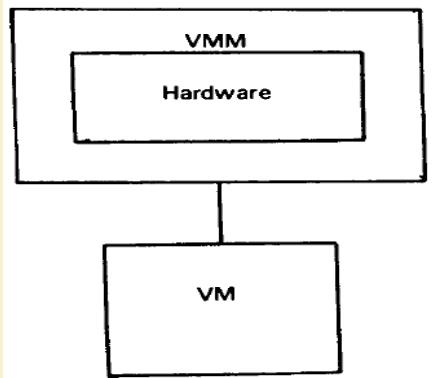
- **privileged** instructions, which cause a trap if executed in user mode, and
- **sensitive** instructions, which change the underlying resources (e.g. doing I/O or changing the page tables) or observe information that indicates the current privilege level (thus exposing the fact that the guest OS 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/eci/2008/courses/n2/Virtualization-Introduction.ppt

VMM and VM

Fig. 1. The virtual machine monitor.



- For any conventional third generation computer, a VMM may be constructed if the set of sensitive instructions for that computer is a subset of the set of privileged instructions
- A conventional third generation computer is recursively virtualizable if it is virtualizable and a VMM without any timing dependencies can be constructed for it.

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

Approaches to Server Virtualization



IIT KHARAGPUR

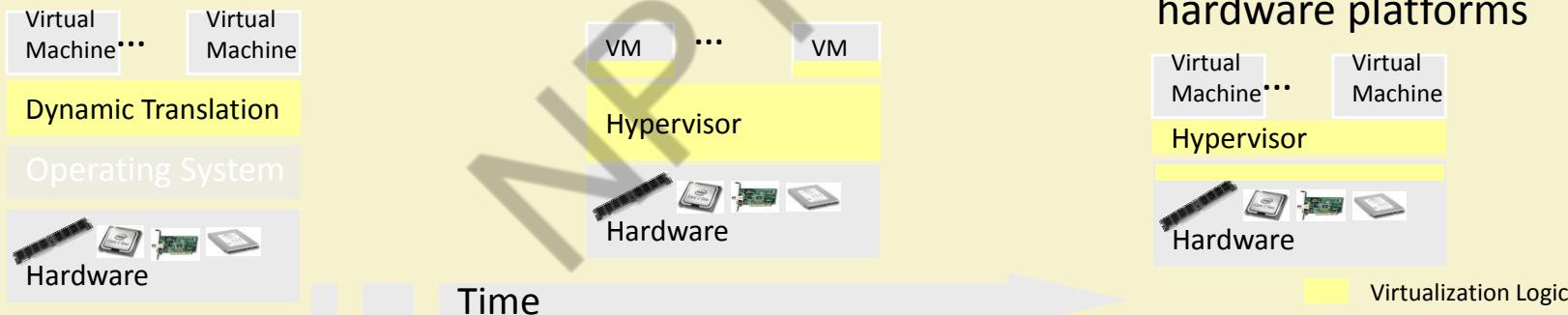


NPTEL

NPTEL ONLINE
CERTIFICATION COURSES

Evolution of Software Solutions

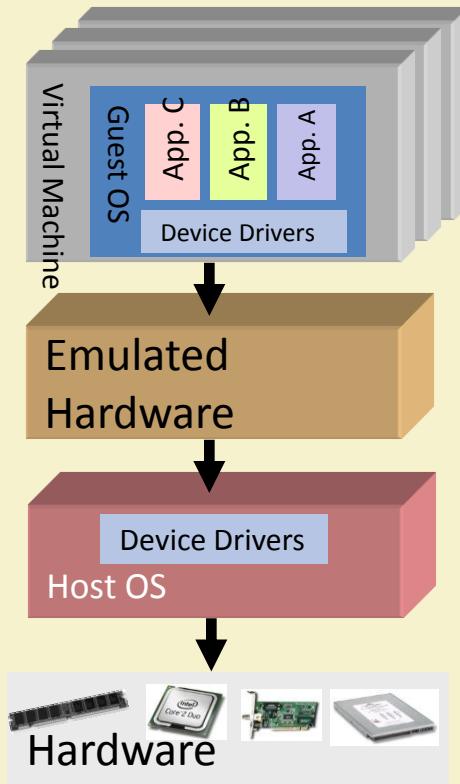
- 1st Generation: Full virtualization (Binary rewriting)
 - Software Based
 - VMware and Microsoft
- 2nd Generation: Para-virtualization
 - Cooperative virtualization
 - Modified guest
 - VMware, Xen
- 3rd Generation: Silicon-based (Hardware-assisted) virtualization
 - Unmodified guest
 - VMware and Xen on virtualization-aware hardware platforms



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

Full Virtualization

- 1st Generation offering of x86/x64 server virtualization
- Dynamic binary translation
 - Emulation layer talks to an operating system which talks to the computer hardware
 - Guest OS doesn't see that it is used in an emulated environment
- All of the hardware is emulated including the CPU
- Two popular open source emulators are QEMU and Bochs



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

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.dc.uba.ar/events/eci/2008/courses/n2/Virtualization-Introduction.ppt



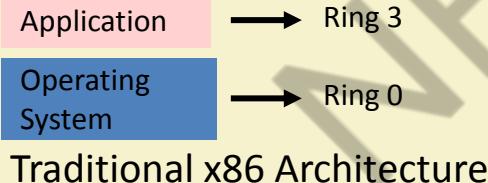
IIT KHARAGPUR



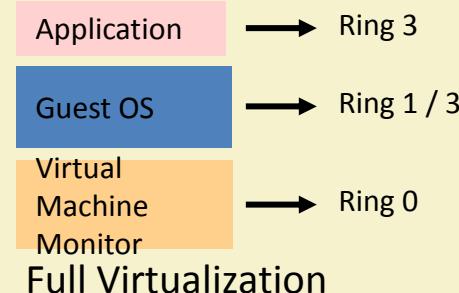
NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

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



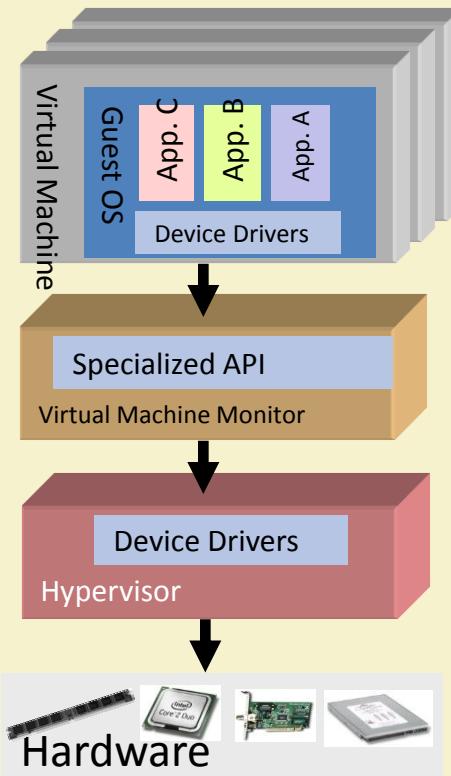
Full Virtualization

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

Para-Virtualization

- Guest OS is modified and thus run kernel-level operations at Ring 1 (or 3)
 - Guest is fully aware of how to process privileged instructions
 - Privileged instruction translation by the VMM is no longer necessary
 - Guest operating system uses a specialized API to talk to the VMM and, in this way, execute the privileged instructions
- VMM is responsible for handling the virtualization requests and putting them to the hardware

Server virtualization approaches



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

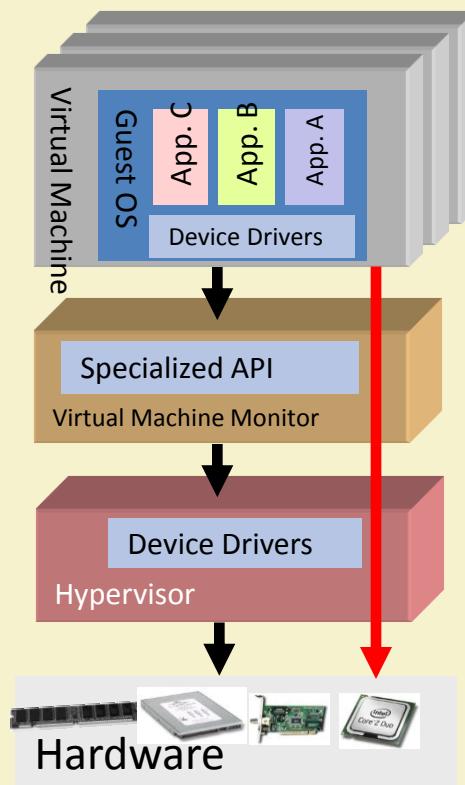
Para-Virtualization

- 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.uba.ar/events/eci/2008/courses/n2/Virtualization-Introduction.ppt

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



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

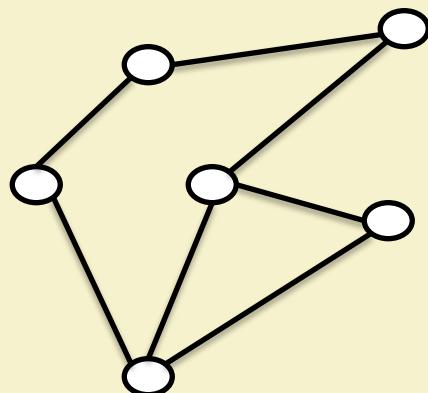
Hardware-assisted virtualization

- 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

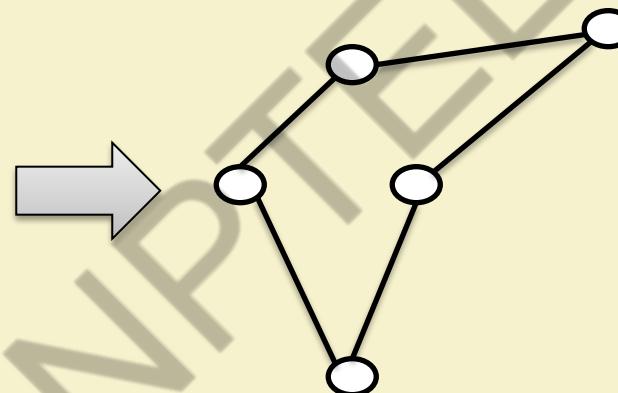
Source: www.dc.uba.ar/events/eci/2008/courses/n2/Virtualization-Introduction.ppt

Network Virtualization

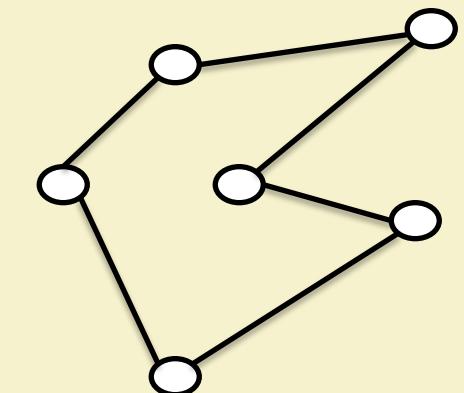
Making a physical network appear as multiple logical ones



Physical Network



Virtualized Network - 1



Virtualized Network - 2

Why Virtualize ?

- Internet is *almost “paralyzed”*
 - Lots of makeshift solutions (e.g. overlays)
 - A new architecture (aka clean-slate) is needed
- Hard to come up with a *one-size-fits-all* architecture
 - Almost impossible to predict what future might unleash
- Why not create an *all-sizes-fit-into-one* instead!
 - Open and expandable architecture
- Testbed for future networking architectures and protocols



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

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

Network Virtualization Model

- Business Model
- Architecture
- Design Principles
- Design Goals



IIT KHARAGPUR



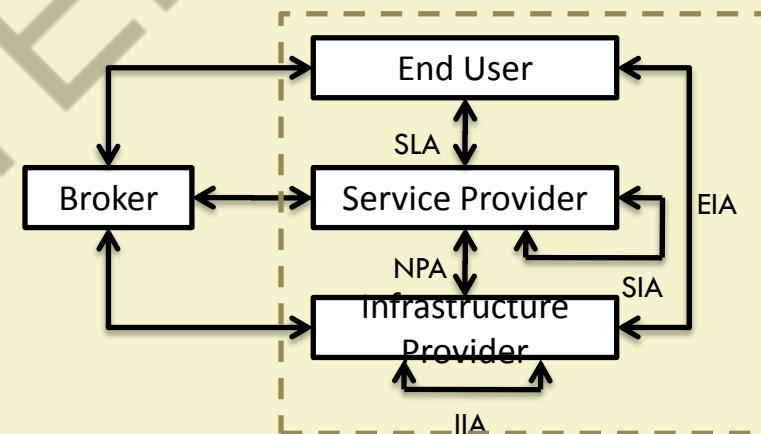
NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Business Model

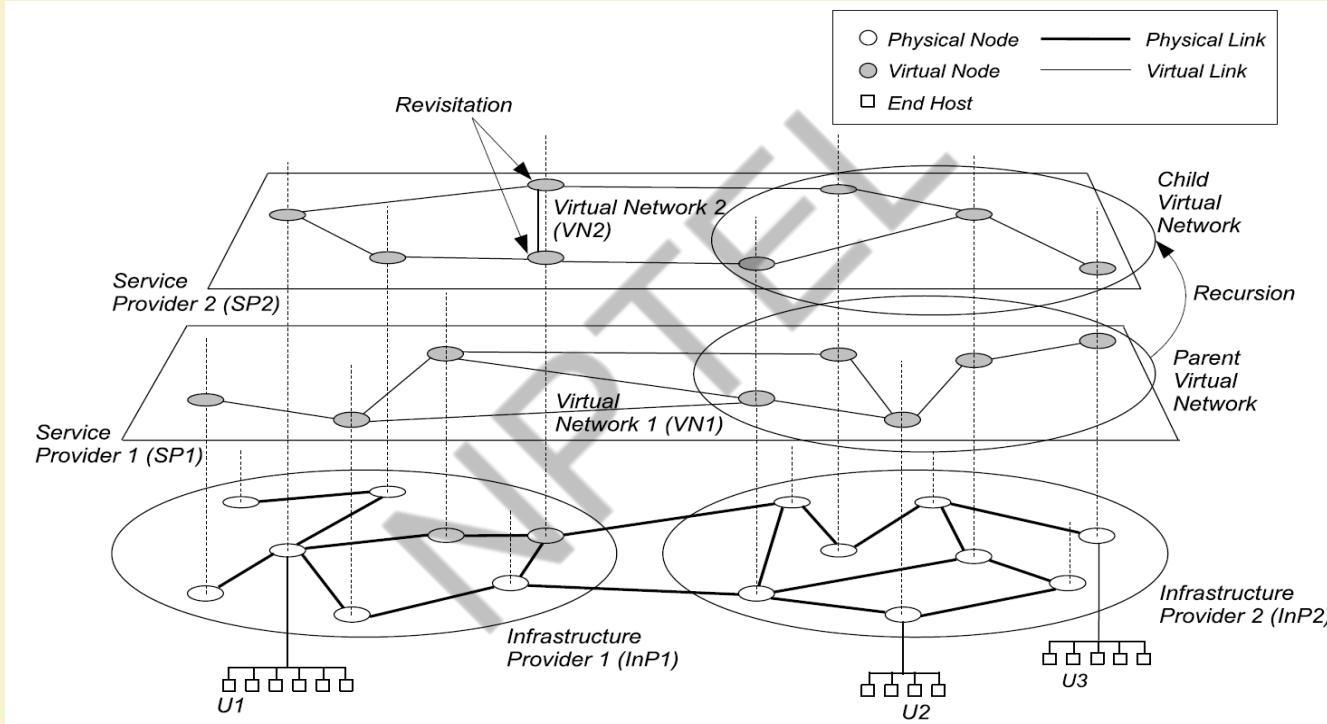
Players

- Infrastructure Providers (*InPs*)
 - Manage underlying physical networks
- Service Providers (*SPs*)
 - Create and manage virtual networks
 - Deploy customized end-to-end services
- End Users
 - Buy and use services from different service providers
- Brokers
 - Mediators/Arbiters

Relationships



Architecture



IIT KHARAGPUR

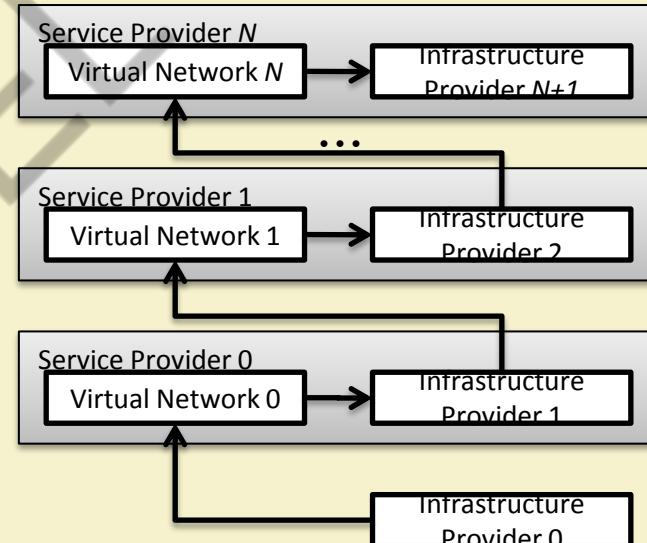


NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Design Principles

- Concurrency of multiple heterogeneous virtual networks
 - Introduces diversity
- Recursion of virtual networks
 - Opens the door for network virtualization economics
- Inheritance of architectural attributes
 - Promotes **value-addition**
- Revisitation of virtual nodes
 - Simplifies network operation and management

Hierarchy of Roles



Design Goals (1)

- Flexibility
 - Service providers can choose
 - arbitrary network topology,
 - routing and forwarding functionalities,
 - customized control and data planes
 - No need for co-ordination with others
 - IPv6 fiasco should never happen again
- Manageability
 - Clear separation of policy from mechanism
 - Defined *accountability* of infrastructure and service providers
 - Modular management



Design Goals (2)

- Scalability
 - Maximize the number of co-existing virtual networks
 - Increase resource utilization and amortize CAPEX and OPEX
- Security, Privacy, and Isolation
 - Complete isolation between virtual networks
 - *Logical and resource*
 - Isolate faults, bugs, and misconfigurations
 - Secured and private

Design Goals (3)

- Programmability
 - Of network elements e.g. routers
 - Answer “*How much*” and “*how*”
 - Easy and effective without being vulnerable to threats
- Heterogeneity
 - Networking technologies
 - Optical, sensor, wireless etc.
 - Virtual networks



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Design Goals (4)

- Experimental and Deployment Facility
 - PlanetLab, GENI, VINI
 - Directly deploy services in real world from the testing phase
- Legacy Support
 - Consider the existing Internet as a member of the collection of multiple virtual Internets
 - *Very important* to keep all concerned parties satisfied

Definition

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**.



Typical Approach

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



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

Thank You!

NPTEL



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

Introduction to XML: *eXtensible Markup Language*

Prof. Soumya K Ghosh

Department of Computer Science and Engineering
IIT KHARAGPUR

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
NPTEL ONLINE
CERTIFICATION COURSES

Presentation Outline

- What is XML (basic introduction)
 - Language rules, basic XML processing
- Defining language dialects
 - DTDs, schemas, and namespaces
- XML processing
 - Parsers and parser interfaces
 - XML-based processing tools
- XML messaging
 - Why, and some issues/example
- Conclusions



IIT KHARAGPUR

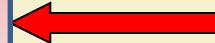


NPTEL ONLINE
CERTIFICATION COURSES

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

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

Example Revisited

element

tags

attribute of this quantity element

```
<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>
```

Hierarchical, structured information



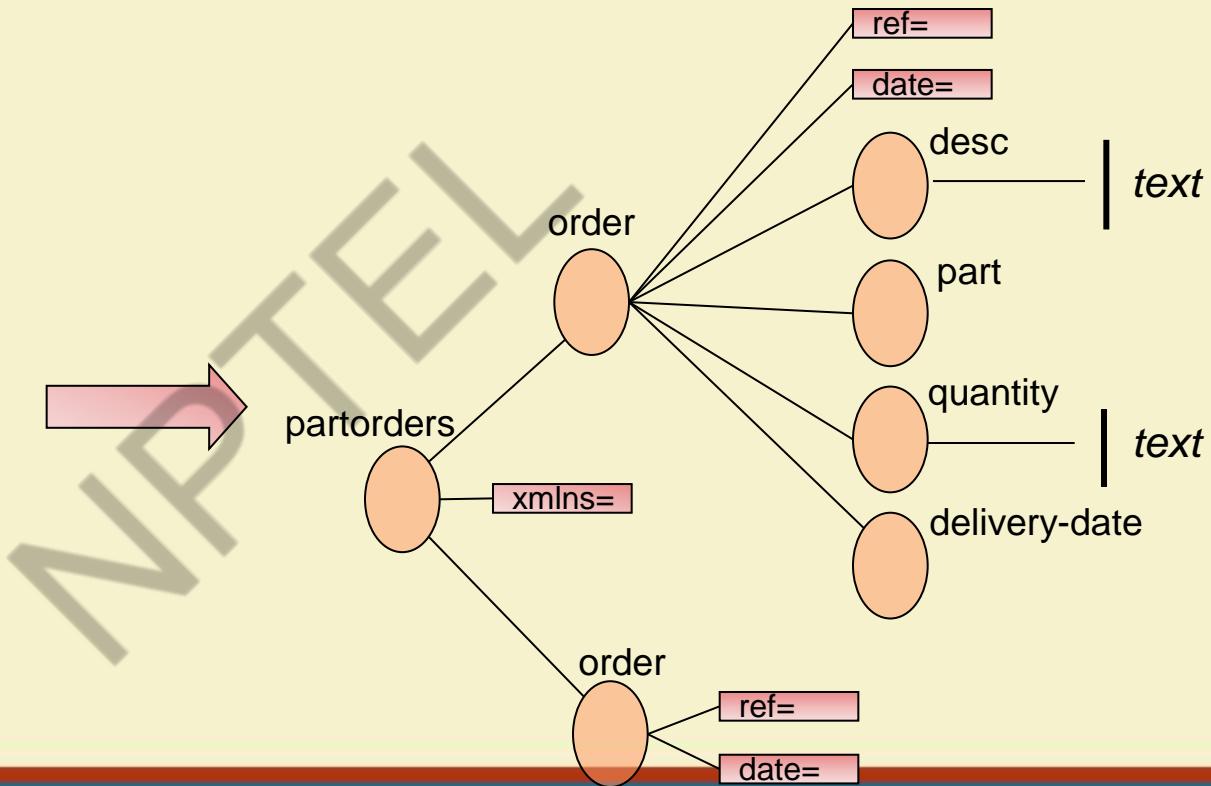
IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

XML Data Model - A Tree

```
<partorders xmlns="...">  
  <order date="..."  
    ref="...">  
    <desc> ..text..  
    </desc>  
    <part />  
    <quantity />  
    <delivery-date />  
  </order>  
  <order ref=".." .../>  
</partorders>
```



IIT KHARAGPUR



NPTEL
ONLINE
CERTIFICATION COURSES

XML: Why it's this way

- **Simple** (like HTML -- but not quite so simple)
 - Strict **syntax** rules, to eliminate syntax errors
 - syntax **defines** structure (hierarchically), and **names** structural parts (element names) -- it is **self-describing data**
- **Extensible** (unlike HTML; vocabulary is not fixed)
 - Can create your own **language** of tags/elements
 - **Strict syntax** ensures that such markup can be reliably processed
- Designed for a **distributed environment** (like HTML)
 - Can have data all over the place: can retrieve and use it reliably
- Can **mix** different data types together (unlike HTML)
 - Can mix one set of tags with another set: resulting data can still be reliably processed



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

XML Processing

```
<?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>
```

xml-simple.xml

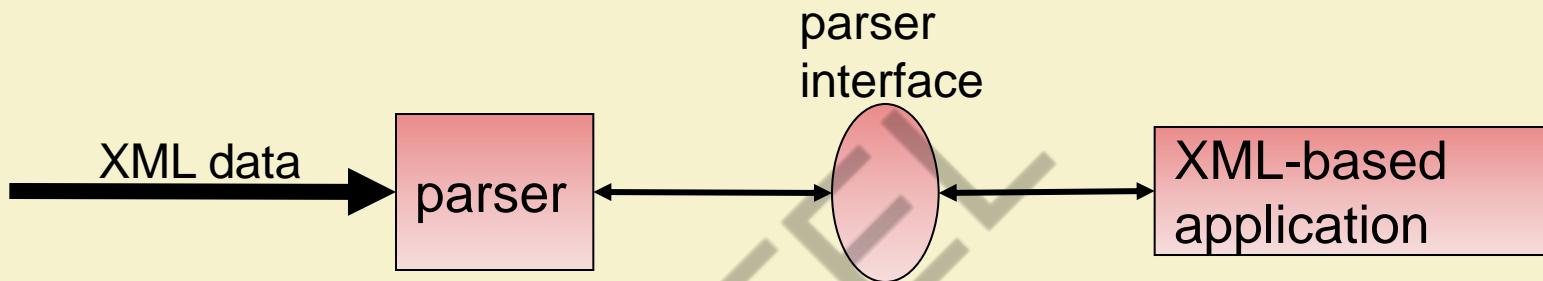


IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

XML Parser Processing Model



- The parser must verify that the XML data is syntactically correct.
- Such data is said to be ***well-formed***
 - The minimal requirement to “be” XML
- A parser **MUST** stop processing if the data isn’t well-formed
 - E.g., stop processing and “throw an exception” to the XML-based application. The XML 1.0 spec ***requires*** this behaviour

XML Processing Rules: Including Parts

```
<?xml version="1.0" encoding="utf-8" ?>  
  Document Type Declaration (DTD)  
  
<!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> "  
    ]>  
  Internal Entity declaration  
  
<transfers>  
  Entity reference  
  &messageHeader;  
  &name;  
  Content omitted . . .  
</transfers>  
  <fundsTransfer date="20010923T12:34:34Z">  
    <from type="intrabank">  
      . . .  
    </from>  
  </fundsTransfer>  
  . . .  
</transfers>
```

xml-simple-intEntity.xml

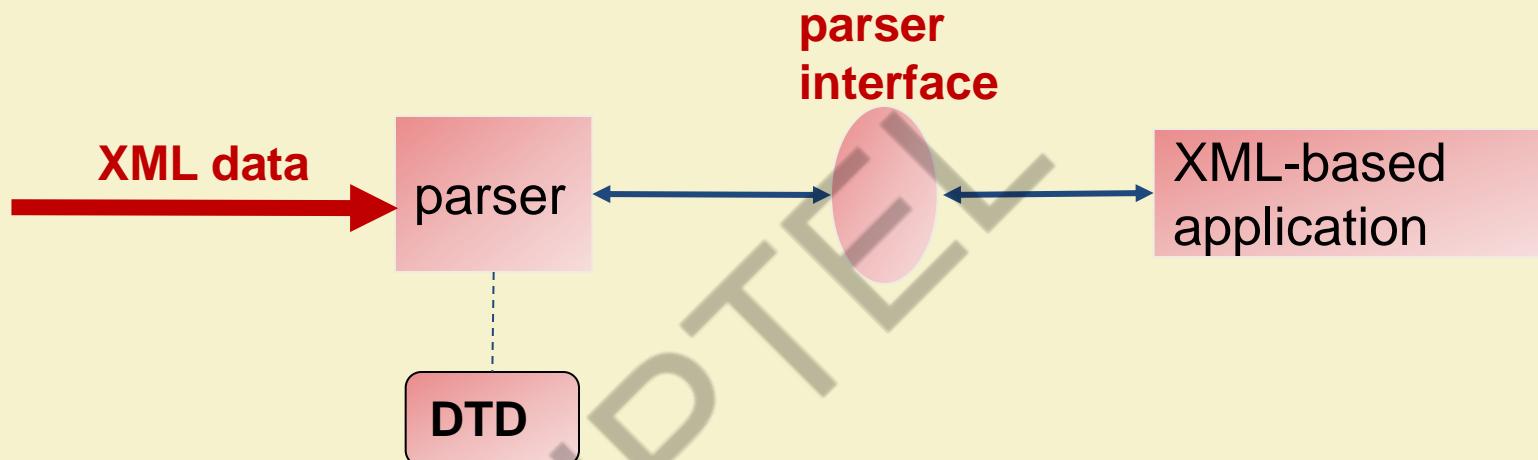


IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

XML Parser Processing Model



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

XML Parsers, DTDs, and Internal Entities

- The parser processes the DTD content, identifies the internal entities, and checks that each entity is well-formed.
- There are explicit syntax rules for DTD content -- well-formed XML must be correct here also.
- The parser then replaces every occurrence of an **entity reference** by the referenced entity (and does so recursively within entities)
- The “resolved” data object is then made available to the XML application



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

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">  
            . . . Content omitted . . .  
    </transfers>
```

External Entity declaration

Location given via a URL

xml-simple-extEntity.xml



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

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
NPTEL ONLINE
CERTIFICATION COURSES

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**



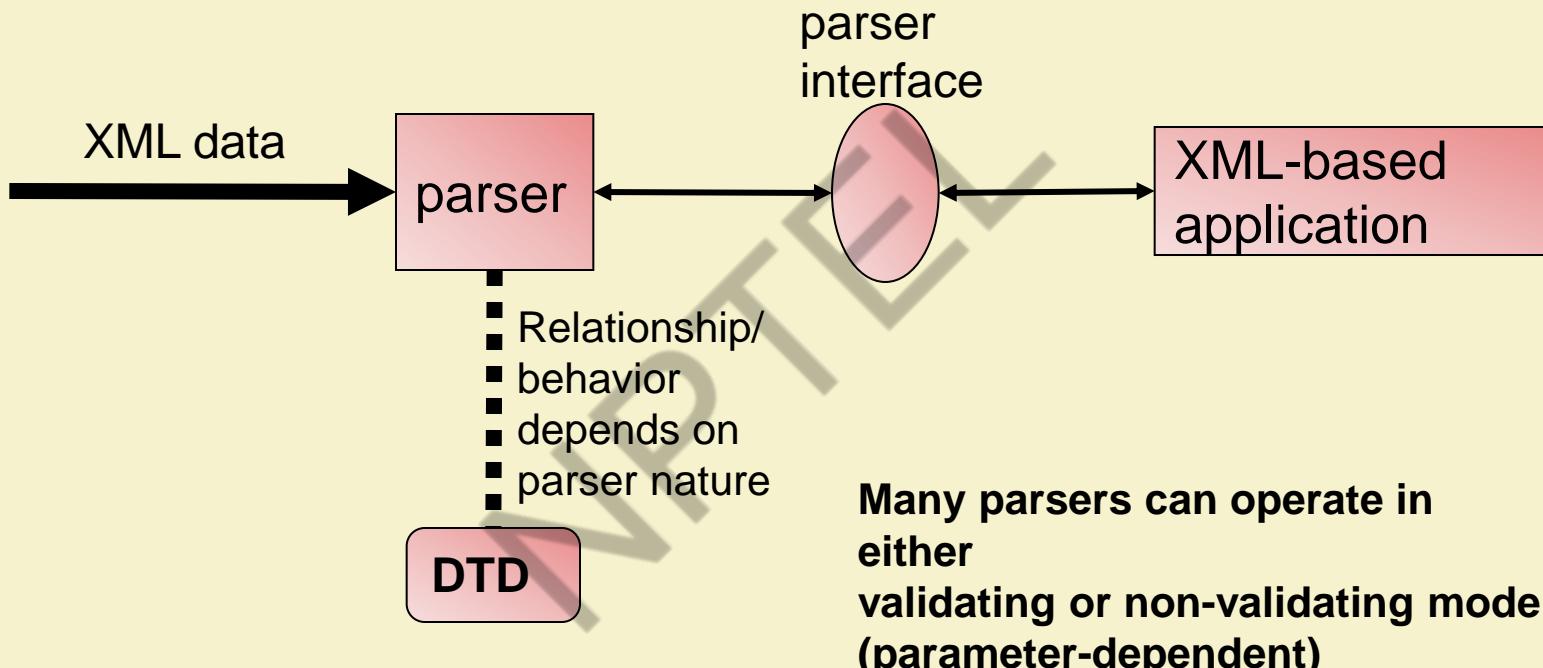
IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES



XML Parser Processing Model



Special Issues: Characters andCharsets

- 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:
 ♠ -- the spades character (♠)

 9824th character in the Unicode character set
- Also, binary data must be encoded as **printable characters**



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Presentation Outline

- What is XML (basic introduction)
 - Language rules, basic XML processing
- Defining language dialects
 - DTDs, schemas, and namespaces
- XML processing
 - Parsers and parser interfaces
 - XML-based processing tools
- XML messaging
 - Why, and some issues/example
- Conclusions



IIT KHARAGPUR



NPTEL
ONLINE
CERTIFICATION COURSES

How do you define language dialects?

- Two ways of doing so:
 - **XML Document Type Declaration (DTD)** -- Part of core XML spec.
 - **XML Schema** -- New XML specification (2001), which allows for stronger constraints on XML documents.
- Adding dialect specifications implies ***two classes*** of XML data:
 - **Well-formed** An XML document that is syntactically correct
 - **Valid** An XML document that is both well-formed *and* consistent with a specific DTD (or Schema)
- What DTDs and/or schema specify:
 - Allowed element and 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 XML models

Example DTD (as part of document)

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

xml-simple-valid.xml



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Example “External” DTD

- Reference is using a variation on the DOCTYPE:

```
<!DOCTYPE transfers SYSTEM  
      "http://www.foo.org/hereitis/simple.dtd" >
```

simple.dtd

```
<transfers>  
  <fundsTransfer date="20010923T12:34:34Z">  
    ... As with previous example ...  
  </fundsTransfer>  
</transfers>
```

- Of course, the DTD file must be there, and accessible.



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Thank You!!

NPTEL



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

Introduction to XML: *eXtensible Markup Language*

Prof. Soumya K Ghosh

Department of Computer Science and Engineering
IIT KHARAGPUR

XML Schemas

- A new specification (2001) for specifying validation rules for XML

Specs:

<http://www.w3.org/XML/Schema>

Best-practice:

<http://www.xfront.com/BestPracticesHomepage.html>

- 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



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

XML Schema version of our DTD (Portion)

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

```

simple.xsd



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

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
NPTEL ONLINE
CERTIFICATION COURSES

Mixing language dialects together

Namespaces let you do this relatively easily:

```
<?xml version= "1.0" encoding= "utf-8" ?>  
  
<html xmlns="http://www.w3.org/1999/xhtml1"  
      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>
```

Default 'space'
is *xhtml*

mt: prefix indicates 'space'
mathml (a different language)



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

Presentation Outline

- What is XML (basic introduction)
 - Language rules, basic XML processing
- Defining language dialects
 - DTDs, schemas, and namespaces
- XML processing
 - Parsers and parser interfaces
 - XML-based processing tools
- XML messaging
 - Why, and some issues/example
- Conclusions



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

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

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
NPTEL ONLINE
CERTIFICATION COURSES

XML Processing: DOM

B) DOM: Document Object Model

- <http://www.w3.org/DOM/>
 - An ***object-based*** interface
 - Parser generates an ***in-memory tree*** corresponding to the document
 - DOM interface defines methods for accessing and modifying the tree
- **Advantages**
 - Very useful for dynamic modification of, access to the tree
 - Useful for querying (I.e. looking for data) that depends on the tree structure [element.childNodes("2").getgetAttributeValue("boobie")]
 - Same interface for many programming languages (C++, Java, ...)
 - **Disadvantages**
 - Can be slow (needs to produce the tree), and may need lots of memory
 - DOM programming interface is a bit awkward, not terribly object oriented

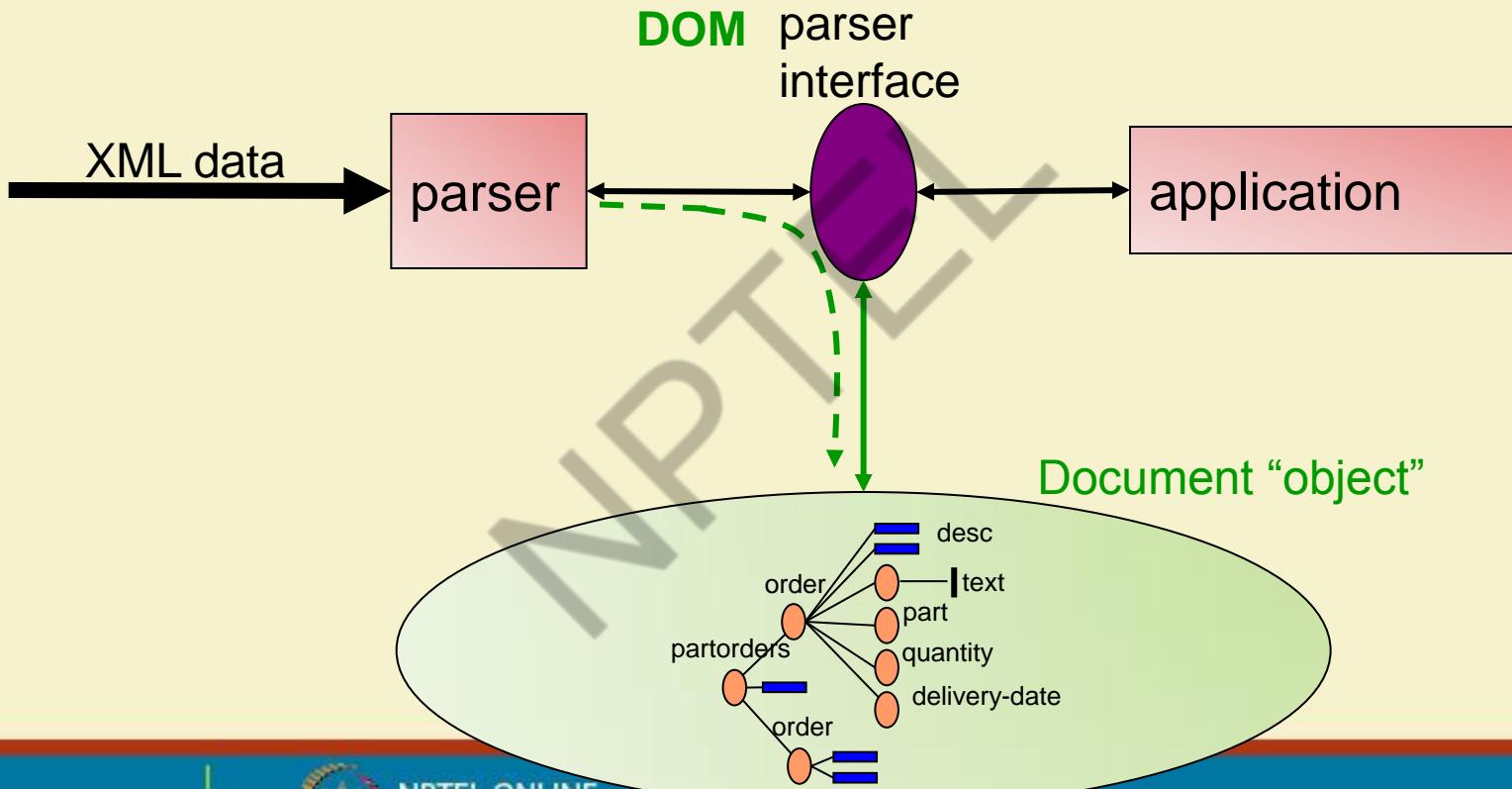


IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

DOM Parser Processing Model



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
NPTEL ONLINE
CERTIFICATION COURSES

C) dom4j: XML framework for Java

- <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!)

XML Processing: dom4j

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 latest version is always found 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
NPTEL ONLINE
CERTIFICATION COURSES

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



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

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)



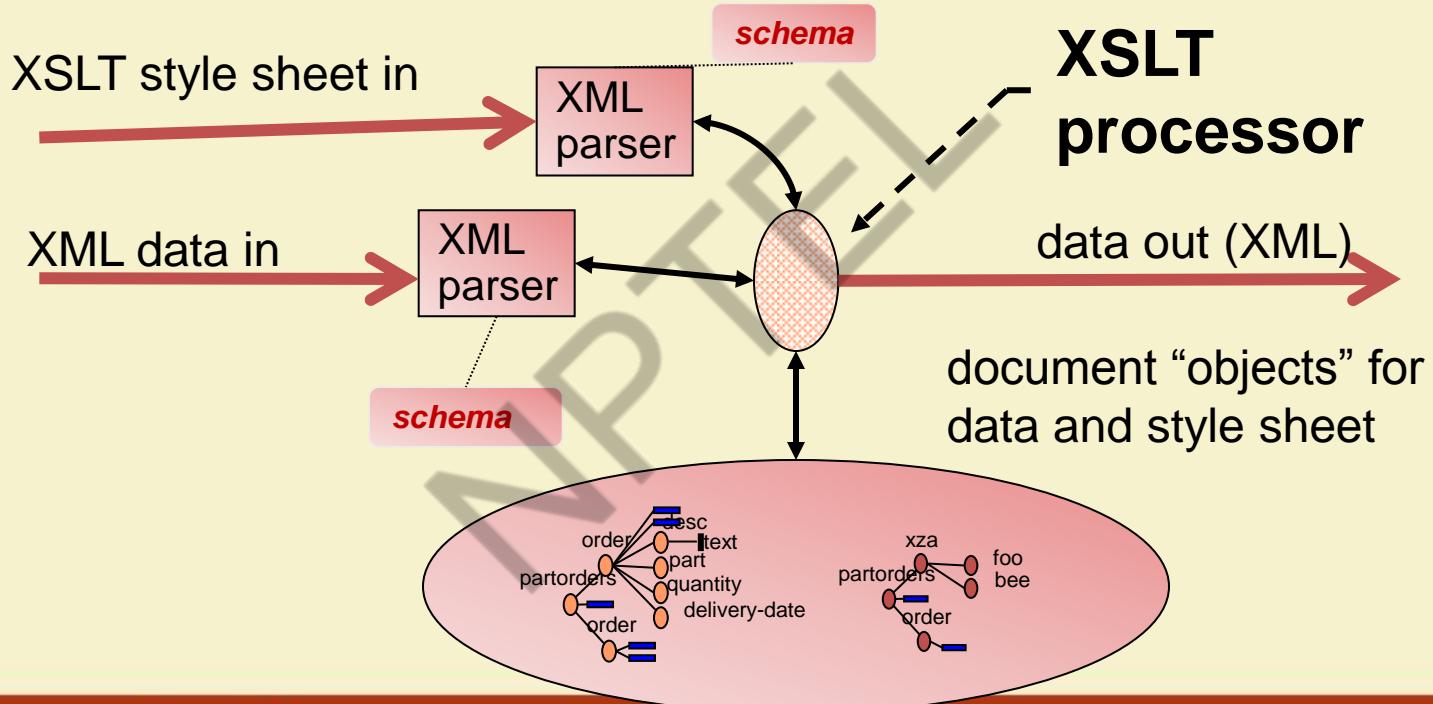
IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

XSLT processing model

- D) XSLT Processing model



Presentation Outline

- What is XML (basic introduction)
 - Language rules, basic XML processing
- Defining language dialects
 - DTDs, schemas, and namespaces
- XML processing
 - Parsers and parser interfaces
 - XML-based processing tools
- XML messaging
 - Why, and some issues/example



IIT KHARAGPUR



NPTEL
ONLINE
CERTIFICATION COURSES

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

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-olk
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



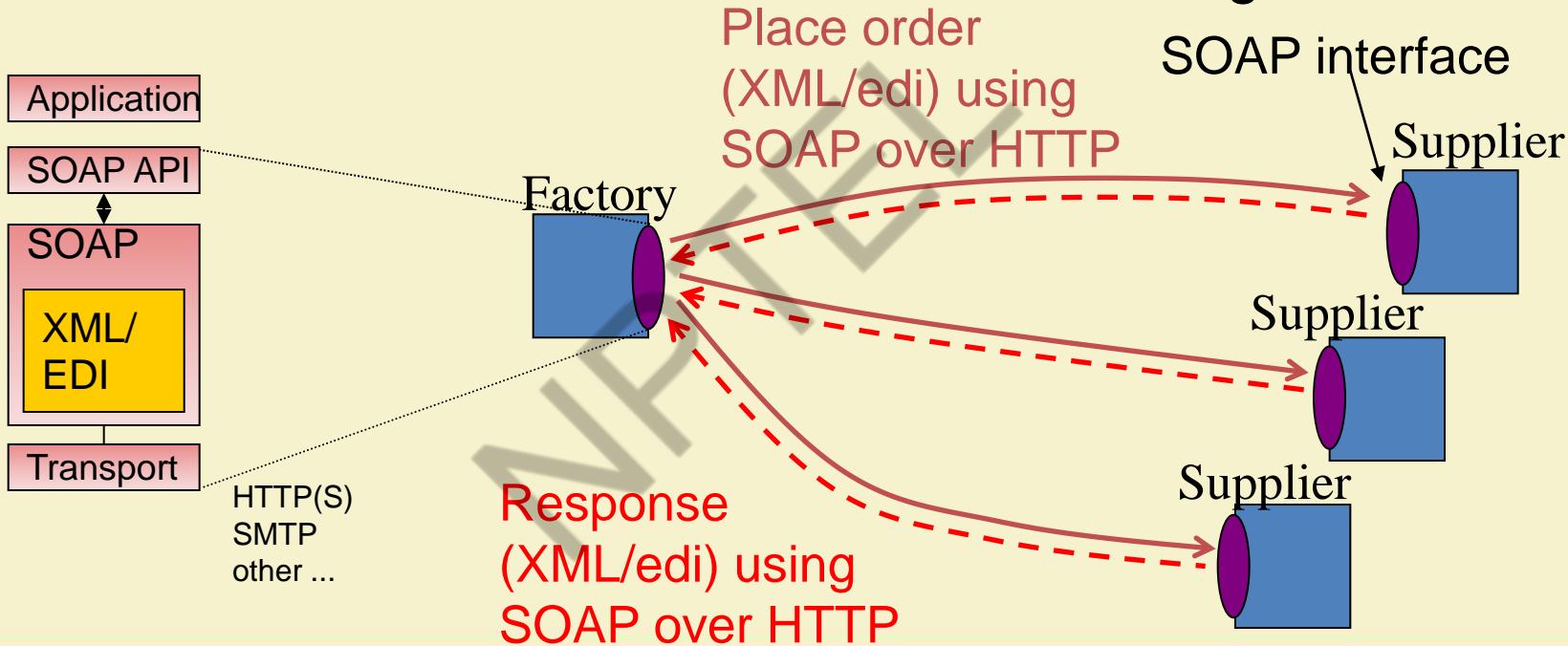
NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Some standards for message format

- 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.

XML Messaging + Processing

- XML as a *universal format* for data exchange



Presentation Outline

- What is XML (basic introduction)
 - Language rules, basic XML processing
- Defining language dialects
 - DTDs, schemas, and namespaces
- XML processing
 - Parsers and parser interfaces
 - XML-based processing tools
- XML messaging
 - Why, and some issues/example
- Conclusions

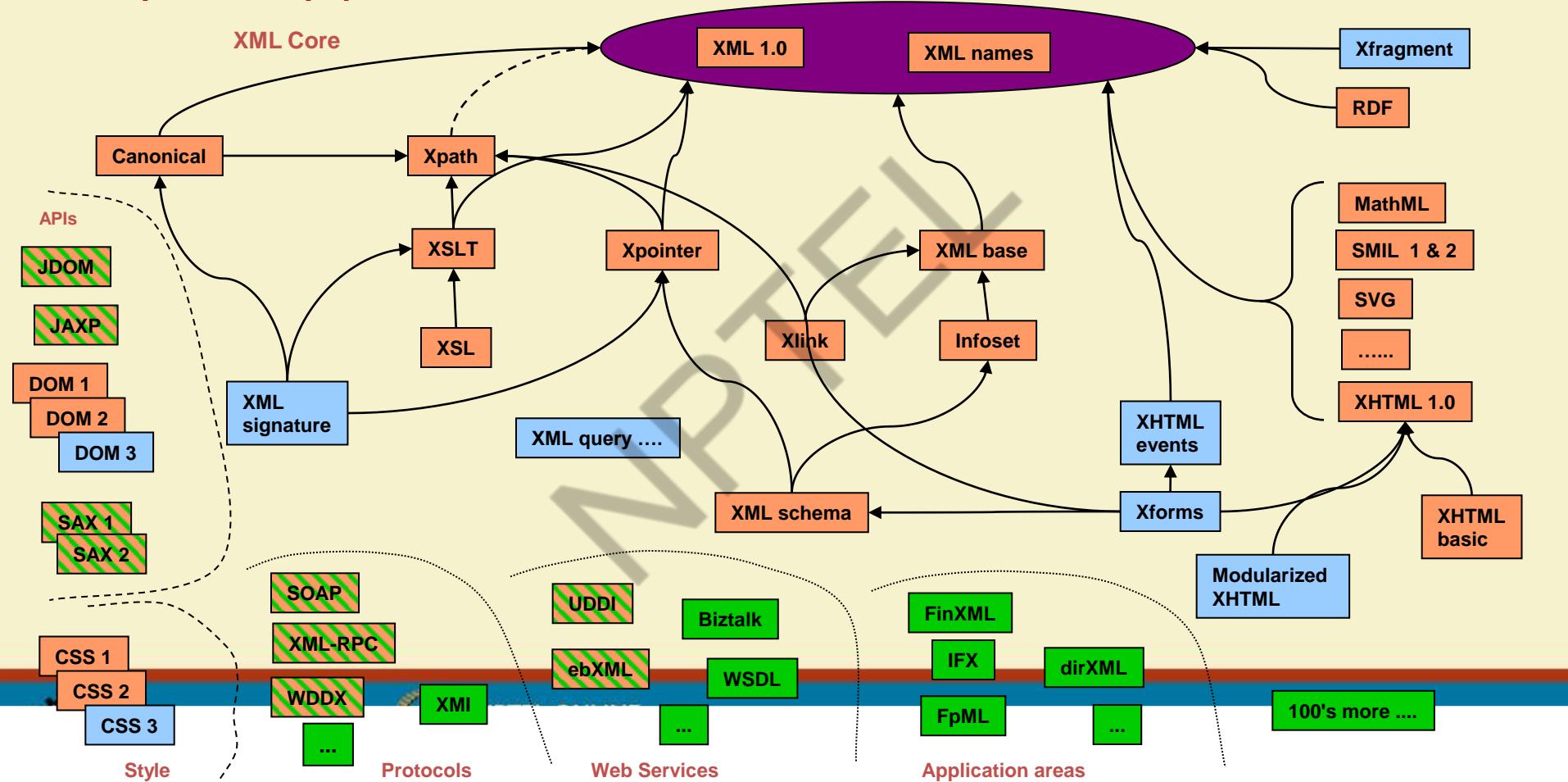


IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

XML (and related) Specifications



Thank You!!

NPTEL



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

CLOUD COMPUTING

Web Services, Service Oriented Architecture

PROF. SOUMYA K. GHOSH

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

IIT KHARAGPUR

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

History!

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



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

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

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

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

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

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

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

Example: Web based purchase

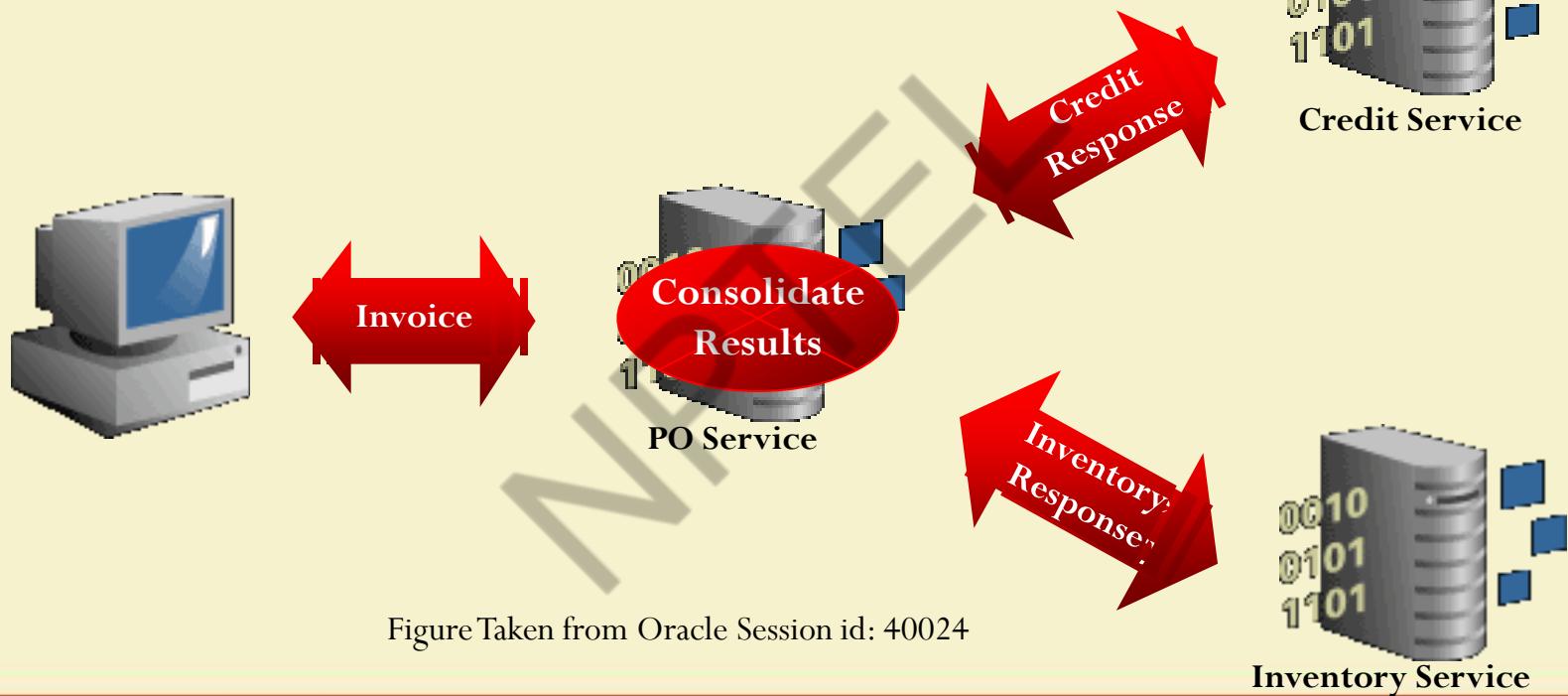
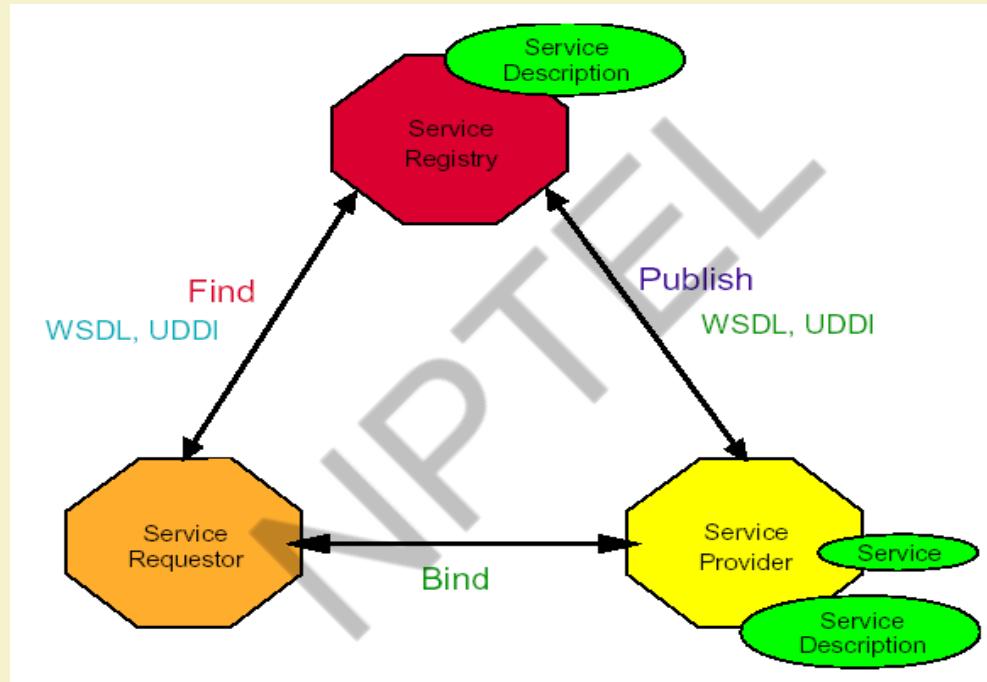


Figure Taken from Oracle Session id: 40024

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/>

Web Service Model



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

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

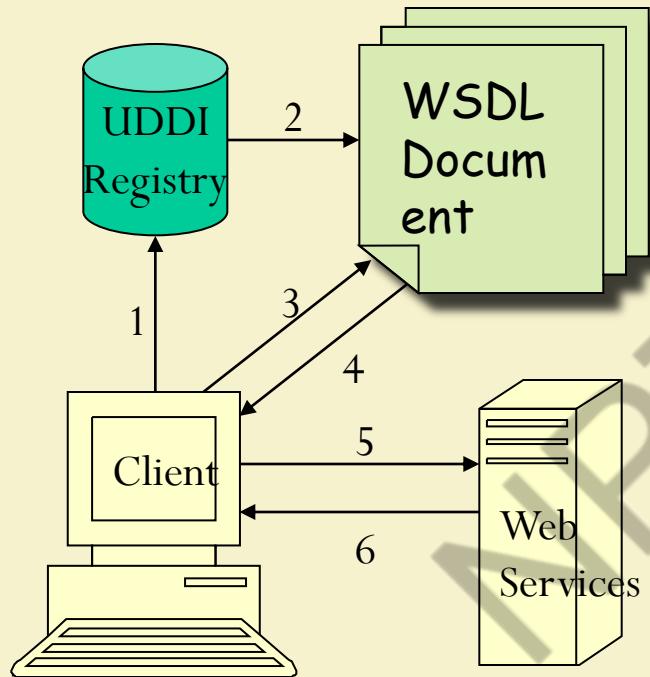
Web Service Model (contd...)

- 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

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.

Steps of Operation



1. Client queries registry to locate service.
2. Registry refers client to WSDL document.
3. Client accesses WSDL document.
4. WSDL provides data to interact with Web service.
5. Client sends SOAP-message request.
6. Web service returns SOAP-message response.

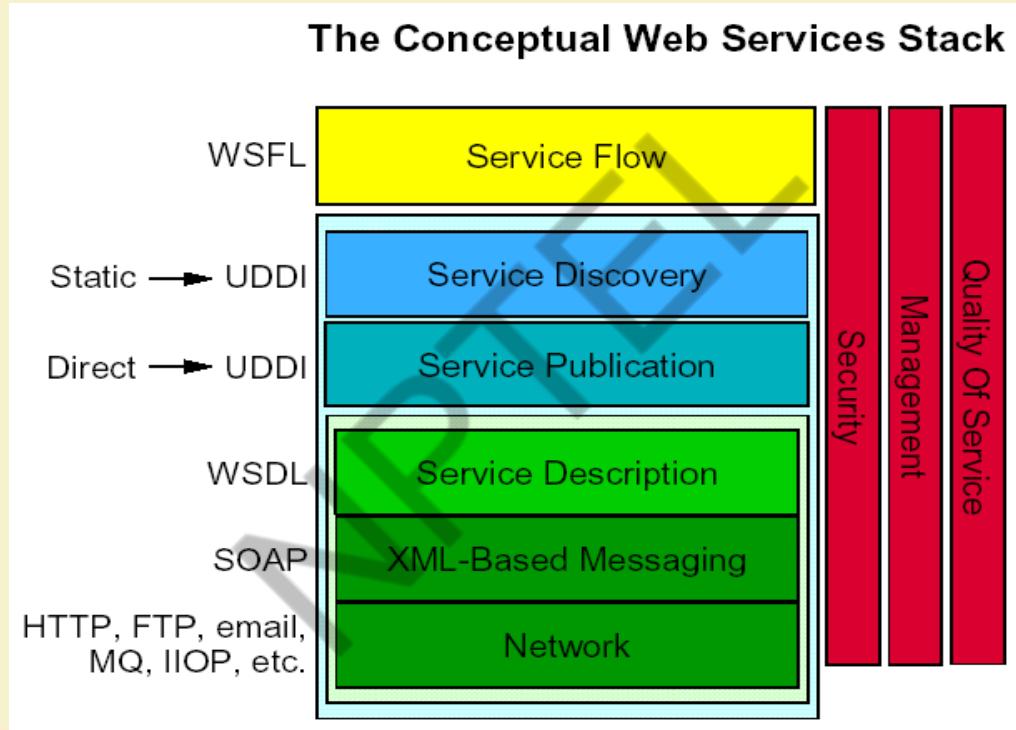


IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

Web Service Stack



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
NPTEL ONLINE
CERTIFICATION COURSES

XML (contd...)

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

XML vs HTML

An HTML example:

```
<html>
<body>
    <h2>John Doe</h2>
    <p>2 Backroads Lane<br>
        New York<br>
        045935435<br>
        john.doe@gmail.com<br>
    </p>
</body>
</html>
```

XML vs HTML (contd...)

- This will be displayed as:

John Doe

2 Backroads Lane

New York

045935435

John.doe@gmail.com

- 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.

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>
```

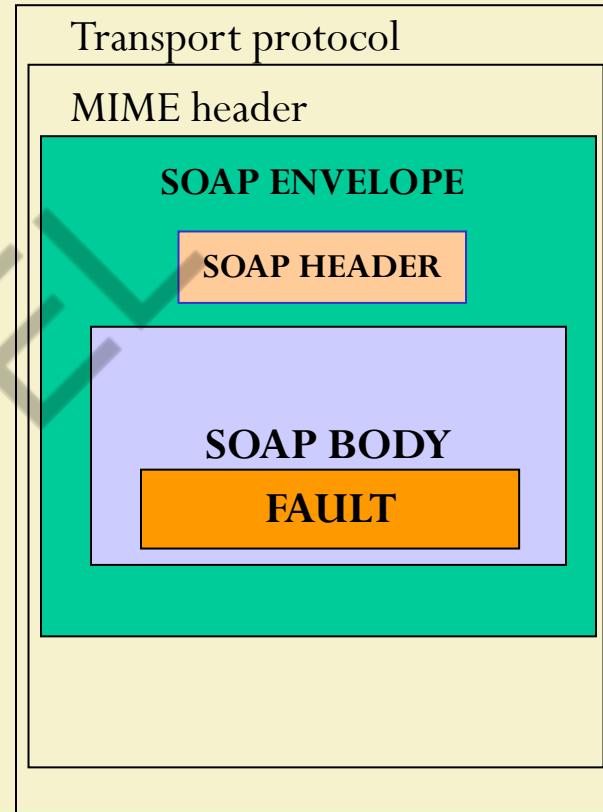
- In this case:
 - The information contained is being marked, but not for displaying.
 - Readable by both human and machines.

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

SOAP Building Blocks

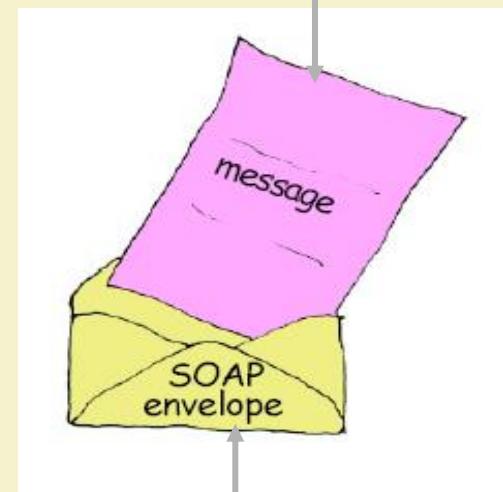
- Envelope (required) – identifies XML document as SOAP message
- Header (optional) – contains header information
- Body (required) – call and response information
- Fault (optional) – errors that occurred while processing message



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

Application-specific
message vocabulary



SOAP Envelope vocabulary



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

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>
```



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

SOAP Response

HTTP/1.1 200 OK

Content-Type: application/soap; charset=utf-8

Content-Length: 126

```
<?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>
```



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

Why SOAP?

- 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

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.

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
NPTEL ONLINE
CERTIFICATION COURSES

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

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.

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.

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>
```

Binding to SOAP

```
<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" />
  <operation>
    <soap:operation
      soapAction="http://example.com/getTerm"/>
    <input>
      <soap:body use="literal"/>
    </input>
    <output>
      <soap:body use="literal"/>
    </output>
  </operation>
</binding>
```



IIT KHARAGPUR



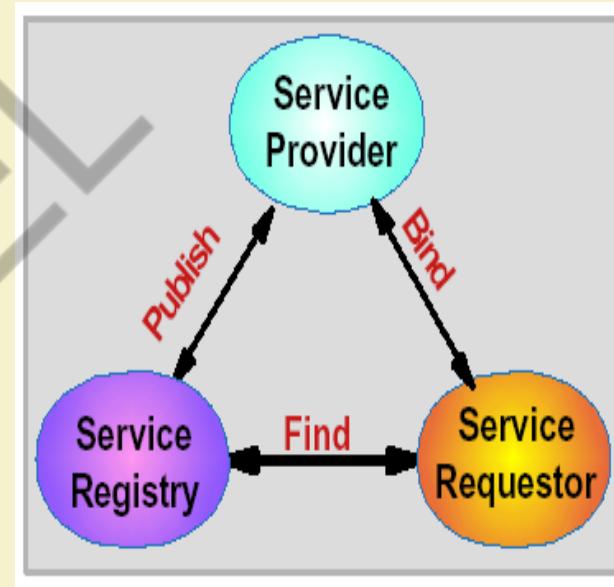
NPTEL ONLINE
CERTIFICATION COURSES

UDDI - Universal Description, Discovery, and Integration

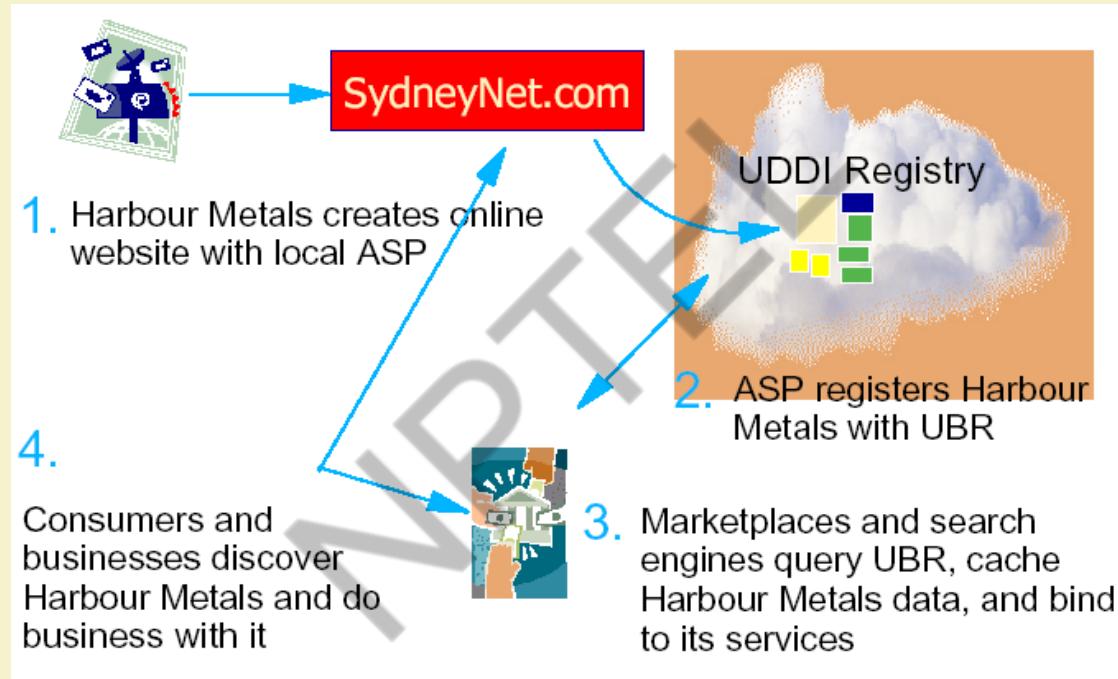
- A framework to define XML-based registries
- Registries are repositories that contain documents that describe business data and also provide search capabilities and programmatic access to remote applications
- Businesses can publish information about themselves and the services they offer
- Can be interrogated by SOAP messages and provides access to WSDL documents describing web services in its directory

UDDI Roles and Operations

- Service Registry
 - Provides support for publishing and locating services
 - Like telephone yellow pages
- Service Provider
 - Provides e-business services
 - Publishes these services through a registry
- Service requestor
 - Finds required services via the Service Broker
 - Binds to services via Service Provider



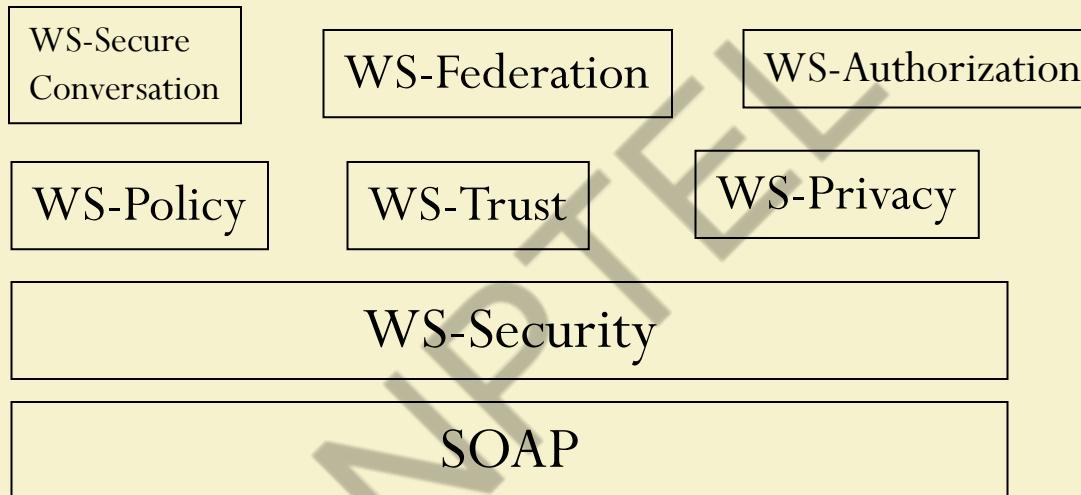
How can UDDI be Used?



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

Web Services Security Architecture



Thank You!

NPTEL



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

CLOUD COMPUTING

SERVICE LEVEL AGREEMENT (SLA)

PROF. SOUMYA K. GHOSH

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
IIT KHARAGPUR

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

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

Web Service SLA

- 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



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

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

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



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

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
NPTEL ONLINE
CERTIFICATION COURSES

Service Level Management

- Monitoring and measuring performance of services based on SLOs
- Provider perspective :
 - Make decisions based on business objectives and technical realties
- Consumer perspective :
 - Decisions about how to use cloud services



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

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
NPTEL ONLINE
CERTIFICATION COURSES

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.

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.

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.

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)*

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

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

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.

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 Deletion		✓	✓
Hardware Erasure and Destruction		✓	✓
Regulatory Compliance	✓	✓	✓
Transparency	✓	✓	✓
Certification	✓	✓	✓
Terminology for Key Performance Indicators		✓	✓
Metrics	✓	✓	✓
Auditability	✓	✓	✓
Monitoring	✓	✓	✓
Machine-Readable SLAs		✓	

Source: "Cloud Computing Use Cases White Paper" Version 4.0

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

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

Example Cloud SLAs (contd...)

Cloud Provider	Service	Type of Cloud Delivery Model	Service Level Agreement Guarantees
Rackspace	Cloud Server	IaaS	<p>Availability regarding the following: Internal Network (100%), Data Center Infrastructure (100%), Load balancers (99.9%)</p> <p>Performance related to service degradation: Server migration, notified 24 hours in advance, and is completed in 3 hours (maximum)</p> <p>Recovery Time: In case of failure, guarantee of restoration/recovery in 1 hour after the problem is identified.</p>
Terremark	vCloud Express	IaaS	<p>Monthly Uptime Percentage (100%) with the following definitions: Service Credit, Credit Request and Payment Procedure, and SLA exclusions</p>

Example Cloud SLAs (contd...)

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

Limitations

- Service measurement
 - Restricted to uptime percentage
 - Measured by taking the mean of service availability observed over a specific period of time
 - Ignores other parameters like stability, capacity, etc.
- Biasness towards vendors
 - Measurement of parameters are mostly established according to vendor's advantage
- Lack of active monitoring on customer's side
 - Customers are given access to some ticketing systems and are responsible for monitoring the outages.
 - Providers do not provide any access to active data streams or audit trails, nor do they report any outages.

Limitations (contd...)

- Gap between *QoS hype* and *SLA offerings* in reality
- QoS in the areas of *governance, reliability, availability, security, and scalability* are not well addressed.
- No formal ways of verifying if the SLA guarantees are complying or not.
- Proper SLA are good for both provider as well as the customer
 - Provider's perspective : Improve upon Cloud infrastructure, fair competition in Cloud market place
 - Customer's perspective : Trust relationship with the provider, choosing appropriate provider for moving respective businesses to Cloud

Expected SLA Parameters

- Infrastructure-as-a-Service (IaaS):
 - CPU capacity, cache memory size, boot time of standard images, storage, scale up (maximum number of VMs for each user), scale down (minimum number of VMs for each user), On demand availability, scale uptime, scale downtime, auto scaling, maximum number of VMs configured on physical servers, availability, cost related to geographic locations, and response time
- Platform-as-a-Service (PaaS):
 - Integration, scalability, billing, environment of deployment (licenses, patches, versions, upgrade capability, federation, etc.), servers, browsers, number of developers

Expected SLA Parameters *(contd...)*

- Software-as-a-Service (SaaS):
 - Reliability, usability, scalability, availability, customizability, Response time
- Storage-as-a-Service :
 - Geographic location, scalability, storage space, storage billing, security, privacy, backup, fault tolerance/resilience, recovery, system throughput, transferring bandwidth, data life cycle management

Thank You!

NPTEL



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

Cloud Computing : Economics

Prof. Soumya K Ghosh

Department of Computer Science and Engineering
IIT KHARAGPUR

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.

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.

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.

A Useful Measure of “Smoothness”

- The coefficient of variation C_V
 - \neq the variance σ^2 nor the correlation coefficient
- Ratio of the standard deviation σ to the absolute value of the mean $|\mu|$
- “Smoother” 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**

Coefficient of variation C_v

- X_1, X_2, \dots, X_n independent random variables for demand
 - Identical standard variation σ and mean μ
- Aggregated demand
 - Mean \rightarrow sum of means: $n \cdot \mu$
 - Variance \rightarrow sum of variances: $n \cdot \sigma^2$
 - Coefficient of variance $\rightarrow \frac{\sqrt{n} \cdot \sigma}{n \cdot \mu} = \frac{\sigma}{\sqrt{n} \cdot \mu} = \frac{1}{\sqrt{n}} C_v$
- Adding n independent demands reduces the C_v by $\frac{1}{\sqrt{n}}$
 - Penalty of insufficient/excess resources grows smaller
 - Aggregating 100 workloads bring the penalty to 10%

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 : $n.X$, variance of sum: $n^2\sigma^2(X)$
 - Mean: $n.\mu$, standard deviation: $n.\sigma(X)$
 - Coefficient of Variance remains constant
- Simultaneous peaks



IIT KHARAGPUR

9/3/2017



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

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 → 3rd parties take care of it

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 😞

Value of Location Independence

Contd...

- Supporting a global user base requires a dispersed service architecture
 - Coordination, consistency, availability, partition-tolerance
 - **Investment implications**

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 10,000/- per day
 - Rent a car for INR 45,000/- a day
 - If you need a car for 2 days in a trip, buying would be much more costly than renting
 - **It depends on the demand**



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 $[B_T : \text{Total Baseline Cost}]$
C	Cloud unit cost $[C_T : \text{Total Cloud Cost}]$
U (=C/B)	Utility Premium $[\text{For rental car example, } U=4.5]$

$$C_T = \int_0^T U \times B \times D(t) dt = A \times U \times B \times T$$

$$B_T = P \times B \times T$$

- 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

Utility Pricing in Real World

- In practice demands are often highly spiky
 - News stories, marketing promotions, product launches, Internet flash floods (Slashdot effect), tax season, Christmas shopping, processing a drone footage for a 1 week border skirmish, 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

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
 - I. 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

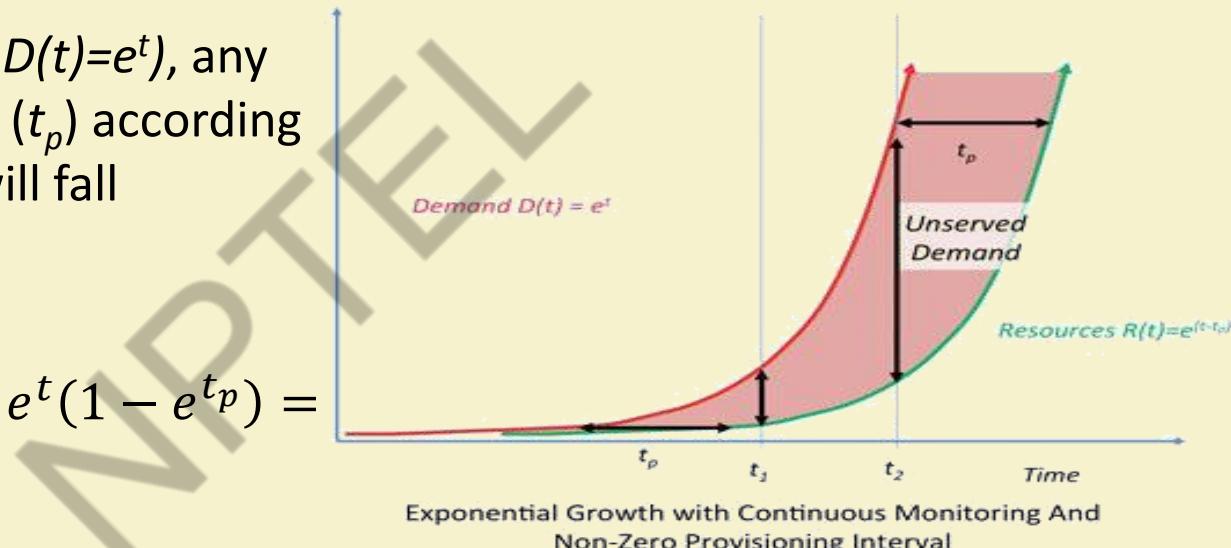
9/3/2017



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

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 $\propto c.k_1 e^t$



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.

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, δ is the delay in provisioning the extra computing recourse on demand

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)

Thank You!!

NPTEL



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES



IIT KHARAGPUR



NPTEL

NPTEL ONLINE
CERTIFICATION COURSES

Cloud Computing : *Managing Data*

Prof. Soumya K Ghosh

Department of Computer Science and Engineering

IIT KHARAGPUR

Introduction

- 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



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

9/3/2017



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Relational Databases

- 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

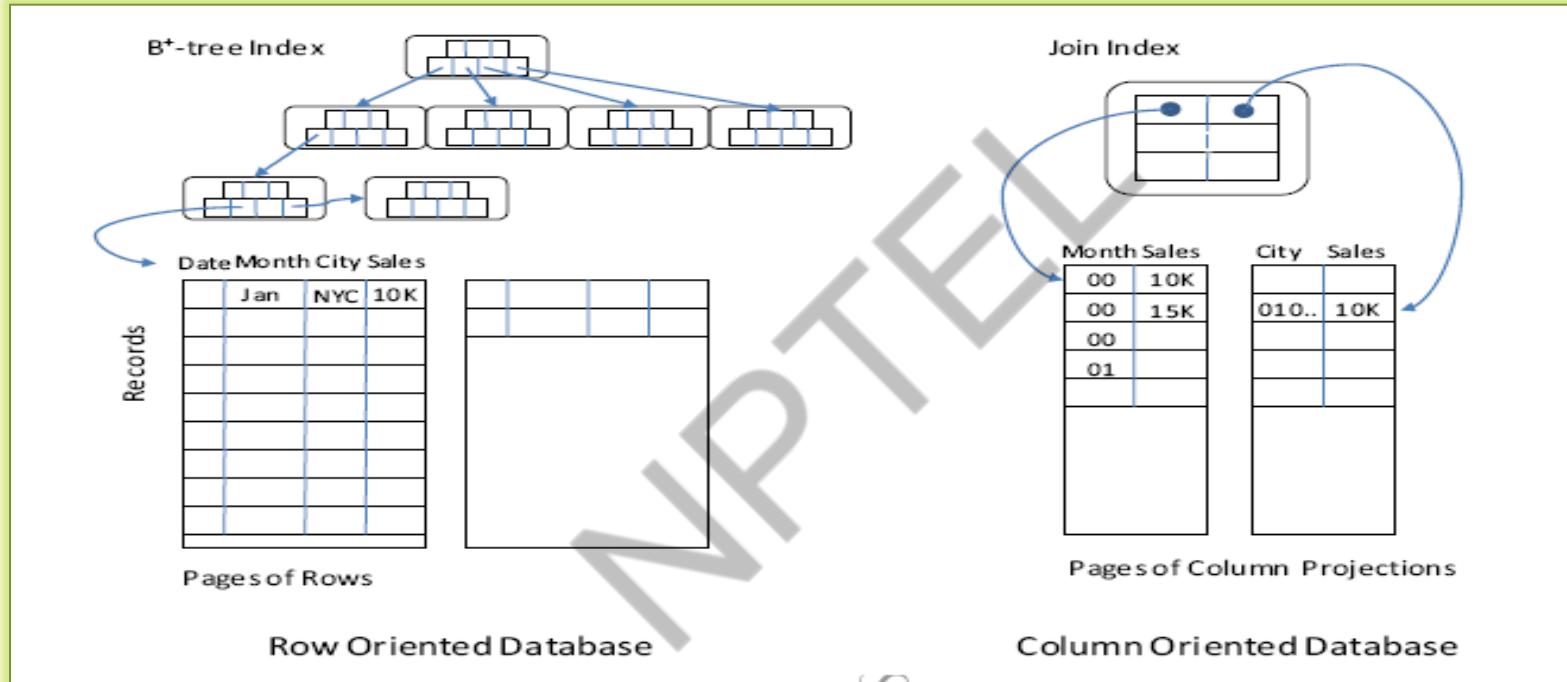
Relational Databases Contd...

- 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

Data Storage Techniques

- 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

Data Storage Techniques Contd...

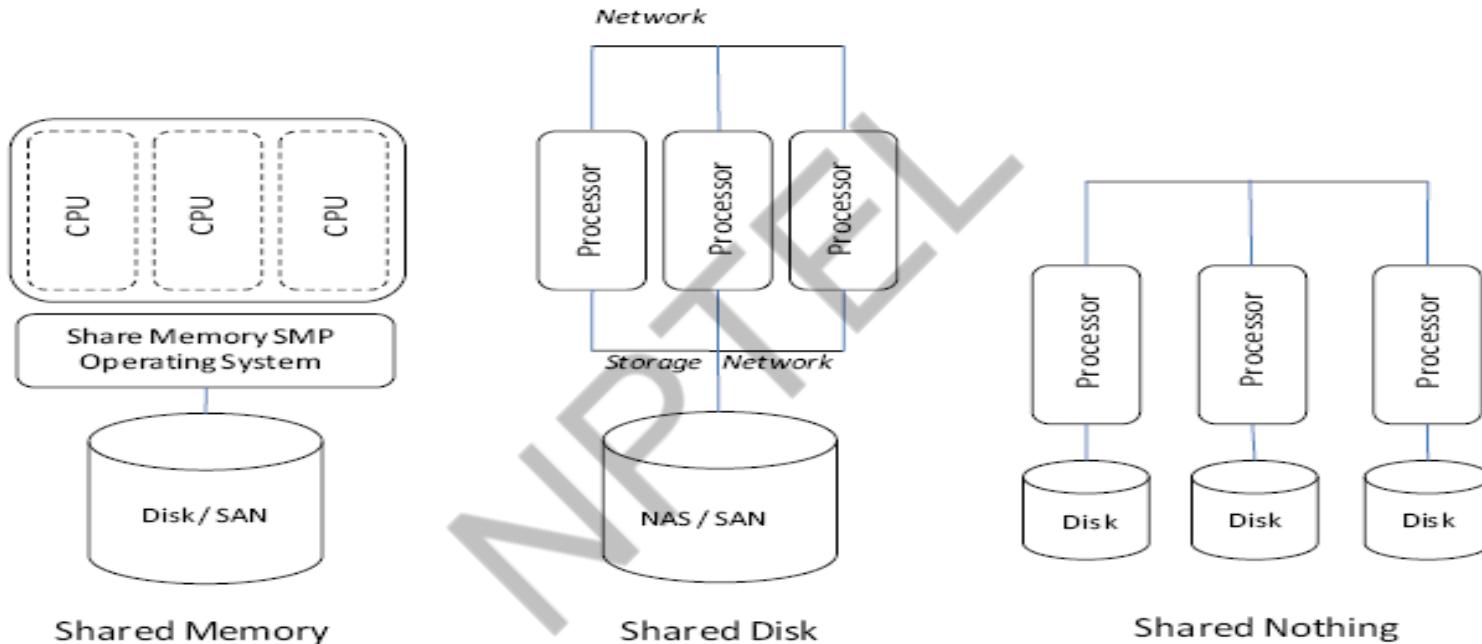


Source: "Enterprise Cloud Computing" by Gautam Shroff

Parallel Database Architectures

- Shared memory
 - Suitable for servers with multiple CPUs
 - Memory address space is shared and managed by a symmetric multi-processing (SMP) operating system
 - SMP:
 - Schedules processes in parallel exploiting all the processors
- Shared nothing
 - Cluster of independent servers each with its own disk space
 - Connected by a network
- Shared disk
 - Hybrid architecture
 - Independent server clusters share storage through high-speed network storage viz. NAS (network attached storage) or SAN (storage area network)
 - Clusters are connected to storage via: standard Ethernet, or faster Fiber Channel or Infiniband connections

Parallel Database Architectures contd...



Source: "Enterprise Cloud Computing" by Gautam Shroff



IIT KHARAGPUR

9/3/2017



NPTEL
ONLINE
CERTIFICATION COURSES

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]

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

9/3/2017

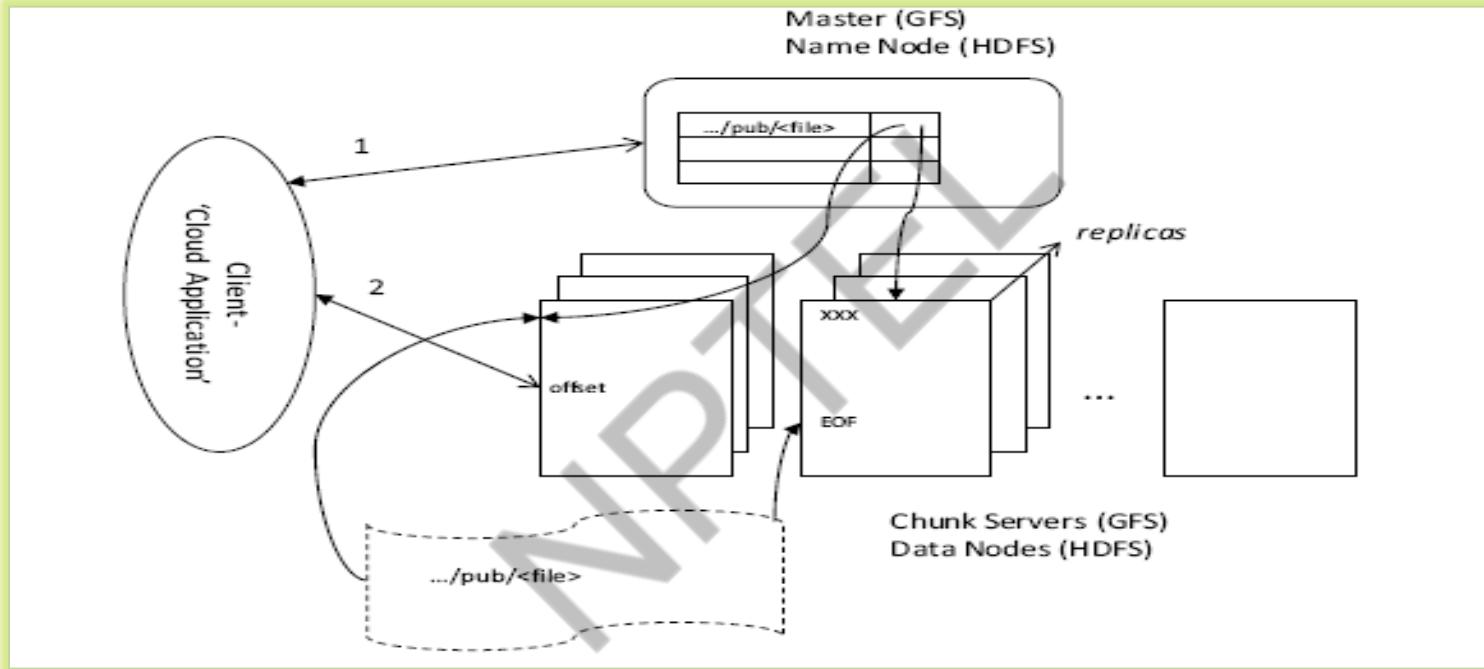


NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Cloud File Systems

- 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

GFS Architecture



Source: "Enterprise Cloud Computing" by Gautam Shroff



IIT KHARAGPUR

9/3/2017



NPTEL
ONLINE
CERTIFICATION COURSES

GFS Architecture Contd...

- Single Master controls file namespace
- Large files are broken up into **chunks** (GFS) or **blocks** (HDFS)
- Typical size of each chunk: 64 MB
 - Stored on commodity (Linux) servers called **Chunk servers** (GFS) or **Data nodes** (HDFS)
 - Replicated **three** times on different:
 - Physical rack
 - Network segment



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

Write/Append Operation in GFS

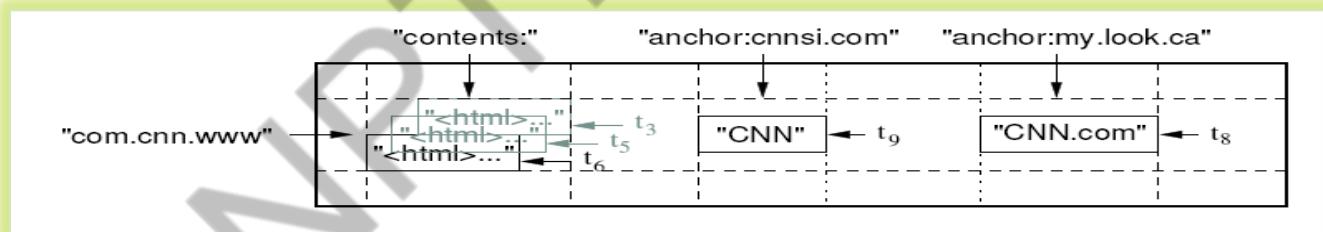
- Client program sends the full path of a file to the **Master** (GFS) or **Name Node** (HDFS)
- Master replies with meta-data for **all** of replicas of the chunk where this data is found.
- Client send data to be appended to all chunk servers
- Chunk server acknowledge the receipt of this data
- Master 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

Fault Tolerance in GFS

- 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

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



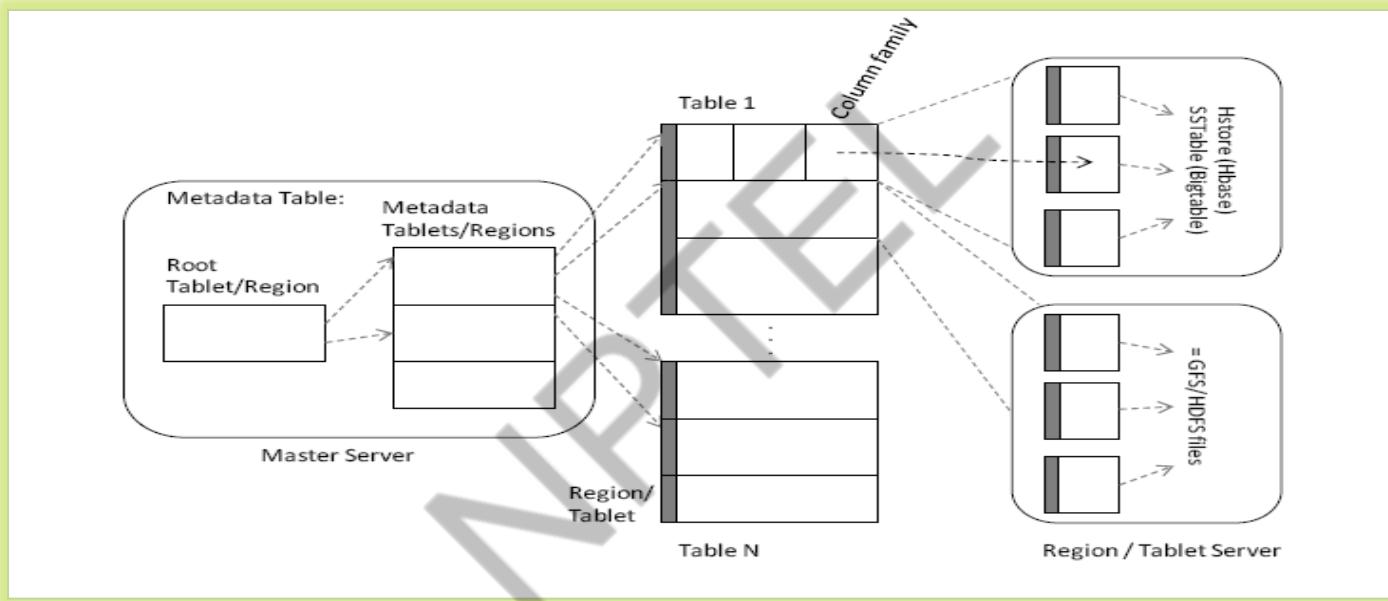
Source: "Enterprise Cloud Computing" by Gautam Shroff

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



BigTable Storage



Source: "Enterprise Cloud Computing" by Gautam Shroff

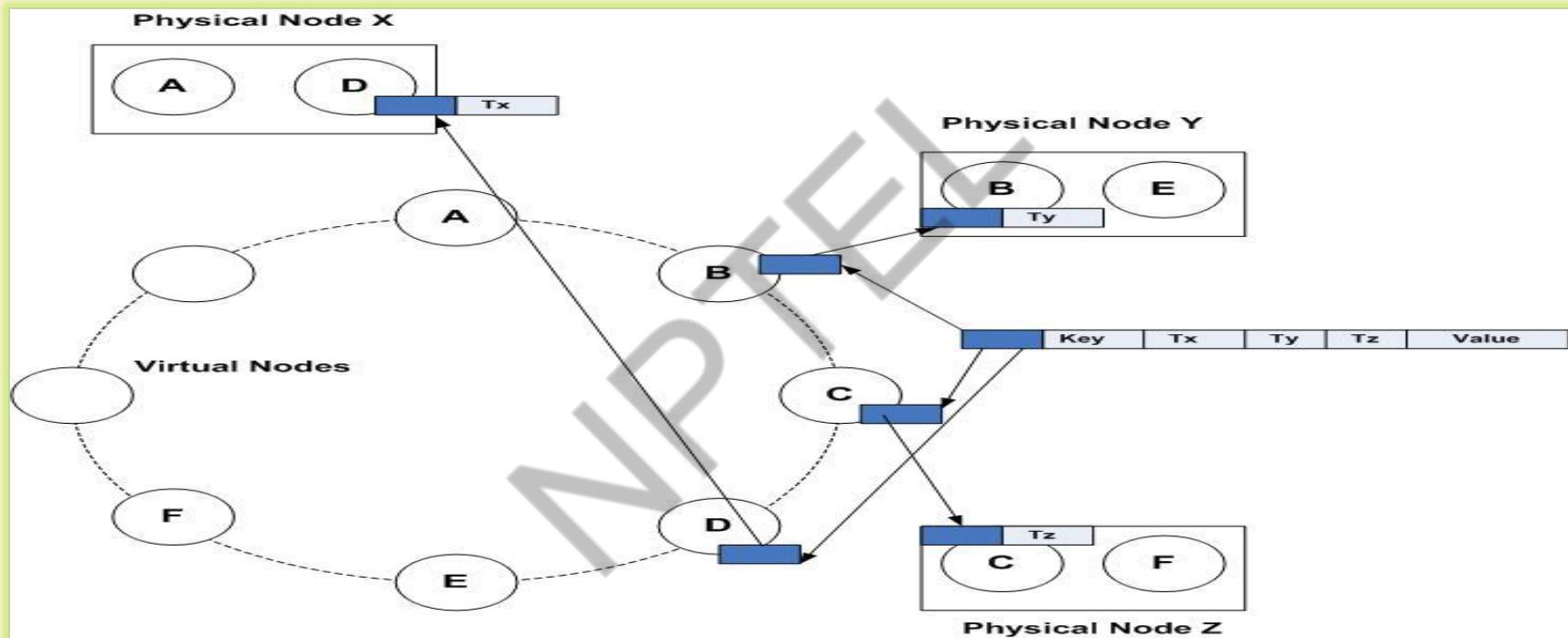
BigTable Storage Contd...

- 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 can simple append

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

Dynamo Architecture



IIT KHARAGPUR

9/3/2017



NPTEL
ONLINE
CERTIFICATION COURSES

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



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

Dynamo Architecture Contd...

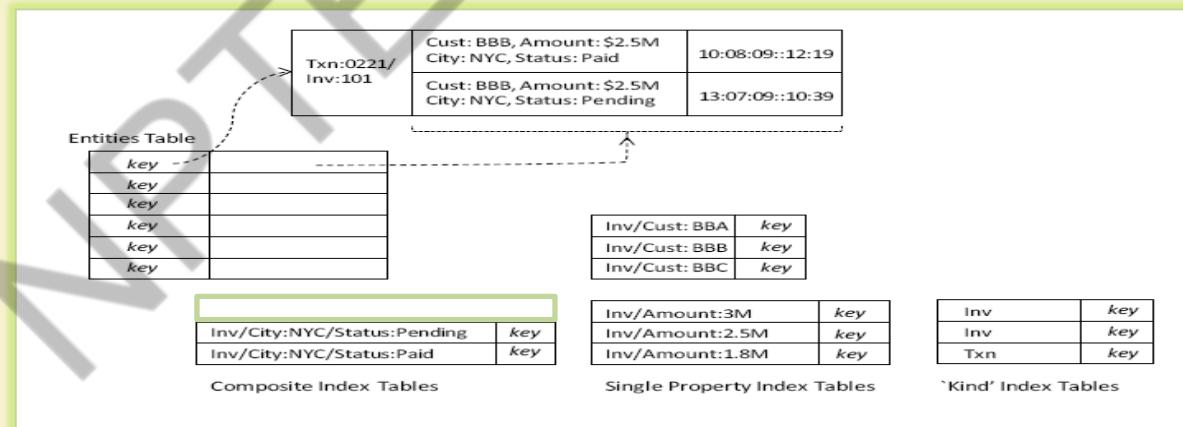
- Distributed object versioning
 - Write creates a new version of an object with its local timestamp incremented
 - Timestamp:
 - Captures history of updates
 - Versions that are superseded by later versions (having larger vector timestamp) are discarded
 - If multiple write operations on same object occurs at the same time, all versions will be maintained and returned to read requests
 - If conflict occurs:
 - Resolution done by application-independent logic

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.

Datastore

- Google and Amazon offer simple transactional <Key, Value> pair database stores
 - Google App Engine's Datastore
 - Amazon' SimpleDB
- All entities (objects) in Datastore reside in one BigTable table
 - Does not exploit column-oriented storage
- **Entities table:** store data as one column family



Source: "Enterprise Cloud Computing" by Gautam Shroff

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

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



Thank You!



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

Cloud Computing : *Introduction to MapReduce*

Prof. Soumya K Ghosh

Department of Computer Science and Engineering
IIT KHARAGPUR

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

9/3/2017



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

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



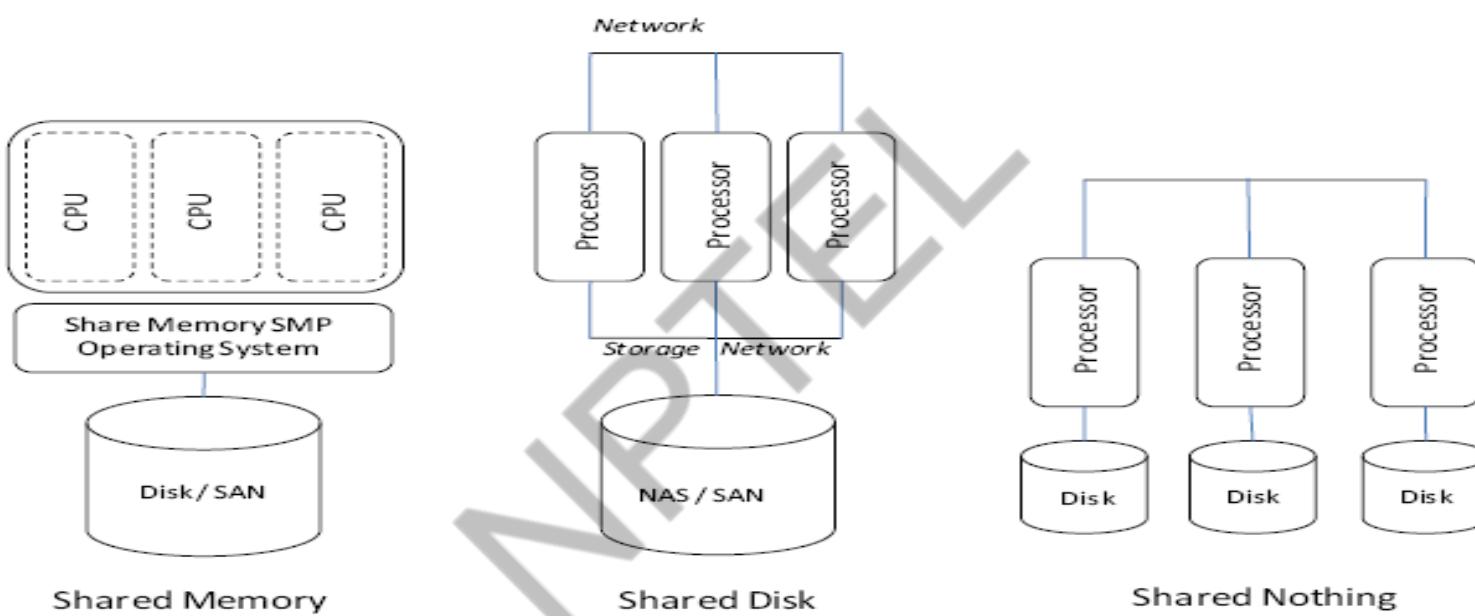
IIT KHARAGPUR

9/3/2017



NPTEL ONLINE
CERTIFICATION COURSES

Parallel Database Architectures



Source: "Enterprise Cloud Computing" by Gautam Shroff



IIT KHARAGPUR

9/3/2017



NPTEL ONLINE
CERTIFICATION COURSES

Parallel Database Architectures Contd...

- Shared memory
 - Suitable for servers with multiple CPUs
 - Memory address space is shared and managed by a symmetric multi-processing (SMP) operating system
 - SMP:
 - Schedules processes in parallel exploiting all the processors
- Shared nothing
 - Cluster of independent servers each with its own disk space
 - Connected by a network
- Shared disk
 - Hybrid architecture
 - Independent server clusters share storage through high-speed network storage viz. NAS (network attached storage) or SAN (storage area network)
 - Clusters are connected to storage via: standard Ethernet, or faster Fiber Channel or Infiniband connections

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:

$$\epsilon = \frac{T}{p T_p}.$$

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

9/3/2017



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

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

9/3/2017



NPTEL ONLINE
CERTIFICATION COURSES

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

9/3/2017



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

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:

$$\epsilon_b = \frac{nmfc}{p \left(\frac{n}{p} mfc + cm + p \frac{m}{p} c \right)} = \frac{nf}{nf + 2p} = \frac{1}{1 + \frac{2p}{nf}}.$$

Illustration Contd...

- Since $p \ll nf$, efficiency of second approach is higher than that of first
- In first 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

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

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:

$$\text{Map: } (k_1, v_1) \rightarrow [(k_2, v_2)].$$

- 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

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



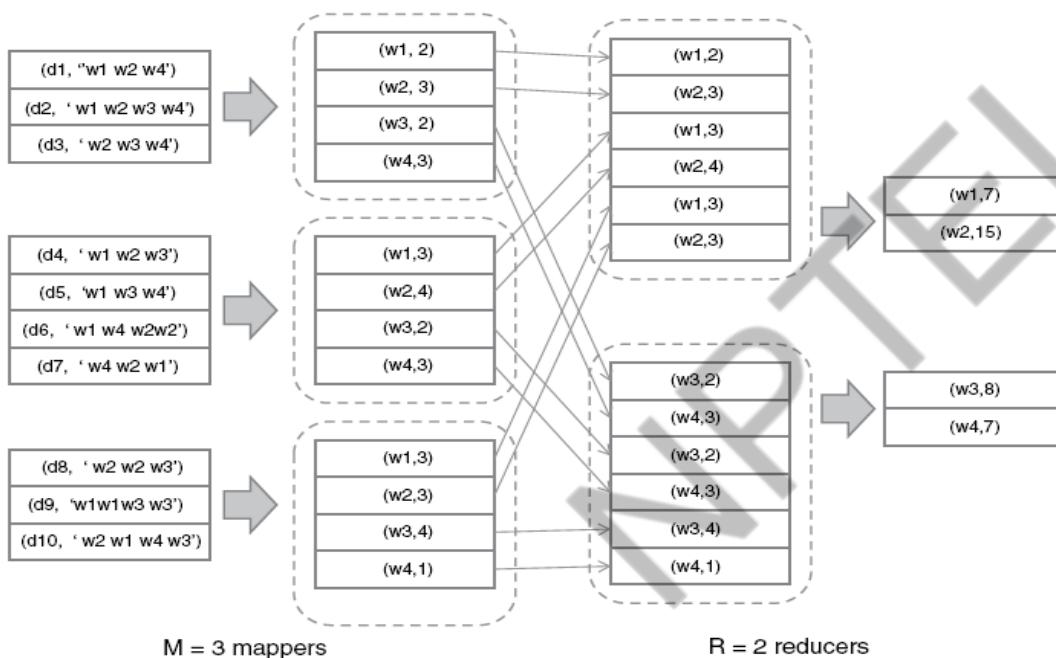
IIT KHARAGPUR

9/3/2017



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

MapReduce: Example



- 3 mappers; 2 reducers
- Map function:

$$(d_k, [w_1 \dots w_n]) \rightarrow [(w_i, c_i)].$$

- Reduce function:

$$(w_i, [c_i]) \rightarrow \left(w_i, \sum_i c_i \right)$$



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

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 = σD
 - Reduce stage:
 - Reducing time = $c_r \sigma D$
 - Data produced as output = $\sigma \mu D$



MapReduce: Efficiency

Contd...

- Considering no overheads in decomposing a task into a map and a reduce stages, we have the following relation:

$$wD = cD + cmD + cr\sigma D + c\sigma\mu D$$

- 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 = $\frac{\sigma D}{P}$

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 = $(\frac{wD}{P} + 2c \frac{\sigma D}{P})$
- Parallel efficiency of the MapReduce implementation:

$$\varepsilon_{MR} = \frac{wD}{P(\frac{wD}{P} + 2c \frac{\sigma D}{P})} = \frac{1}{1 + \frac{2c}{w}\sigma}$$

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: $(d_k, [w_1 \dots w_n]) \rightarrow [(w_i, dk)]$
 - The reduce step groups the pairs by word and creates an index entry for each word: $[(w_i, dk)] \rightarrow (wi, [d_{i1} \dots dim])$
- Relational operations using MapReduce
 - Execute SQL statements (relational joins/group by) on large data sets
 - Advantages over parallel database
 - Large scale
 - Fault-tolerance

Thank You!



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

CLOUD COMPUTING

OPENSTACK:

PROF. SOUMYA K. GHOSH

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
IIT KHARAGPUR

What is OpenStack?

OpenStack is a cloud operating system that controls large pools of compute, storage, and networking resources throughout a datacenter, all managed through a dashboard that gives administrators control while empowering their users to provision resources through a web interface.

Source: OpenStack, <http://www.doc.openstack.org>

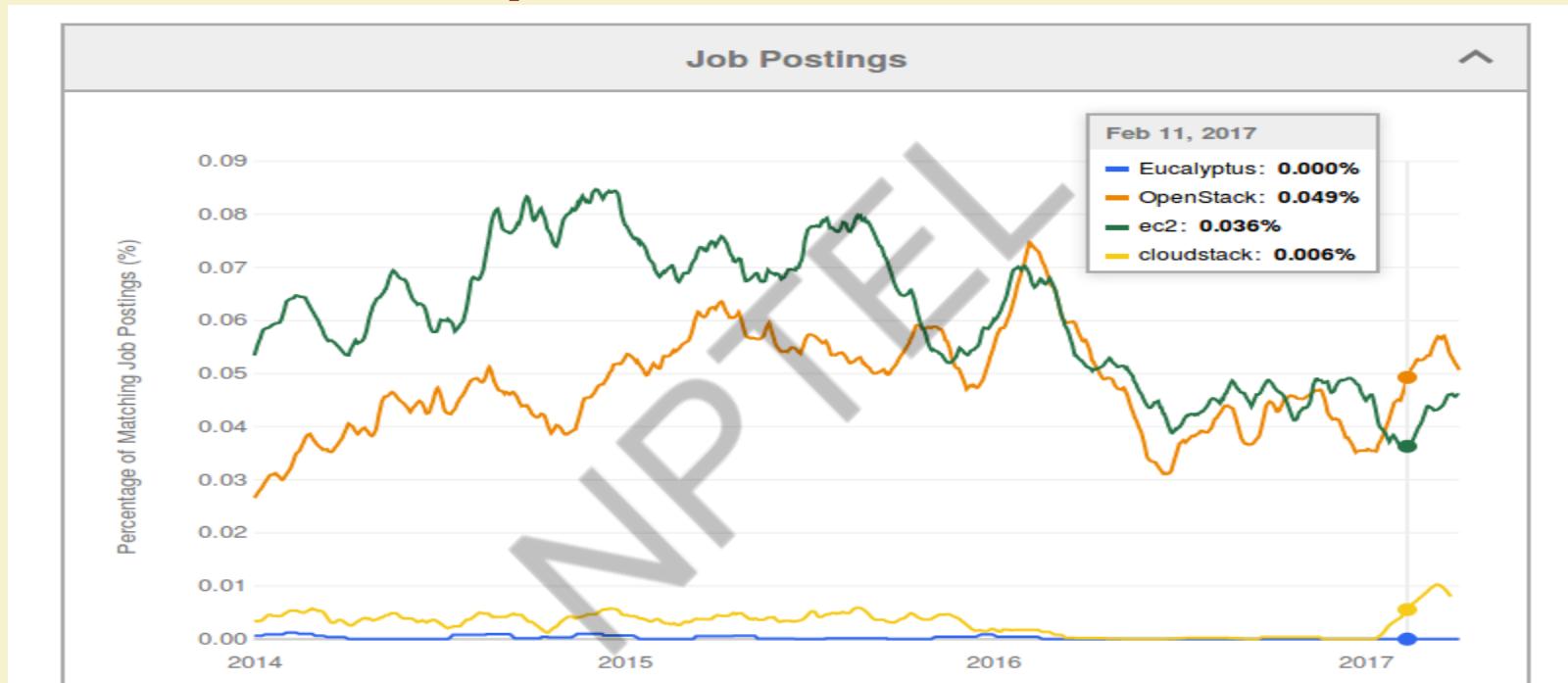


IIT KHARAGPUR



NPTEL
ONLINE
CERTIFICATION COURSES

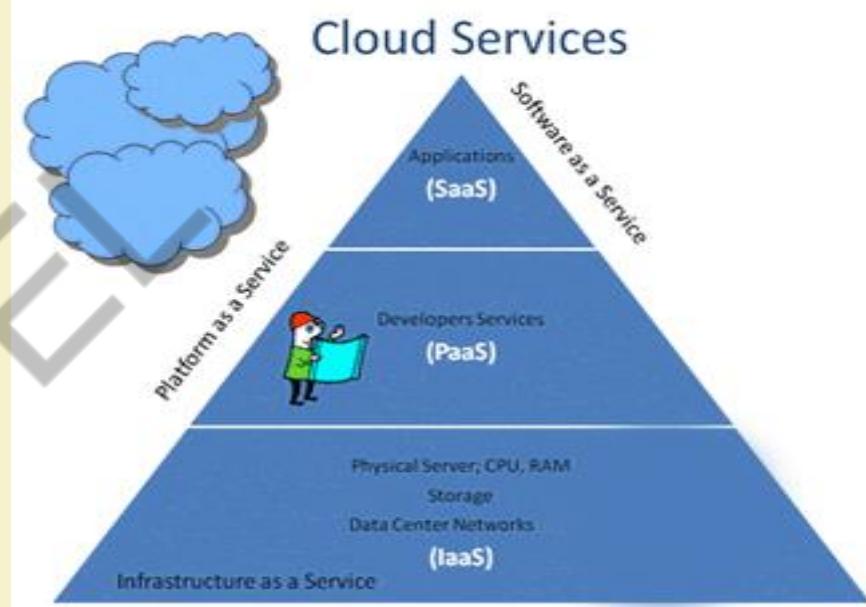
Job Trend for Openstack



Source: <http://www.indeed.com>, Accessed on:July-2017

OpenStack Capability

- Software as Service (SaaS)
 - Browser or Thin Client access
- Platform as Service (PaaS)
 - On top of IaaS e.g. Cloud Foundry
- Infrastructure as Service (IaaS)
 - Provision Compute, Network, Storage



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

OpenStack History

Series	Status	Initial Release Date	Next Phase	EOL Date
Queens	<i>Future</i>	TBD		TBD
Pike	Under Development	TBD		TBD
Ocata	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
Diablo	EOL	2011-09-22		2013-05-06
Cactus	Deprecated	2011-04-15		
Bexar	Deprecated	2011-02-03		
Austin	Deprecated	2010-10-21		

*Started as a collaboration between NASA and Rackspace

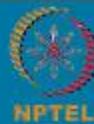
OpenStack Major Components

- Service - Compute
- Project - Nova

Manages the lifecycle of compute instances in an OpenStack environment. Responsibilities include spawning, scheduling and decommissioning of virtual machines on demand.



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

OpenStack Major Components

- 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.



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

OpenStack Major Components

- Service - Object storage
 - Project - Swift
-
- 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.



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

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



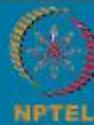
NPTEL
ONLINE
CERTIFICATION COURSES

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

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

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

OpenStack Major Components

- Service - Dashboard
- Project - Horizon
- 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.

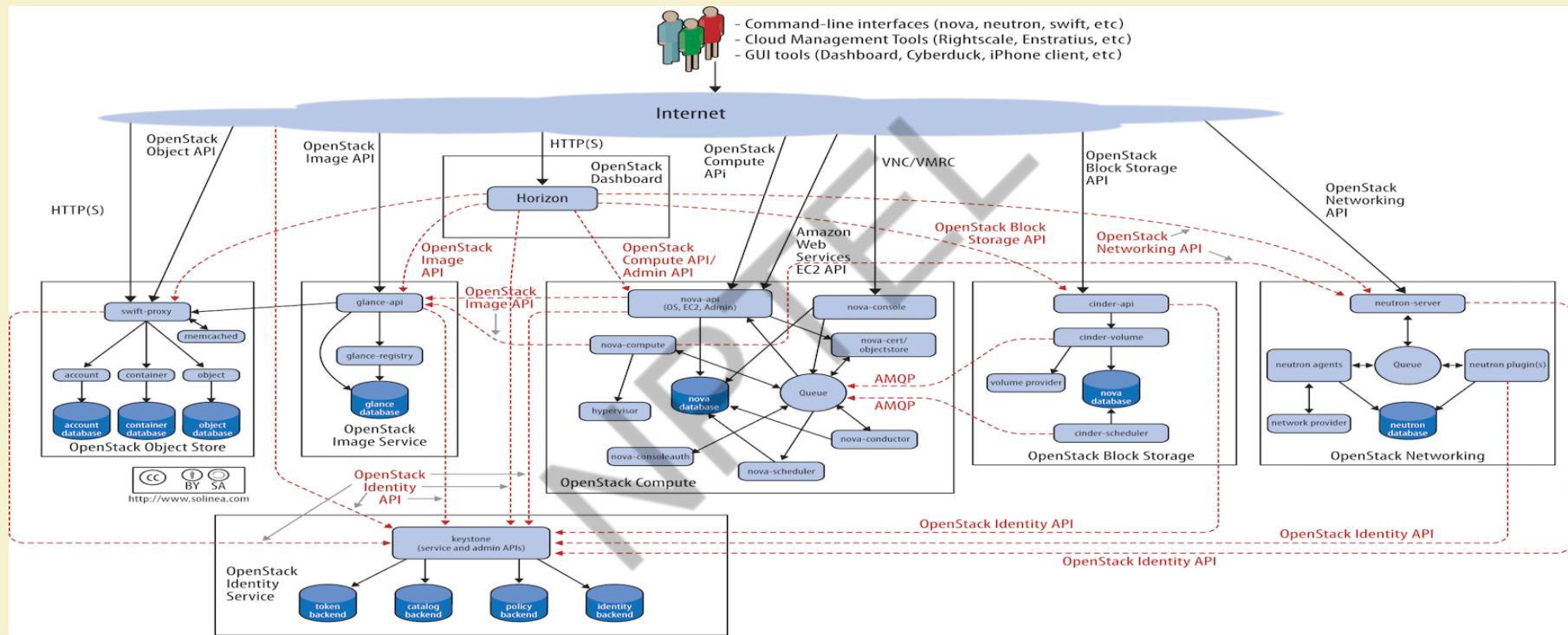


IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

Architecture of Openstack



Openstack Work Flow

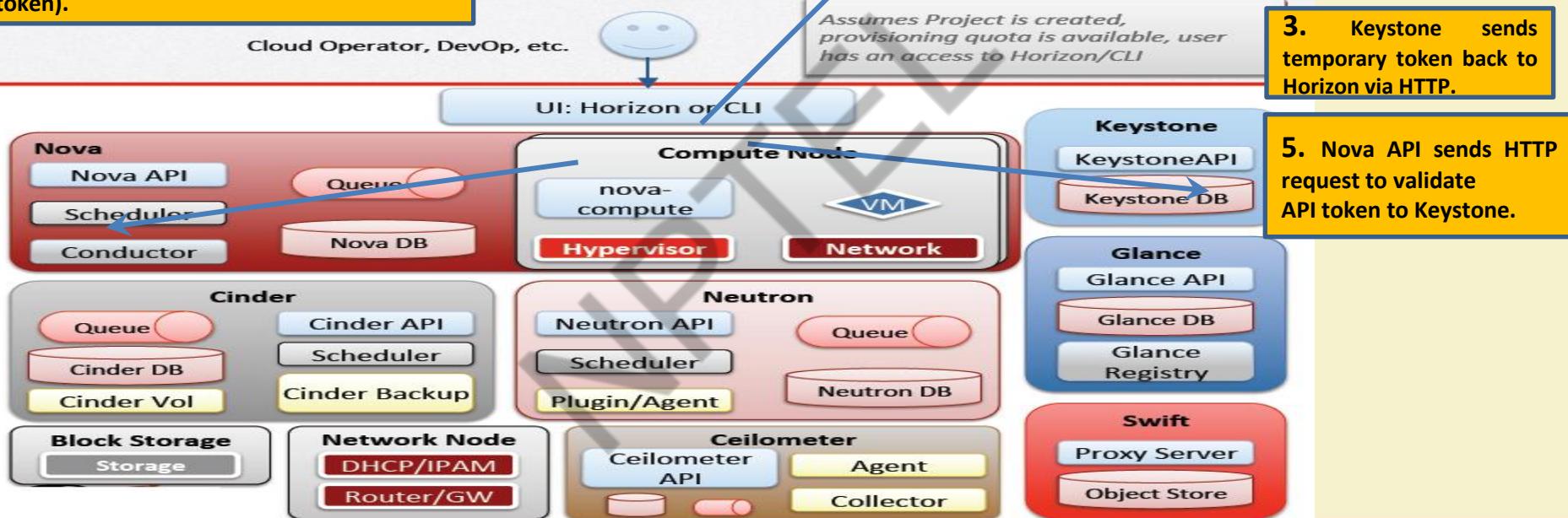
4. Keystone sends temporary token back to Horizon via HTTP. Horizon sends POST request to Nova API(signed with given token).

1. User logs in to UI Specifies VM params: name, flavor, keys, etc. and hits "Create" button

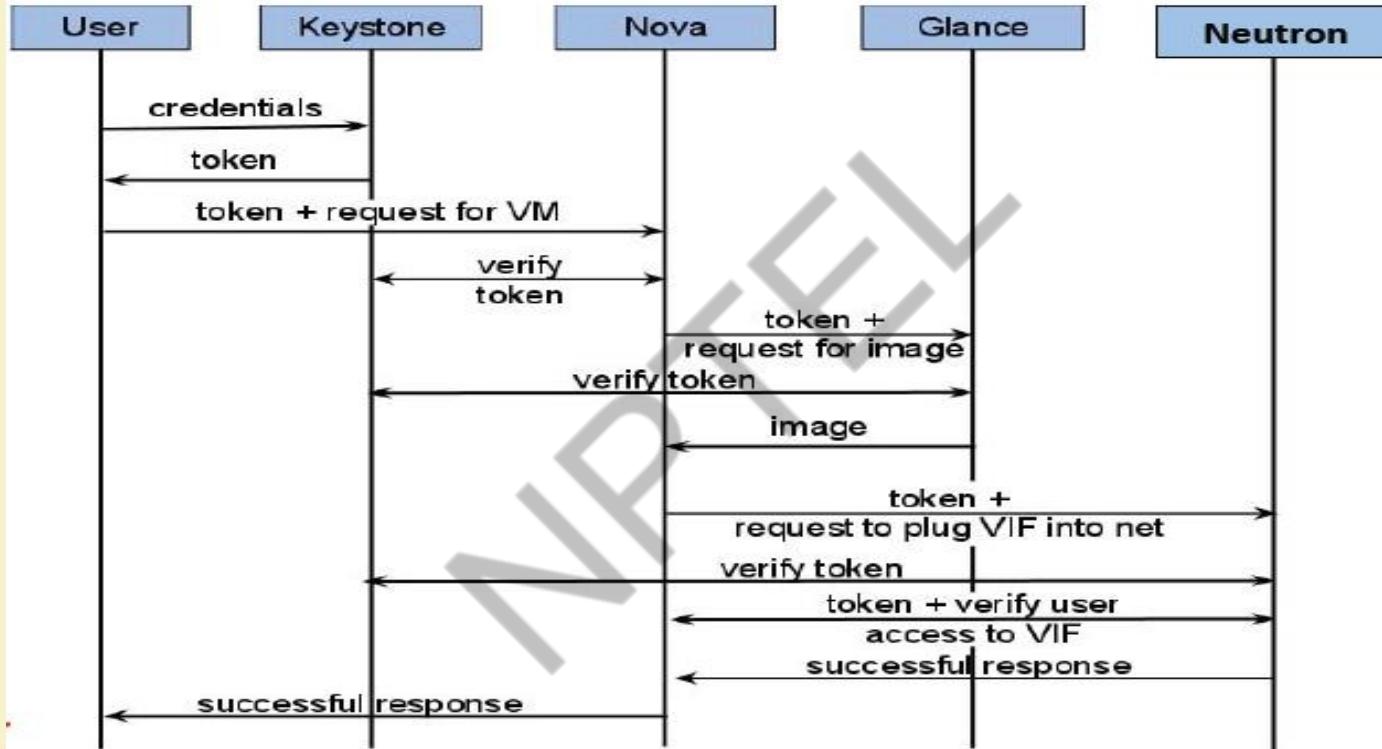
2. Horizon sends HTTP request to Keystone. Auth info is specified in HTTP headers.

3. Keystone sends temporary token back to Horizon via HTTP.

5. Nova API sends HTTP request to validate API token to Keystone.



Auth Token Usage



Provisioning Flow

- Nova API makes rpc.cast to Scheduler. It publishes a short message to scheduler queue with VM info.
- Scheduler picks up the message from MQ.
- Scheduler fetches information about the whole cluster from database, filters, selects compute node and updates DB with its ID
- Scheduler publishes message to the compute queue (based on host ID) to trigger VM provisioning
- Nova Compute gets message from MQ
- Nova Compute makes rpc.call to Nova Conductor for information on VM from DB
- Nova Compute makes a call to Neutron API to provision network for the instance
- Neutron configures IP, gateway, DNS name, L2 connectivity etc.
- It is assumed a volume is already created. Nova Compute contacts Cinder to get volume data. Can also attach volumes after VM is built.

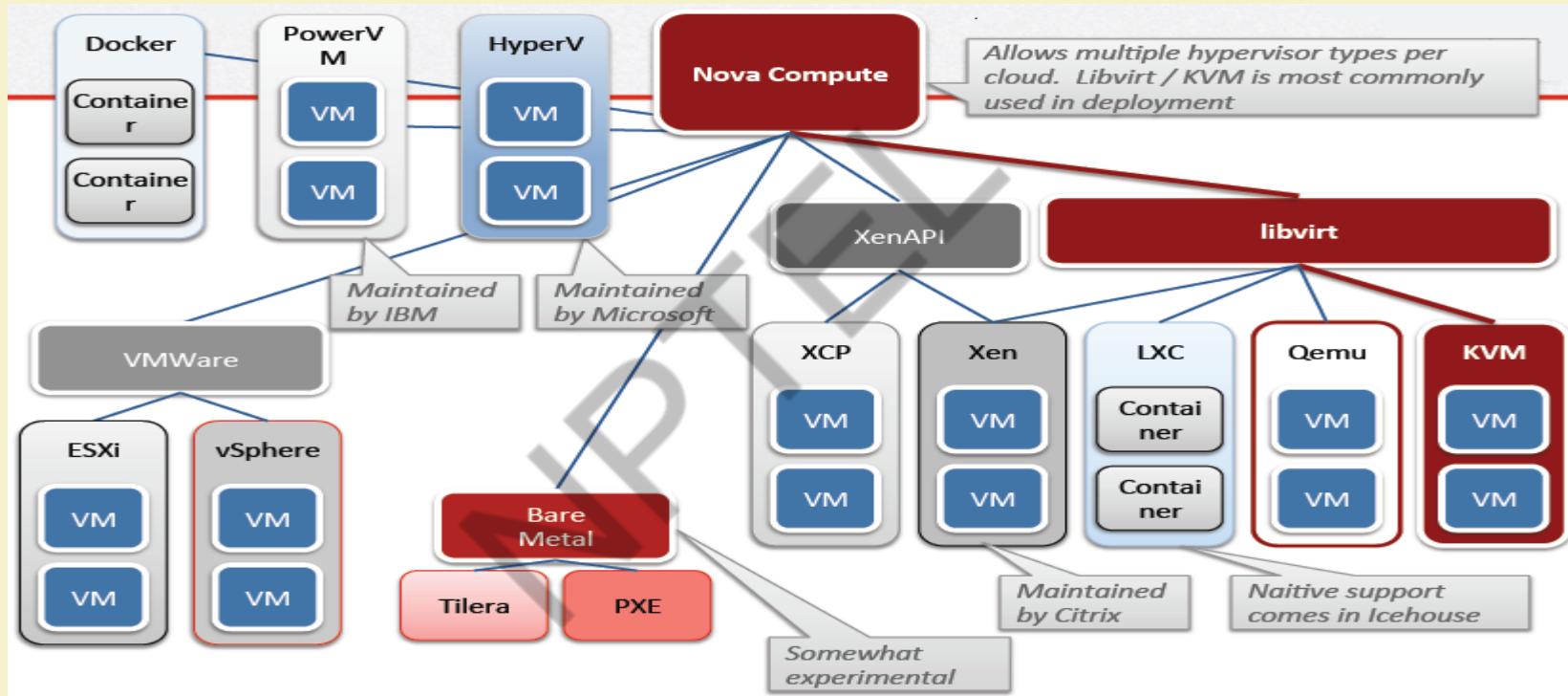


IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Nova Compute Driver

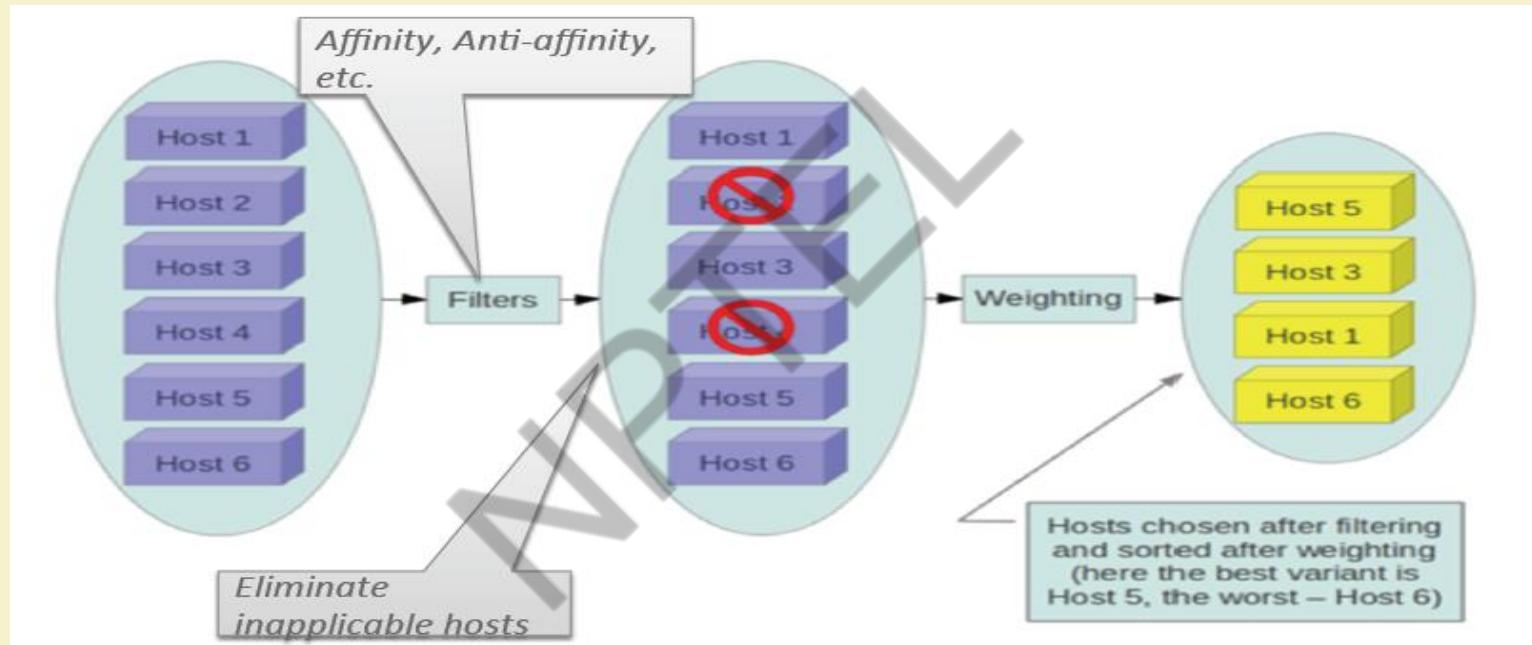


IIT KHARAGPUR



NPTEL
ONLINE
CERTIFICATION COURSES

Nova scheduler filtering

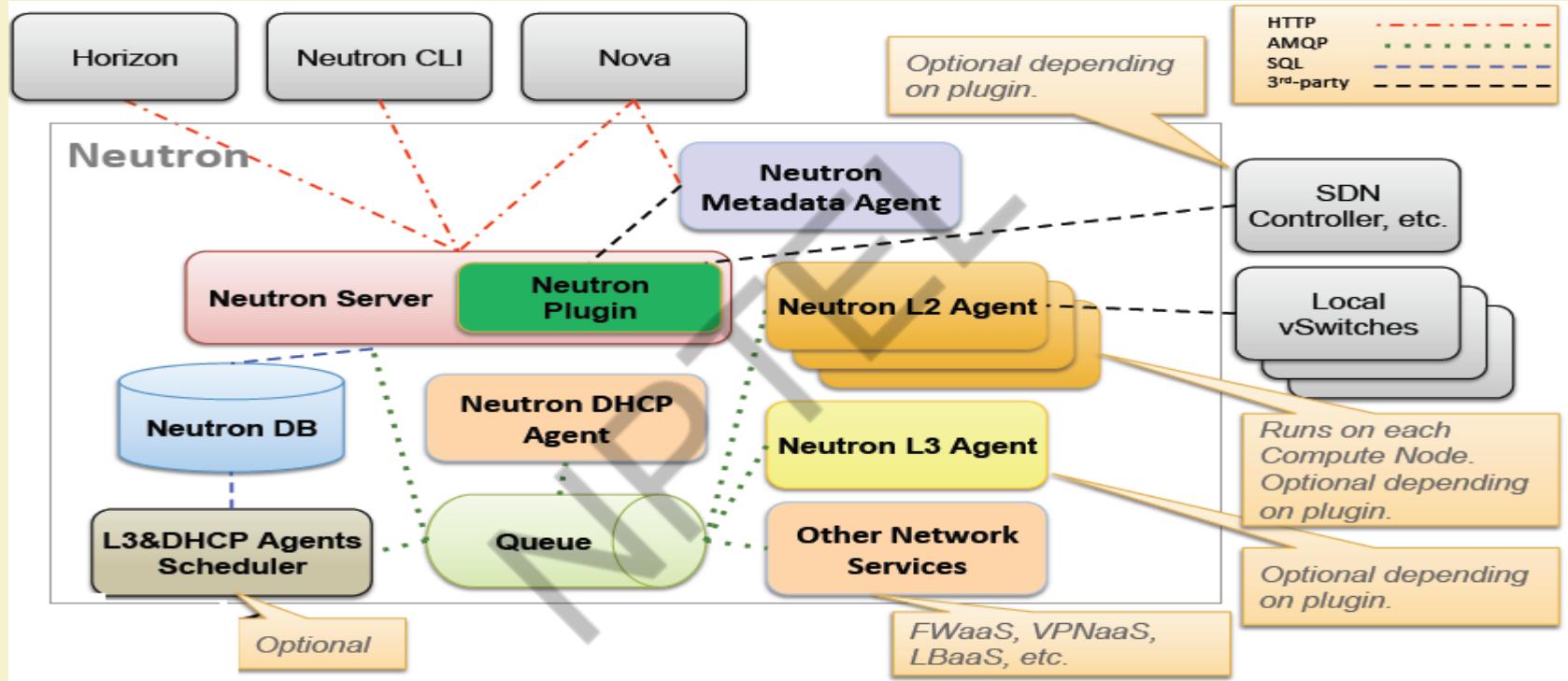


IIT KHARAGPUR

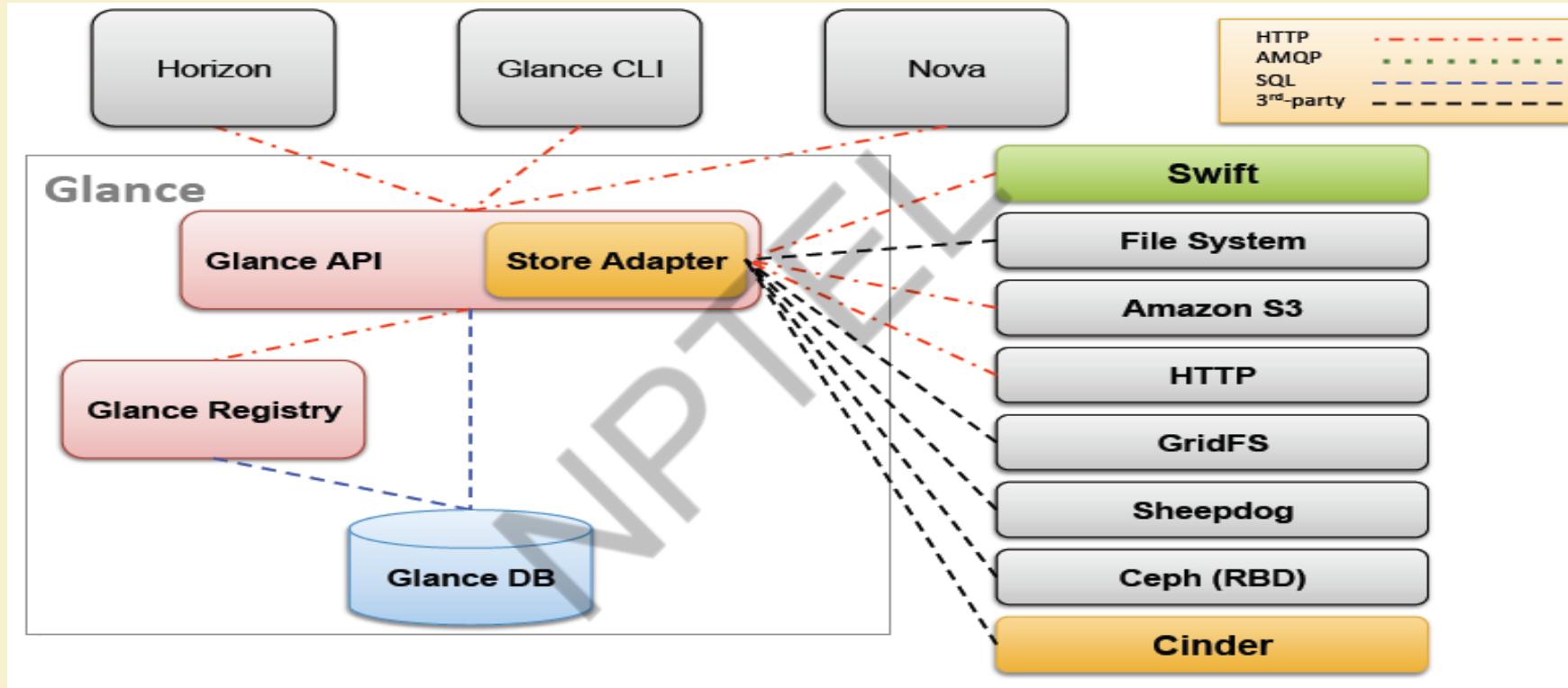


NPTEL
ONLINE
CERTIFICATION COURSES

Neutron Architecture



Glance Architecture

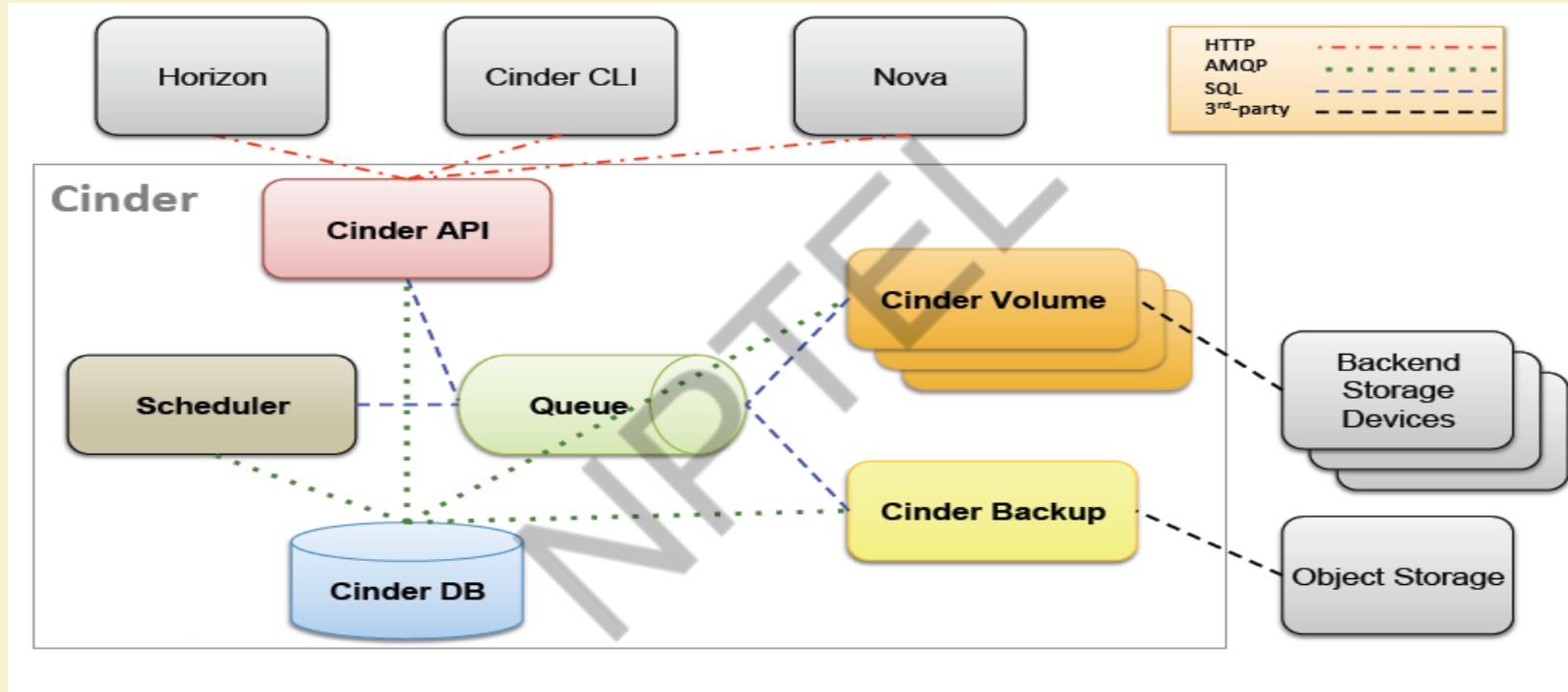


IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

Cinder Architecture

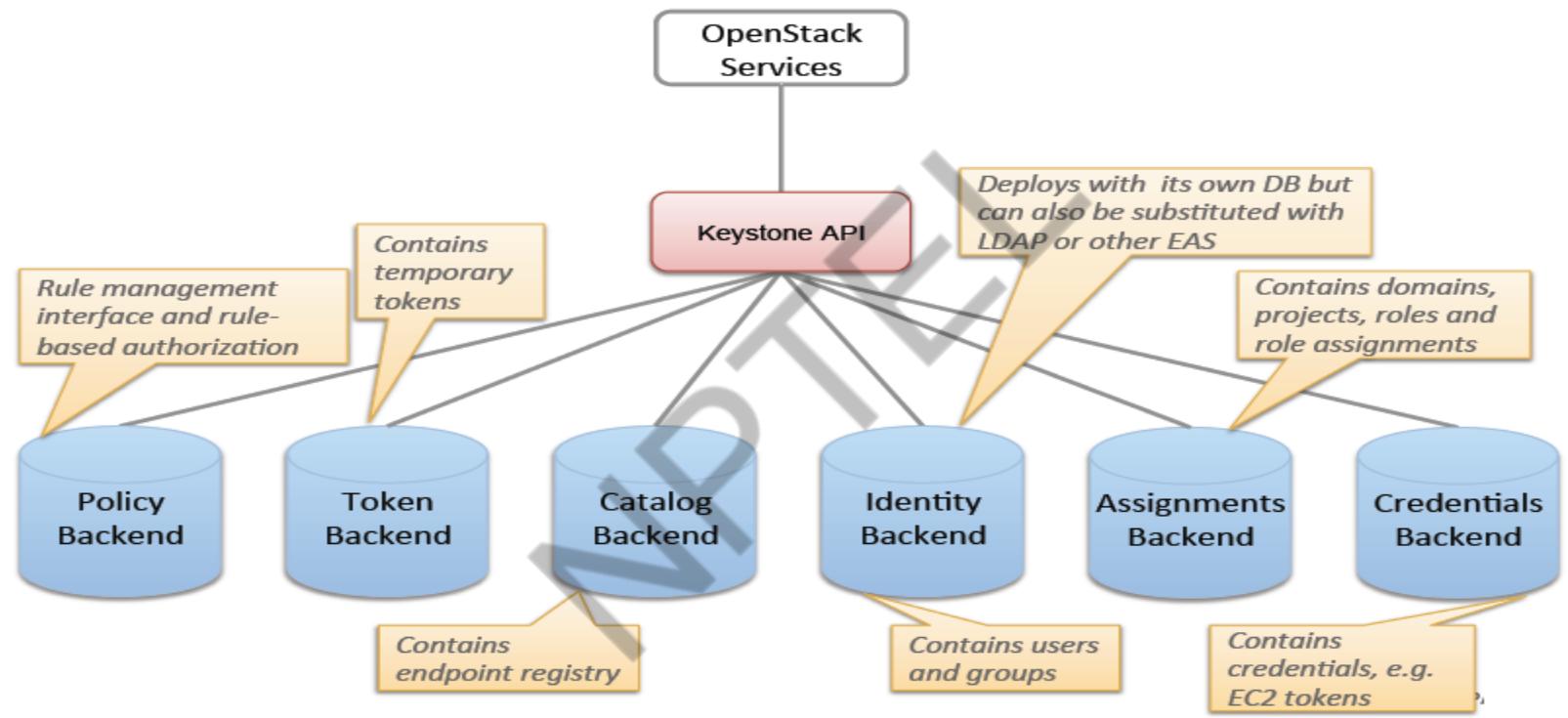


IIT KHARAGPUR



NPTEL
ONLINE
CERTIFICATION COURSES

Keystone Architecture



OpenStack Storage Concepts

- **Ephemeral storage:**
 - Persists until VM is terminated
 - Accessible from within VM as local file system
 - Used to run operating system and/or scratch space
 - Managed by Nova
- **Block storage:**
 - Persists until specifically deleted by user
 - Accessible from within VM as a block device (e.g. /dev/vdc)
 - Used to add additional persistent storage to VM and/or run operating system
 - Managed by Cinder
- **Object storage:**
 - Persists until specifically deleted by user
 - Accessible from anywhere
 - Used to store files, including VM images
 - Managed by Swift



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

Summary

- 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



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

Thank You!

NPTEL



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

CLOUD COMPUTING

Private Cloud Implementation using OpenStack

PROF. SOUMYA K. GHOSH

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

IIT KHARAGPUR

Overview

- *Meghamala @IITKgp* on OpenStack Cloud
- VM Creation
- Accessing VM by User
- VM Termination



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES



Meghamala - a one stop solution to your computational needs.

The IIT Kharagpur cloud gives you compute and storage with one click.

[Know more](#)

Welcome to Meghamala!

Meghamala is an initiative by the Indian Institute of Technology, Kharagpur to provide on demand computational and storage resources to the institute research community. It is built using OpenStack Cloud Computing platform.

Meghamala has been set up in the Computer and Informatics Centre, IIT Kharagpur. The hardware of the system includes :

- Blade servers
- SAN Storage
- NAS

Please visit the various sections of this website to know more about Meghamala.

Latest News



MAR 23, 2016

MegHadoop

MegHadoop, a Hadoop cluster on Meghamala is up and available for use.



AUG 12, 2015

MeghaData

MeghaData, a data storage service is under beta testing.



APR 25, 2015

Inauguration

Inauguration and Workshop on Meghamala was carried out on 30th April 2015.

Services offered by Meghamala

Meghamala was conceptualized to address the computational needs of the research community at IIT Kharagpur.

To meet these demands, Meghamala offers the following services :

- **VMS4U -- Compute Nodes**

Provision a virtual machine on demand and use it as a desktop or run your workload on it. The following virtual machine configurations are available :

- **IITKG_P_regular**

- 2 VCPUs
 - 4 GB RAM
 - 45 GB ephemeral storage

- **IITKG_P_large**

- 4 VCPUs
 - 8 GB RAM
 - 45 GB ephemeral storage

- **IITKG_P_xLarge**

- 8 VCPUs
 - 16 GB RAM
 - 60 GB ephemeral storage

The virtual machines can have the following guest operating systems.

- Ubuntu 14.04
 - Centos 7
 - Fedora 20

- **Storage on the House**

- Persistent storage provided on request

[Click here to request for a VM](#)

- **MegHadoop**

- Hadoop cluster running on Meghamala

Latest News



MAR 23, 2016

MegHadoop

MegHadoop, a Hadoop cluster on Meghamala is up and available for use.



AUG 12, 2015

MeghaData

MeghaData, a data storage service is under beta testing.



APR 25, 2015

Inauguration

Inauguration and Workshop on Meghamala was carried out on 30th April 2015.



MAR 17, 2015

Installation Complete

Hardware and software installed. Testing in progress.



MAR 13, 2015

GUI on Meghamala

VM images with GUI have been created on Meghamala.

VMs4U - Request form

Name of faculty

Department

Designation

Phone/Mobile no.

E-mail

Purpose

Preferred VM Name

VM Type

- IITKGP_regular IITKGP_large IITKGP_xlarge

Number of VMs

1

Operating system

Ubuntu 14.04

Persistent storage of 20 GB required

- Yes No

VM required till (max 60 days)

Enter the code above here

Can't read the image? click [here](#) to refresh

Please note that the VMs should be used only for academic purposes. Neither the Meghamala team nor IIT Kharagpur is responsible for the contents of your VMs. It is important to highlight that the presence of inappropriate material may lead to immediate termination of the VM(s).

Steps to follow



Fill out this form.
Fill out the form on the left and click on Submit.



Get hard copy signed.
Print the generated PDF and sign it. You may save a copy for future reference.



Submit signed hard copy.
Submit the signed hard copy to the professor-in-charge, Meghamala.

Submit Query

Meghamala team

- Students

Current Members

- Shubham Jain, 4th year Dual Degree (Computer Science and Engineering)
- Shreyans Pagariya, 4th year Dual Degree (Computer Science and Engineering)
- Arindam Roy, PhD Scholar (Advanced Technology Development Center)
- Rajesh Basak, PhD Scholar (Computer Science and Engineering)
- Debopriyo Banerjee, PhD Scholar (Computer Science and Engineering)
- Chandan Misra, PhD Scholar (Advanced Technology Development Center)

Past Members

- Harshit Gupta, Dual Degree (Computer Science and Engineering)
- Nikhil Agrawal, Dual Degree (Computer Science and Engineering)
- Ashish Kale, M.Tech (Computer Science and Engineering)
- Major Sujeeet Deshmukh, M.Tech. (Information Technology)

- CIC Engineers

- Alokes Chattopadhyay
- Alok Baran Das

- Faculty

- Soumya K. Ghosh (Dept. of Computer Science and Engineering)
- Shamik Sural (Dept. of Computer Science and Engineering)

We plan to add more team members as time progresses. After all, the key aim of the project remains to make people learn.

Latest News



MAR 23, 2016

MegHadoop

MegHadoop, a Hadoop cluster on Meghamala is up and available for use.



AUG 12, 2015

MeghaData

MeghaData, a data storage service is under beta testing.



APR 25, 2015

Inauguration

Inauguration and Workshop on Meghamala was carried out on 30th April 2015.



MAR 17, 2015

Installation Complete

Hardware and software installed. Testing in progress.



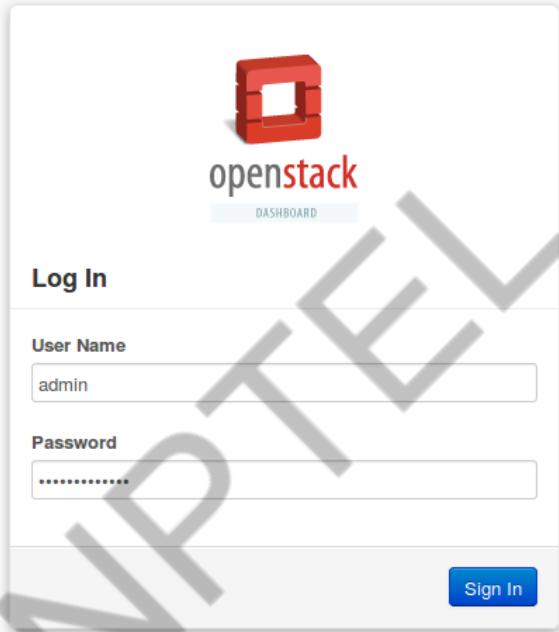
MAR 13, 2015

GUI on Meghamala

VM images with GUI have been created on Meghamala.

Meghamala - IITKgp Cloud

(using OpenStack)



Horizon Login Page

Project

Compute

Overview

Instances

Volumes

Images

Access & Security

Network

Object Store

Orchestration

Admin

Overview

Usage Summary

Select a period of time to query its usage:

From: To: Submit

The date should be in YYYY-mm-dd format.

Active Instances: 30 Active RAM: 304GB This Period's VCPU-Hours: 679.47 This Period's GB-Hours: 64662.52

Usage

[Download CSV Summary](#)

Project Name	VCPUs	Disk	RAM	VCPUs Hours	Disk GB Hours
admin	128	2855	304GB	679.47	64662.52

Displaying 1 item

Project

Compute

Overview

Instances

Volumes

Images

Access & Security

Network

Object Store

Orchestration

Admin

Overview

Limit Summary



Instances
Used Inf of No Limit



VCPUs
Used Inf of No Limit



RAM
Used Inf.0PB of No Limit



Floating IPs
Used 43 of 250



Security Groups
Used 1 of No Limit



Volumes
Used 21 of 200



Volume Storage
Used 3.0TB of 3.7TB

Usage Summary

Select a period of time to query its usage:

From: To:

The date should be in YYYY-mm-dd format.

Active Instances: 30 Active RAM: 304GB This Period's VCPU-Hours: 680.30 This Period's GB-Hours: 64741.88

Usage

Instance Name	VCPUs	Disk	RAM	Uptime
nik_windows	2	45	4GB	2 years, 2 months
Ravi_Teja_2	8	160	16GB	1 year, 4 months
4	...	16GB	1 year, 11 months	
5	4GB	16GB	1 year, 10 months	

Graphical representation of resource usage

Project

Compute

Overview

Instances

Volumes

Images

Access & Security

Network

Object Store

Orchestration

Admin

Instances

Instances

Instance Name

Filter



Filter

+ Launch Instance

Soft Reboot Instances

Terminate Instances

<input type="checkbox"/>	Instance Name	Image Name	IP Address	Size	Key Pair	Status	Availability Zone	Task	Power State	Uptime	Actions	
<input type="checkbox"/>	ccTest	Centos_7_GUI	192.164.0.1 10.4.0.1	IITKGP_regular 4GB RAM 2 VCPU 45.0GB Disk	-	Active	nova	None	Running	2 months, 2 weeks	<button>Create Snapshot</button> <button>More</button>	
<input type="checkbox"/>	TestDiskPartition	Ubuntu_14_04_x2go_45G	192.164.0.2 10.4.0.2	IITKGP_regular 4GB RAM 2 VCPU 45.0GB Disk	-	Active	nova	None	Running	3 months, 2 weeks	<button>Create Snapshot</button> <button>More</button>	
<input type="checkbox"/>	centosForSify	CentOS_6.5_GUI	192.164.0.3 10.4.0.3	IITKGP_regular 4GB RAM 2 VCPU 45.0GB Disk	-	Shutoff	nova	None	Shutdown	7 months	<button>Start Instance</button> <button>More</button>	
<input type="checkbox"/>	CL1_R_SERVER1	Ubuntu_New_X2Go	192.164.0.4 10.4.0.4	IITKGP_xxlarge 32GB RAM 8 VCPU 60.0GB Disk	-	Active	nova	None	Running	9 months, 1 week	<button>Create Snapshot</button> <button>More</button>	
<input type="checkbox"/>	Harshit_Utkarsh_LARGE	Ubuntu_14_04_x2go_60G	192.164.0.5 10.4.0.5	IITKGP_xlarge 16GB RAM 8 VCPU 60.0GB Disk	-	Active	nova	None	Running	1 year, 2 months	<button>Create Snapshot</button> <button>More</button>	
<input type="checkbox"/>	cc16_test1	Ubuntu_14_04_x2go_45G	192.164.0.6 10.4.0.6	IITKGP_regular 4GB RAM 2 VCPU 45.0GB Disk	-	Shutoff	nova	None	Shutdown	1 year, 4 months	<button>Start Instance</button> <button>More</button>	
<input type="checkbox"/>	MeghdooPNewMaster	CentOS_6.5_GUI	192.164.0.7 10.4.0.7	IITKGP_MeghdooP_Bigger 48GB RAM 8 VCPU 600.0GB Disk	-	Active	nova	None	Running	1 year, 4 months	<button>Create Snapshot</button> <button>More</button>	
<input type="checkbox"/>	MeghdooP_18	CentOS_6.5_GUI	192.164.0.8 10.4.0.8	MeghdooP_new 8GB RAM 4 VCPU 90.0GB Disk	-	Active	nova	None	Running	1 year, 5 months	<button>Create Snapshot</button> <button>More</button>	
<input type="checkbox"/>	MeghdooP_19			192.164.0.9 10.4.0.9	w 8GB 90.0GB	-	Active	nova	None	Running	1 year, 5 months	<button>Create Snapshot</button> <button>More</button>

Details of Instances

Project

Compute

Overview

Instances

Volumes

Images

Access & Security

Network

Object Store

Orchestration

Admin

Volumes & Snapshots

[Volumes](#) [Volume Snapshots](#)

Volumes

Filter



Filter

[+ Create Volume](#)[Delete Volumes](#)

<input type="checkbox"/>	Name	Description	Size	Status	Type	Attached To	Availability Zone	Actions
<input type="checkbox"/>	checkcentos_vol	created on 30-12-2016 for downloading...	200GB	In-Use	-	Attached to CheckCentos on /dev/vdb	nova	Edit Volume More
<input type="checkbox"/>	CL1_R_VOL1		100GB	In-Use	-	Attached to CL1_R_SERVER1 on /dev/vdc	nova	Edit Volume More
<input type="checkbox"/>	cc16		5GB	In-Use	-	Attached to cc16_test1 on /dev/vdb	nova	Edit Volume More
<input type="checkbox"/>	cc16_test1		2GB	Available	-		nova	Edit Volume More
<input type="checkbox"/>	DebopriyoTestTwitter_vol	Volume reduced to 1TB from 2TB	1024GB	In-Use	-	Attached to DebopriyoTwitterTest on /dev/vdb	nova	Edit Volume More
<input type="checkbox"/>	Meghadoop_20_Vol	-	110GB	In-Use	-	Attached to Meghadoop_20 on /dev/vdb	nova	Edit Volume More
<input type="checkbox"/>	Meghadoop_19_Vol	-	110GB	In-Use	-	Attached to Meghadoop_19 on /dev/vdb	nova	Edit Volume More
<input type="checkbox"/>	Meghadoop_18_Vol	-	110GB	In-Use	-	Attached to Meghadoop_18 on /dev/vdb	nova	Edit Volume More
<input type="checkbox"/>	Meghadoop_17_Vol	-	110GB	In-Use	-	Attached to Meghadoop_17 on /dev/vdb	nova	Edit Volume More
<input type="checkbox"/>	Meghadoop_16_Vol	-	110GB	In-Use	-	Attached to Meghadoop_16 on /dev/vdb	nova	Edit Volume More

Cinder- details of Volumes

Project

Compute

Overview

Instances

Volumes

Images

Access & Security

Network

Object Store

Orchestration

Admin

Images

Images

[Project \(16\)](#) [Shared with Me \(0\)](#) [Public \(14\)](#)[+ Create Image](#)[Delete Images](#)

<input type="checkbox"/>	Image Name	Type	Status	Public	Protected	Format	Actions
<input type="checkbox"/>	Meghadoop_snapshot_ready	Snapshot	Active	Yes	No	QCOW2	Launch More
<input type="checkbox"/>	CentOS_6.5_GUI	Image	Active	Yes	No	QCOW2	Launch More
<input type="checkbox"/>	Stacksync1_10_4_2_30_01092015	Snapshot	Active	No	No	QCOW2	Launch More
<input type="checkbox"/>	stacksync_working	Snapshot	Active	No	No	QCOW2	Launch More
<input type="checkbox"/>	Ubuntu_14_04_x2go_60G	Image	Active	Yes	No	QCOW2	Launch More
<input type="checkbox"/>	Ubuntu_14_04_x2go_45G	Image	Active	Yes	No	QCOW2	Launch More
<input type="checkbox"/>	Ubuntu_14_04_x2go_20G	Image	Active	Yes	No	QCOW2	Launch More
<input type="checkbox"/>	Ubuntu_New_X2Go	Image	Active	Yes	No	QCOW2	Launch More
<input type="checkbox"/>	Windows_7_x64	Image	Active	Yes	No	QCOW2	Launch More
<input type="checkbox"/>	Fedora_20_GUI	Image	Active	Yes	No	QCOW2	Launch More
<input type="checkbox"/>	Centos_7_GUI	Image	Active	Yes	No	QCOW2	Launch More

Glance- Overview of available images in Meghamala cloud

Project

Compute

Overview

Instances

Volumes

Images

Access & Security

Network

Object Store

Orchestration

Admin

Manage Security Group Rules: default

Security Group Rules

[+ Add Rule](#)[Delete Rules](#)

<input type="checkbox"/>	Direction	Ether Type	IP Protocol	Port Range	Remote	Actions
<input type="checkbox"/>	Egress	IPv4	Any	-	0.0.0.0/0 (CIDR)	Delete Rule
<input type="checkbox"/>	Ingress	IPv4	Any	-	default	Delete Rule
<input type="checkbox"/>	Ingress	IPv6	Any	-	default	Delete Rule
<input type="checkbox"/>	Egress	IPv6	Any	-	::/0 (CIDR)	Delete Rule
<input type="checkbox"/>	Ingress	IPv4	ICMP	-	0.0.0.0/0 (CIDR)	Delete Rule
<input type="checkbox"/>	Ingress	IPv4	TCP	1 - 65535	0.0.0.0/0 (CIDR)	Delete Rule
<input type="checkbox"/>	Ingress	IPv4	TCP	3389 (RDP)	0.0.0.0/0 (CIDR)	Delete Rule
<input type="checkbox"/>	Ingress	IPv4	TCP	27017	0.0.0.0/0 (CIDR)	Delete Rule

Displaying 8 items

Neutron- Network Access Rules of a Security Group

Project

Admin

System Panel

Overview

Hypervisors

Host Aggregates

Instances

Volumes

Flavors

Images

Networks

Routers

System Info

Identity Panel

All Hypervisors

Hypervisor Summary



vCPU Usage
Used 128 of 144



Memory Usage
Used 305GB of 377GB



Disk Usage
Used 2.6TB of 3.1TB

Hypervisors

Hostname	Type	VCPUs (total)	VCPUs (used)	RAM (total)	RAM (used)	Storage (total)	Storage (used)	Instances
node-77.domain.tld	QEMU	48	52	125GB	104GB	1.0TB	930.0GB	13
node-62.domain.tld	QEMU	48	26	125GB	84GB	1.0TB	985.0GB	5
node-79.domain.tld	QEMU	48	50	125GB	116GB	1.0TB	940.0GB	12

Displaying 3 items

Nova-vCPU, RAM, Storage details of Hypervisors

Project

Admin

System Panel

Overview

Hypervisors

Host Aggregates

Instances

Volumes

Flavors

Images

Networks

Routers

System Info

Identity Panel

Flavors

Flavors

 Filter

Filter

+ Create Flavor

Delete Flavors

<input type="checkbox"/>	Flavor Name	VCPUs	RAM	Root Disk	Ephemeral Disk	Swap Disk	ID	Public	Actions
<input type="checkbox"/>	m1.tiny	1	512MB	1GB	0GB	0MB	1	Yes	<button>Edit Flavor</button> More
<input type="checkbox"/>	m1.small	1	2048MB	20GB	0GB	0MB	2	Yes	<button>Edit Flavor</button> More
<input type="checkbox"/>	m1.medium	2	4096MB	40GB	0GB	0MB	3	Yes	<button>Edit Flavor</button> More
<input type="checkbox"/>	IITKGP_regular	2	4096MB	45GB	0GB	0MB	66e4a1a7-249a-4853-925d-6b59e1118b4f	Yes	<button>Edit Flavor</button> More
<input type="checkbox"/>	RamOverCommitTest	2	16384MB	2GB	0GB	0MB	206e40e2-dfba-432a-8bac-61e80147a5ca	Yes	<button>Edit Flavor</button> More
<input type="checkbox"/>	IITKGP_large	4	8192MB	45GB	0GB	0MB	a0266a30-b6b1-4d82-8468-1e4b643dfc51	Yes	<button>Edit Flavor</button> More
<input type="checkbox"/>	m1.large	4	8192MB	80GB	0GB	0MB	4	Yes	<button>Edit Flavor</button> More
<input type="checkbox"/>	Megadoop	4	8192MB	90GB	0GB	1024MB	1cc3f7a3-7678-4139-b51a-e72a6b0a42b4	Yes	<button>Edit Flavor</button> More
<input type="checkbox"/>	Megadoop_new	4	8192MB	90GB	0GB	0MB	dc1aaa5b-d6e8-435d-b994-7172606c9312	Yes	<button>Edit Flavor</button> More
<input type="checkbox"/>	IITKGP_xlarge	8	16384MB	60GB	0GB	0MB	36031ddf-12b0-406c-9343-221567593cff	Yes	<button>Edit Flavor</button> More
<input type="checkbox"/>	m1.xlarge	8	16384MB	160GB	0GB	0MB	5	Yes	<button>Edit Flavor</button> More

Nova- Different flavors of VMs in Meghamala

Project

Admin

System Panel

Overview

Hypervisors

Host Aggregates

Instances

Volumes

Flavors

Images

Networks

Routers

System Info

Identity Panel

Images

Images

Image Name =

Filter



Filter

+ Create Image

Delete Images

<input type="checkbox"/>	Image Name	Type	Status	Public	Protected	Format	Actions
<input type="checkbox"/>	Meghadoop_snapshot_ready	Snapshot	Active	Yes	No	QCOW2	<button>Edit</button> <button>More</button>
<input type="checkbox"/>	CentOS_6.5_GUI	Image	Active	Yes	No	QCOW2	<button>Edit</button> <button>More</button>
<input type="checkbox"/>	Stacksync1_10_4_2_30_01092015	Snapshot	Active	No	No	QCOW2	<button>Edit</button> <button>More</button>
<input type="checkbox"/>	stacksync_working	Snapshot	Active	No	No	QCOW2	<button>Edit</button> <button>More</button>
<input type="checkbox"/>	Ubuntu_14_04_x2go_60G	Image	Active	Yes	No	QCOW2	<button>Edit</button> <button>More</button>
<input type="checkbox"/>	Ubuntu_14_04_x2go_45G	Image	Active	Yes	No	QCOW2	<button>Edit</button> <button>More</button>
<input type="checkbox"/>	Ubuntu_14_04_x2go_20G	Image	Active	Yes	No	QCOW2	<button>Edit</button> <button>More</button>
<input type="checkbox"/>	Ubuntu_New_X2Go	Image	Active	Yes	No	QCOW2	<button>Edit</button> <button>More</button>
<input type="checkbox"/>	Windows_7_x64	Image	Active	Yes	No	QCOW2	<button>Edit</button> <button>More</button>
<input type="checkbox"/>	Fedora_20_GUI	Image	Active	Yes	No	QCOW2	<button>Edit</button> <button>More</button>
<input type="checkbox"/>	Centos_7_GUI	Image	Active	Yes	No	QCOW2	<button>Edit</button> <button>More</button>
<input type="checkbox"/>	Ubu					QCOW2	<button>Edit</button> <button>More</button>
<input type="checkbox"/>	Cen					QCOW2	<button>Edit</button> <button>More</button>

Images of Cloud Instance in Meghamala

Project

Admin

System Panel

Overview

Hypervisors

Host Aggregates

Instances

Volumes

Flavors

Images

Networks

Routers

System Info

Identity Panel

System Info

Services

Compute Services

Network Agents

Default Quotas

Compute Services

Filter



Filter

Name	Host	Zone	Status	State	Updated At
nova-consoleauth	node-61.domain.tld	internal	enabled	up	0 minutes
nova-conductor	node-61.domain.tld	internal	enabled	up	0 minutes
nova-scheduler	node-61.domain.tld	internal	enabled	up	0 minutes
nova-cert	node-61.domain.tld	internal	enabled	up	0 minutes
nova-compute	node-77.domain.tld	nova	enabled	up	0 minutes
nova-compute	node-62.domain.tld	nova	enabled	up	0 minutes
nova-compute	node-79.domain.tld	nova	enabled	up	0 minutes
nova-console	node-61.domain.tld	internal	enabled	up	0 minutes

Displaying 8 items

Compute Services in Meghamala

VM Creation

Project

Compute

Overview

Instances

Volumes

Images

Access & Security

Network

Object Store

Orchestration

Admin

Instances

Launch Instance

Details *

Access & Security *

Networking *

Post-Creation

Advanced Options

Availability Zone

nova

Instance Name *

Cloud_npTEL_1

Flavor *

m1.tiny

m1.tiny

m1.small

m1.medium

IITKGP_regular

Meghdooop

m1.large

IITKGP_large

Meghdooop_new

IITKGP_xlarge_Meghdooop

RamOverCommitTest

IITKGP_xlarge

m1.xlarge

IITKGP_xxlarge

IITKGP_Meghdooop_Bigger

Specify the details for launching an instance.

The chart below shows the resources used by this project in relation to the project's quotas.

Flavor Details

Name m1.tiny

VCPU

1

Root Disk

1 GB

Ephemeral Disk

0 GB

Total Disk

1 GB

RAM

512 MB

Project Limits

Number of Instances

inf of No Limit Used

Number of VCPUs

inf of No Limit Used

Total RAM

inf of No Limit MB Used

Cancel

Launch

<input type="checkbox"/>	Meghdooop_18	CentOS_6.5_GUI	192.164.111.105 10.4.2.52	Meghdooop_new 8GB RAM 4 VCPU 90.0GB Disk	-	Active	nova	None	Running	1 year, 5 months	Create Snapshot	More ▾
<input type="checkbox"/>	Meghdooop_19	CentOS_6.5_GUI	192.164.111.106 10.4.2.53	Meghdooop_new 8GB RAM 4 VCPU 90.0GB Disk	-	Active	nova	None	Running	1 year, 5 months	Create Snapshot	More ▾

Project

Compute

Overview

Instances

Volumes

Images

Access & Security

Network

Object Store

Orchestration

Admin

Instances

Launch Instance

Details *

Access & Security *

Networking *

Post-Creation

Advanced Options

Availability Zone

nova

Instance Name *

Cloud_npTEL_1

Flavor *

IITKGP_regular

Some flavors not meeting minimum image requirements have been disabled.

Instance Count *

1

Instance Boot Source *

Boot from image

Image Name

CentOS 6.5 GUI (1.0 GB)

Specify the details for launching an instance.

The chart below shows the resources used by this project in relation to the project's quotas.

Flavor Details

Name	IITKGP_regular
------	----------------

VCPUs	2
-------	---

Root Disk	45 GB
-----------	-------

Ephemeral Disk	0 GB
----------------	------

Total Disk	45 GB
------------	-------

RAM	4,096 MB
-----	----------

Project Limits

Number of Instances

inf of No Limit Used

Number of VCPUs

inf of No Limit Used

Total RAM

inf of No Limit MB Used

Cancel

Launch

<input type="checkbox"/>	Meghdooop_18	CentOS_6.5_GUI	192.164.111.105 10.4.2.52	Meghdooop_new 8GB RAM 4 VCPU 90.0GB Disk	-	Active	nova	None	Running	1 year, 5 months	Create Snapshot More ▾
<input type="checkbox"/>	Meghdooop_19	CentOS_6.5_GUI	192.164.111.106 10.4.2.53	Meghdooop_new 8GB RAM 4 VCPU 90.0GB Disk	-	Active	nova	None	Running	1 year, 5 months	Create Snapshot More ▾

Instances

	Instance Name	Image	Details *	Access & Security *	Networking *	Post-Creation	Advanced Options	Uptime	Actions
<input type="checkbox"/>	ccTest	CentOS_6.5_GUI	Creating	Ubuntu_14_04_x2go_60G	Selected Networks	Choose network from Available networks to Selected Networks by push button or drag and drop, you may change nic order by drag and drop as well.		2 months, 2 weeks	<button>Create Snapshot</button> <button>More</button>
<input type="checkbox"/>	TestDiskPartition	Ubuntu_14_04_x2go_60G	Creating	Ubuntu_14_04_x2go_60G	Available networks			3 months, 2 weeks	<button>Create Snapshot</button> <button>More</button>
<input type="checkbox"/>	centosForGity	CentOS_6.5_GUI	Creating	Ubuntu_14_04_x2go_60G				7 months	<button>Start Instance</button> <button>More</button>
<input type="checkbox"/>	CL1_R_SERVER1	Ubuntu_New_X2Go	192.164.111.130 10.4.2.28	IITKGP_xlarge 62GB RAM 8 VCPU 60.0GB Disk	-	Active	nova	None	Running
<input type="checkbox"/>	Harshit_Utkarsh_LARGE	Ubuntu_14_04_x2go_60G	192.164.111.129 10.4.2.17	IITKGP_xlarge 16GB RAM 8 VCPU 60.0GB Disk	-	Active	nova	None	Running
<input type="checkbox"/>	cc16_test1	Ubuntu_14_04_x2go_45G	192.164.111.113 10.4.2.18	IITKGP_regular 4GB RAM 2 VCPU 45.0GB Disk	-	Shutoff	nova	None	Shutdown
<input type="checkbox"/>	MeghadoopNewMaster	CentOS_6.5_GUI	192.164.111.111 10.4.2.55	IITKGP_Meghadoop_Bigger 48GB RAM 8 VCPU 600.0GB Disk	-	Active	nova	None	Running
<input type="checkbox"/>	Meghadoop_18	CentOS_6.5_GUI	192.164.111.105 10.4.2.52	Meghadoop_new 8GB RAM 4 VCPU 90.0GB Disk	-	Active	nova	None	Running
<input type="checkbox"/>	Meghadoop_19	CentOS_6.5_GUI	192.164.111.106 10.4.2.53	Meghadoop_new 8GB RAM 4 VCPU 90.0GB Disk	-	Active	nova	None	Running

10.4.2.5/dashboard/project/instances/#

Project

Compute

Overview

Instances

Volumes

Images

Access & Security

Network

Object Store

Orchestration

Admin

Instances

Instance Name



Filter

+ Launch Instance

Soft Reboot Instances

Terminate Instances

	Instance Name	Image Name	IP Address	Size	Key Pair	Status	Availability Zone	Task	Power State	Uptime	Actions
	ccTest	Centos_7_GUI	192.164.111.133 10.4.2.26	IITKGP_regular 4GB RAM 2 VCPU 45.0GB Disk	-	Active	nova	None	Running	2 months, 2 weeks	<button>Create Snapshot</button> <button>More</button>
	TestDiskPartition	Ubuntu_14_04_x2go_45G	192.164.111.132	IITKGP_regular 4GB RAM 2 VCPU 45.0GB Disk	-	Active	nova	None	Running	3 months, 2 weeks	<button>Create Snapshot</button> <button>More</button>
	centosForSify	CentOS_6.5_GUI	192.164.111.131 10.4.2.21	IITKGP_regular 4GB RAM 2 VCPU 45.0GB Disk	-	Shutoff	nova	None	Shutdown	7 months	<button>Start Instance</button> <button>More</button>
	CL1_R_SERVER1	Ubuntu_New_X2Go	192.164.111.130 10.4.2.28	IITKGP_xxlarge 32GB RAM 8 VCPU 60.0GB Disk	-	Active	nova	None	Running	9 months, 1 week	<button>Create Snapshot</button> <button>More</button>
	Harshit_Utkarsh_LARGE	Ubuntu_14_04_x2go_60G	192.164.111.129 10.4.2.17	IITKGP_xlarge 16GB RAM 8 VCPU 60.0GB Disk	-	Active	nova	None	Running	1 year, 2 months	<button>Create Snapshot</button> <button>More</button>
	cc16_test1	Ubuntu_14_04_x2go_45G	192.164.111.113 10.4.2.18	IITKGP_regular 4GB RAM 2 VCPU 45.0GB Disk	-	Shutoff	nova	None	Shutdown	1 year, 4 months	<button>Start Instance</button> <button>More</button>
	MeghadoopNewMaster	CentOS_6.5_GUI	192.164.111.111 10.4.2.55	IITKGP_Meghadoop_Bigger 48GB RAM 8 VCPU 600.0GB Disk	-	Active	nova	None	Running	1 year, 4 months	<button>Create Snapshot</button> <button>More</button>
	Meghadoop_18	CentOS_6.5_GUI	192.164.111.105 10.4.2.52	Meghadoop_new 8GB RAM 4 VCPU 90.0GB Disk	-	Active	nova	None	Running	1 year, 5 months	<button>Create Snapshot</button> <button>More</button>
	Meghadoop_19	CentOS_6.5_GUI	192.164.111.106 10.4.2.53	Meghadoop_new 8GB RAM 4 VCPU 90.0GB Disk	-	Active	nova	None	Running	1 year, 5 months	<button>Create Snapshot</button> <button>More</button>



Project

Compute

Overview

Instances

Volumes

Images

Access & Security

Network

Object Store

Orchestration

Admin

Instances

Instances

Instance Name

Filter

Filter

+ Launch Instance

Soft Reboot Instances

Terminate Instances

<input type="checkbox"/>	Instance Name	Image Name	IP Address	Size	Key Pair	Status	Availability Zone	Task	Power State	Uptime	Actions
<input type="checkbox"/>	Cloud_npTEL_1	CentOS_6.5_GUI	192.164.111.149	IITKGP_regular 4GB RAM 2 VCPU 45.0GB Disk	-	Active	nova	None	Running	1 minute	<button>Create Snapshot</button> <button>More</button>
<input type="checkbox"/>	ccTest	Centos_7_GUI	192.164.0.1 10.4.0.1	IITKGP_regular 4GB RAM 2 VCPU 45.0GB Disk	-	Active	nova	None	Running	2 months, 2 weeks	<button>Associate Floating IP</button> <button>Edit Instance</button> <button>Edit Security Groups</button> <button>Console</button> <button>View Log</button> <button>Pause Instance</button> <button>Suspend Instance</button> <button>Resize Instance</button> <button>Soft Reboot Instance</button> <button>Hard Reboot Instance</button> <button>Shut Off Instance</button> <button>Rebuild Instance</button> <button>Terminate Instance</button>
<input type="checkbox"/>	TestDiskPartition	Ubuntu_14_04_x2go_45G	192.164.0.2 10.4.0.2	IITKGP_regular 4GB RAM 2 VCPU 45.0GB Disk	-	Active	nova	None	Running	3 months, 2 weeks	
<input type="checkbox"/>	centosForSify	CentOS_6.5_GUI	192.164.0.3 10.4.0.3	IITKGP_regular 4GB RAM 2 VCPU 45.0GB Disk	-	Shutoff	nova	None	Shutdown	7 months	
<input type="checkbox"/>	CL1_R_SERVER1	Ubuntu_New_X2Go	192.164.0.4 10.4.0.4	IITKGP_xxlarge 32GB RAM 8 VCPU 60.0GB Disk	-	Active	nova	None	Running	9 months, 1 week	
<input type="checkbox"/>	Harshit_Utkarsh_LARGE	Ubuntu_14_04_x2go_60G	192.164.0.5 10.4.0.5	IITKGP_xlarge 16GB RAM 8 VCPU 60.0GB Disk	-	Active	nova	None	Running	1 year, 2 months	
<input type="checkbox"/>	cc16_test1	Ubuntu_14_04_x2go_45G	192.164.0.6 10.4.0.6	IITKGP_regular 4GB RAM 2 VCPU 45.0GB Disk	-	Shutoff	nova	None	Shutdown	1 year, 4 months	<button>Start Instance</button> <button>More</button>
<input type="checkbox"/>	MeghadoopNewMaster	CentOS_6.5_GUI	192.164.0.7 10.4.0.7	IITKGP_MeghdooP_Bigger 48GB RAM 8 VCPU 600.0GB Disk	-	Active	nova	None	Running	1 year, 4 months	<button>Create Snapshot</button> <button>More</button>
<input type="checkbox"/>	Meghadoop_18	CentOS_6.5_GUI	192.164.0.8 10.4.0.8	Meghadoop_new 8GB RAM 4 VCPU 90.0GB Disk	-	Active	nova	None	Running	1 year, 5 months	<button>Create Snapshot</button> <button>More</button>

Project

Compute

Overview

Instances

Volumes

Images

Access & Security

Network

Object Store

Orchestration

Admin

Instances

Instances

Instance Name

Filter



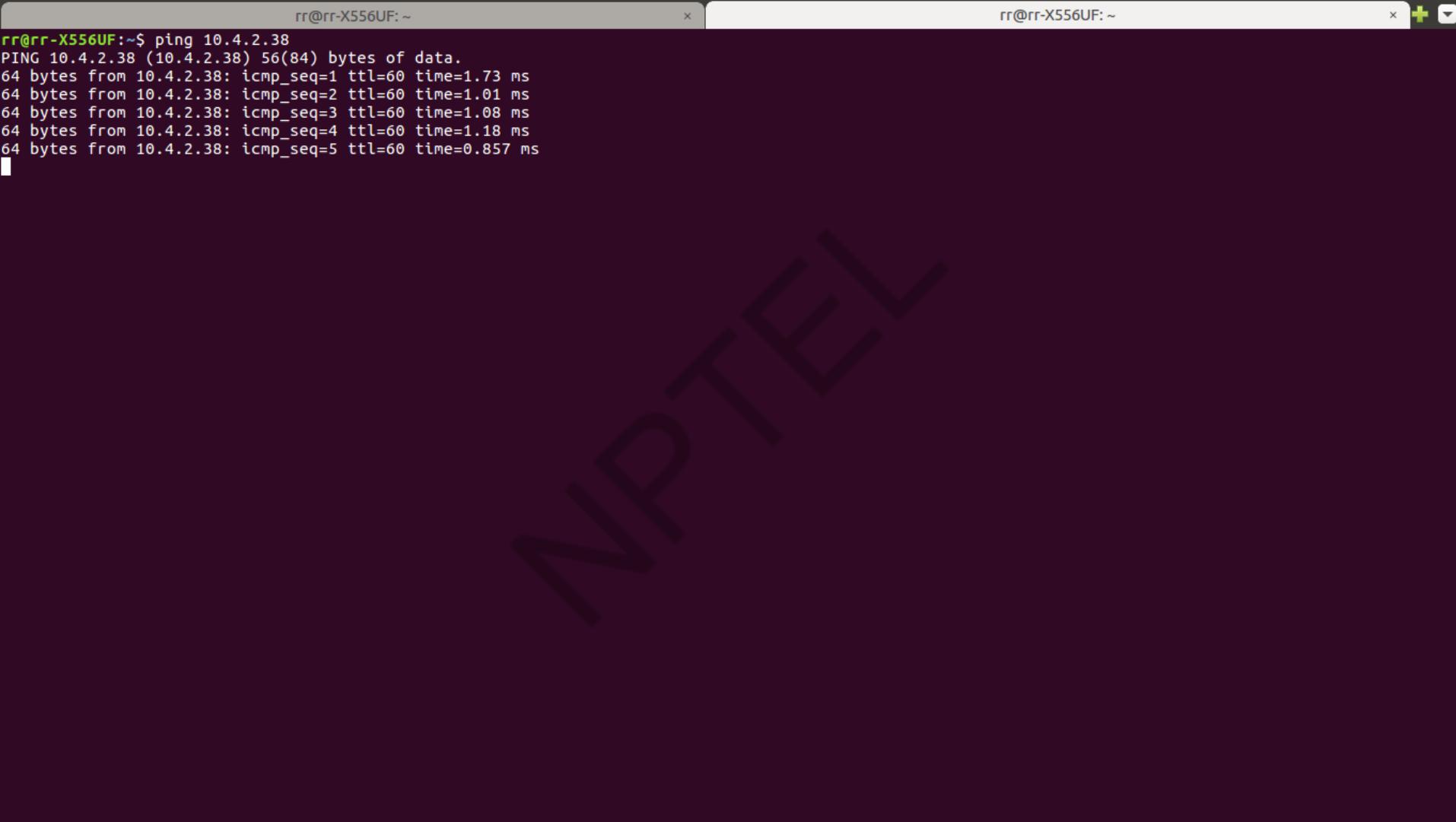
Filter

+ Launch Instance

Soft Reboot Instances

Terminate Instances

	Instance Name	Image Name	IP Address	Size	Key Pair	Status	Availability Zone	Task	Power State	Uptime	Actions
<input type="checkbox"/>	Cloud_npTEL_1	CentOS_6.5_GUI	192.164.111.149 10.4.2.38	IITKGP_regular 4GB RAM 2 VCPU 45.0GB Disk	-	Active	nova	None	Running	2 minutes	<button>Create Snapshot</button> <button>More</button>
<input type="checkbox"/>	ccTest	Centos_7_GUI	192.164.0.1 10.4.0.1	IITKGP_regular 4GB RAM 2 VCPU 45.0GB Disk	-	Active	nova	None	Running	2 months, 2 weeks	<button>Create Snapshot</button> <button>More</button>
<input type="checkbox"/>	TestDiskPartition	Ubuntu_14_04_x2go_45G	192.164.0.2 10.4.0.2	IITKGP_regular 4GB RAM 2 VCPU 45.0GB Disk	-	Active	nova	None	Running	3 months, 2 weeks	<button>Create Snapshot</button> <button>More</button>
<input type="checkbox"/>	centosForSify	CentOS_6.5_GUI	192.164.0.3 10.4.0.3	IITKGP_regular 4GB RAM 2 VCPU 45.0GB Disk	-	Shutoff	nova	None	Shutdown	7 months	<button>Start Instance</button> <button>More</button>
<input type="checkbox"/>	CL1_R_SERVER1	Ubuntu_New_X2Go	192.164.0.4 10.4.0.4	IITKGP_xxlarge 32GB RAM 8 VCPU 60.0GB Disk	-	Active	nova	None	Running	9 months, 1 week	<button>Create Snapshot</button> <button>More</button>
<input type="checkbox"/>	Harshit_Utkarsh_LARGE	Ubuntu_14_04_x2go_60G	192.164.0.5 10.4.0.5	IITKGP_xlarge 16GB RAM 8 VCPU 60.0GB Disk	-	Active	nova	None	Running	1 year, 2 months	<button>Create Snapshot</button> <button>More</button>
<input type="checkbox"/>	cc16_test1	Ubuntu_14_04_x2go_45G	192.164.0.6 10.4.0.6	IITKGP_regular 4GB RAM 2 VCPU 45.0GB Disk	-	Shutoff	nova	None	Shutdown	1 year, 4 months	<button>Start Instance</button> <button>More</button>
<input type="checkbox"/>	MeghdooPNewMaster	CentOS_6.5_GUI	192.164.0.7 10.4.0.7	IITKGP_Meghdoo_Bigger 48GB RAM 8 VCPU 600.0GB Disk	-	Active	nova	None	Running	1 year, 4 months	<button>Create Snapshot</button> <button>More</button>
<input type="checkbox"/>	MeghdooP_18	CentOS_6.5_GUI	192.164.0.8 10.4.0.8	MeghdooP_new 8GB RAM 4 VCPU 90.0GB Disk	-	Active	nova	None	Running	1 year, 5 months	<button>Create Snapshot</button> <button>More</button>



Accessing VM by User

Session preferences - cloud-nptel

Session Connection Input/Output Media Shared foldersSession name: cloud-nptel

<< change icon

Path: /

Server

Host: 10.4.2.38

Login: centos

SSH port: 22

Use RSA/DSA key for ssh connection: Try auto login (via SSH Agent or default SSH key) Kerberos 5 (GSSAPI) authentication Delegation of GSSAPI credentials to the server Use Proxy server for SSH connectionSession type

XFCE

Command: OKCancelDefaults

Accessing of newly created VM through X2Go Client

Applications Menu

centos - File Browser

04:53



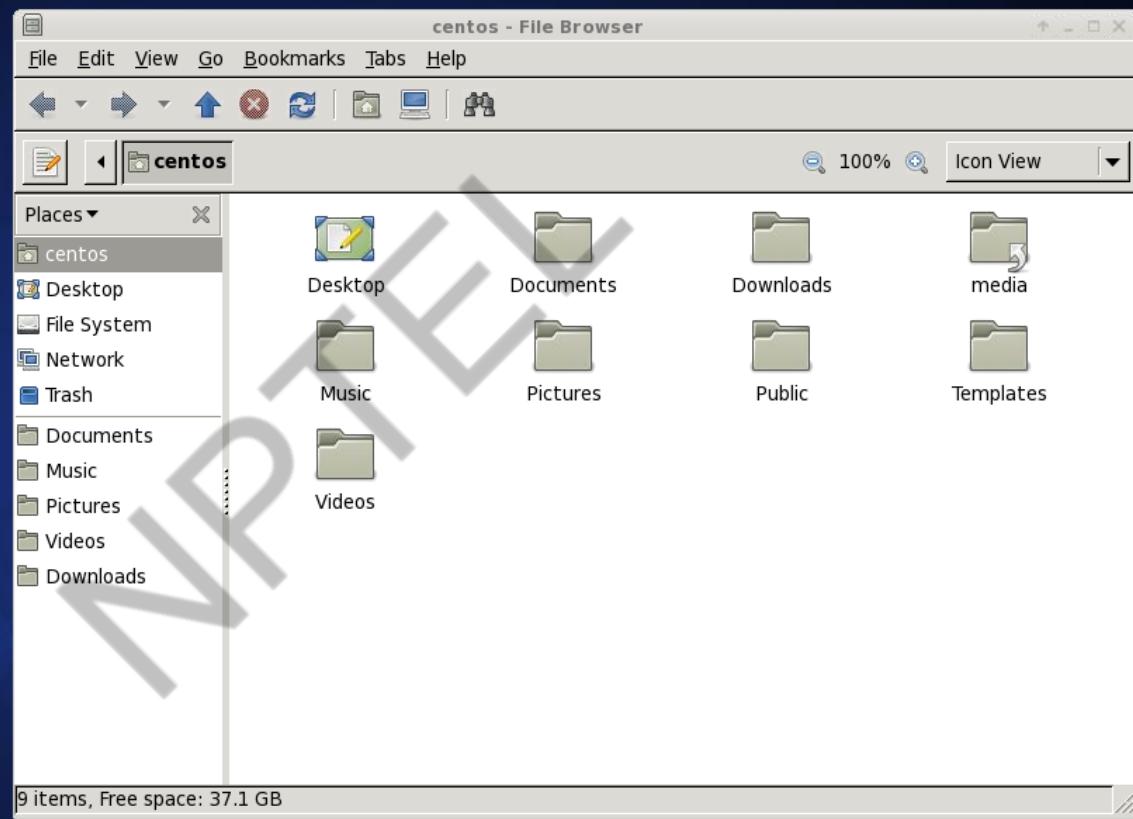
Computer



centos's Home



Trash

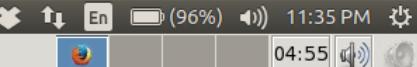


Accessing newly created VM - 'cloud-nptel'

Applications Menu

Google - Mozilla Firefox

centos - File Browser



04:55



Google - Mozilla Firefox

Google

https://www.google.co.in/?gfe_rd=cr&ei=7thbWcQt6_LwB4XjpYAM&gws_rd=ssl

Search



Gmail Images



Sign in

Google India

Google Search

I'm Feeling Lucky

Come here often? Make Google your homepage.

[Yes, show me](#)

Google.co.in offered in: हिन्दी वांग्ना தமிழ் மராதி தமிழ் ગુજરાતી କ୍ଷେତ୍ର ମହାଜ୍ଞ ପੰਜਾਬੀ

[Advertising](#)[Business](#)[About](#)[Privacy](#)[Terms](#)[Settings](#)[Use Google.com](#)



Prof. Soumya K. Ghosh - YouTube - Mozilla Firefox

<https://www.youtube.com/watch?v=OL8prdrpjOg>

Search

05:01

soumya k ghosh



Prof. Soumya K. Ghosh

Up next

Autoplay 

VM Termination

Project

Compute

Overview

Instances

Volumes

Images

Access & Security

Network

Object Store

Orchestration

Admin

Instances

Instances

Instance Name

Filter

Filter

+ Launch Instance

Soft Reboot Instances

Terminate Instances

	Instance Name	Image Name	IP Address	Size	Key Pair	Status	Availability Zone	Task	Power State	Uptime	Actions
<input checked="" type="checkbox"/>	Cloud_npTEL_1	CentOS_6.5_GUI	192.164.111.149 10.4.2.38	IITKGP_regular 4GB RAM 2 VCPU 45.0GB Disk	-	Active	nova	None	Running	2 minutes	<button>Create Snapshot</button> <button>More</button>
<input type="checkbox"/>	ccTest	Centos_7_GUI	192.164.0.1 10.4.0.1	IITKGP_regular 4GB RAM 2 VCPU 45.0GB Disk	-	Active	nova	None	Running	2 months, 2 weeks	<button>Create Snapshot</button> <button>More</button>
<input type="checkbox"/>	TestDiskPartition	Ubuntu_14_04_x2go_45G	192.164.0.2 10.4.0.2	IITKGP_regular 4GB RAM 2 VCPU 45.0GB Disk	-	Active	nova	None	Running	3 months, 2 weeks	<button>Create Snapshot</button> <button>More</button>
<input type="checkbox"/>	centosForSify	CentOS_6.5_GUI	192.164.0.3 10.4.0.3	IITKGP_regular 4GB RAM 2 VCPU 45.0GB Disk	-	Shutoff	nova	None	Shutdown	7 months	<button>Start Instance</button> <button>More</button>
<input type="checkbox"/>	CL1_R_SERVER1	Ubuntu_New_X2Go	192.164.0.4 10.4.0.4	IITKGP_xxlarge 32GB RAM 8 VCPU 60.0GB Disk	-	Active	nova	None	Running	9 months, 1 week	<button>Create Snapshot</button> <button>More</button>
<input type="checkbox"/>	Harshit_Utkarsh_LARGE	Ubuntu_14_04_x2go_60G	192.164.0.5 10.4.0.5	IITKGP_xlarge 16GB RAM 8 VCPU 60.0GB Disk	-	Active	nova	None	Running	1 year, 2 months	<button>Create Snapshot</button> <button>More</button>
<input type="checkbox"/>	cc16_test1	Ubuntu_14_04_x2go_45G	192.164.0.6 10.4.0.6	IITKGP_regular 4GB RAM 2 VCPU 45.0GB Disk	-	Shutoff	nova	None	Shutdown	1 year, 4 months	<button>Start Instance</button> <button>More</button>
<input type="checkbox"/>	MeghdooPNewMaster	CentOS_6.5_GUI	192.164.0.7 10.4.0.7	IITKGP_Meghdoo_Bigger 48GB RAM 8 VCPU 600.0GB Disk	-	Active	nova	None	Running	1 year, 4 months	<button>Create Snapshot</button> <button>More</button>
<input type="checkbox"/>	MeghdooP_18	CentOS_6.5_GUI	192.164.0.8 10.4.0.8	MeghdooP_new 8GB RAM 4 VCPU 90.0GB Disk	-	Active	nova	None	Running	1 year, 5 months	<button>Create Snapshot</button> <button>More</button>

Project

Compute

Overview

Instances

Volumes

Images

Access & Security

Network

Object Store

Orchestration

Admin

Instances

Confirm Terminate Instances

You have selected "Cloud_ntpel_1". Please confirm your selection. This action cannot be undone.

Cancel

Terminate Instances

Instance Name	Image	Created	Flavor	Network	Status	Host	Port	Uptime	Actions
<input checked="" type="checkbox"/> Cloud_ntpel_1	CentOS_6.5_Gui	10.4.2.38 10.4.2.26	IITKGP_regular 4GB RAM 2 VCPU 45.0GB Disk	-	Active	nova	None	2 minutes	<button>Create Snapshot</button> <button>More</button>
<input type="checkbox"/> ccTest	Centos_7_GUI	192.164.111.133 10.4.2.26	IITKGP_regular 4GB RAM 2 VCPU 45.0GB Disk	-	Active	nova	None	2 months, 2 weeks	<button>Create Snapshot</button> <button>More</button>
<input type="checkbox"/> TestDiskPartition	Ubuntu_14_04_x2go_45G	192.164.111.132	IITKGP_regular 4GB RAM 2 VCPU 45.0GB Disk	-	Active	nova	None	3 months, 2 weeks	<button>Create Snapshot</button> <button>More</button>
<input type="checkbox"/> centosForSify	CentOS_6.5_GUI	192.164.111.131 10.4.2.21	IITKGP_regular 4GB RAM 2 VCPU 45.0GB Disk	-	Shutoff	nova	None	7 months	<button>Start Instance</button> <button>More</button>
<input type="checkbox"/> CL1_R_SERVER1	Ubuntu_New_X2Go	192.164.111.130 10.4.2.28	IITKGP_xxlarge 32GB RAM 8 VCPU 60.0GB Disk	-	Active	nova	None	9 months, 1 week	<button>Create Snapshot</button> <button>More</button>
<input type="checkbox"/> Harshit_Utkarsh_LARGE	Ubuntu_14_04_x2go_60G	192.164.111.129 10.4.2.17	IITKGP_xlarge 16GB RAM 8 VCPU 60.0GB Disk	-	Active	nova	None	1 year, 2 months	<button>Create Snapshot</button> <button>More</button>
<input type="checkbox"/> cc16_test1	Ubuntu_14_04_x2go_45G	192.164.111.113 10.4.2.18	IITKGP_regular 4GB RAM 2 VCPU 45.0GB Disk	-	Shutoff	nova	None	1 year, 4 months	<button>Start Instance</button> <button>More</button>
<input type="checkbox"/> MegahadoopNewMaster	CentOS_6.5_GUI	192.164.111.111 10.4.2.55	IITKGP_Megahadoop_Bigger 48GB RAM 8 VCPU 600.0GB Disk	-	Active	nova	None	1 year, 4 months	<button>Create Snapshot</button> <button>More</button>
<input type="checkbox"/> Megahadoop_18	CentOS_6.5_GUI	192.164.111.105 10.4.2.52	Megahadoop_new 8GB RAM 4 VCPU 90.0GB Disk	-	Active	nova	None	1 year, 5 months	<button>Create Snapshot</button> <button>More</button>

Success: Scheduled termination of
Instance: Cloud_ntpel_1

Project

Compute

Overview

Instances

Volumes

Images

Access & Security

Network

Object Store

Orchestration

Admin

Instances

Instances

Instance Name

Filter

Filter

+ Launch Instance

Soft Reboot Instances

Terminate Instances

<input type="checkbox"/>	Instance Name	Image Name	IP Address	Size	Key Pair	Status	Availability Zone	Task	Power State	Uptime	Actions
<input type="checkbox"/>	Cloud_ntpel_1	CentOS_6.5_GUI	192.164.111.149 10.4.2.38	IITKGP_regular 4GB RAM 2 VCPU 45.0GB Disk	-	Active	nova	 Deleting	Running	34 minutes	
<input type="checkbox"/>	ccTest	Centos_7_GUI	192.164.0.1 10.4.0.1	IITKGP_regular 4GB RAM 2 VCPU 45.0GB Disk	-	Active	nova	None	Running	2 months, 2 weeks	<button>Create Snapshot</button> <button>More</button>
<input type="checkbox"/>	TestDiskPartition	Ubuntu_14_04_x2go_45G	192.164.0.2 10.4.0.2	IITKGP_regular 4GB RAM 2 VCPU 45.0GB Disk	-	Active	nova	None	Running	3 months, 2 weeks	<button>Create Snapshot</button> <button>More</button>
<input type="checkbox"/>	centosForSify	CentOS_6.5_GUI	192.164.0.3 10.4.0.3	IITKGP_regular 4GB RAM 2 VCPU 45.0GB Disk	-	Shutoff	nova	None	Shutdown	7 months	<button>Start Instance</button> <button>More</button>
<input type="checkbox"/>	CL1_R_SERVER1	Ubuntu_New_X2Go	192.164.0.4 10.4.0.4	IITKGP_xxlarge 32GB RAM 8 VCPU 60.0GB Disk	-	Active	nova	None	Running	9 months, 1 week	<button>Create Snapshot</button> <button>More</button>
<input type="checkbox"/>	Harshit_Utkarsh_LARGE	Ubuntu_14_04_x2go_60G	192.164.0.5 10.4.0.5	IITKGP_xlarge 16GB RAM 8 VCPU 60.0GB Disk	-	Active	nova	None	Running	1 year, 2 months	<button>Create Snapshot</button> <button>More</button>
<input type="checkbox"/>	cc16_test1	Ubuntu_14_04_x2go_45G	192.164.0.6 10.4.0.6	IITKGP_regular 4GB RAM 2 VCPU 45.0GB Disk	-	Shutoff	nova	None	Shutdown	1 year, 4 months	<button>Start Instance</button> <button>More</button>
<input type="checkbox"/>	MeghadoopNewMaster	CentOS_6.5_GUI	192.164.0.7 10.4.0.7	IITKGP_Meghdoo_Bigger 48GB RAM 8 VCPU 600.0GB Disk	-	Active	nova	None	Running	1 year, 4 months	<button>Create Snapshot</button> <button>More</button>
<input type="checkbox"/>	Meghadoop_18	CentOS_6.5_GUI	192.164.0.8 10.4.0.8	Meghadoop_new 8GB RAM 4 VCPU 90.0GB Disk	-	Active	nova	None	Running	1 year, 5 months	<button>Create Snapshot</button> <button>More</button>

Instances

Instances

Instance Name

Filter

Filter

+ Launch Instance

Soft Reboot Instances

Terminate Instances

	Instance Name	Image Name	IP Address	Size	Key Pair	Status	Availability Zone	Task	Power State	Uptime	Actions
<input type="checkbox"/>	ccTest	Centos_7_GUI	192.164.0.1 10.4.0.1	IITKG_P_regular 4GB RAM 2 VCPU 45.0GB Disk	-	Active	nova	None	Running	2 months, 2 weeks	<button>Create Snapshot</button> <button>More</button>
<input type="checkbox"/>	TestDiskPartition	Ubuntu_14_04_x2go_45G	192.164.0.2 10.4.0.2	IITKG_P_regular 4GB RAM 2 VCPU 45.0GB Disk	-	Active	nova	None	Running	3 months, 2 weeks	<button>Create Snapshot</button> <button>More</button>
<input type="checkbox"/>	centosForSify	CentOS_6.5_GUI	192.164.0.3 10.4.0.3	IITKG_P_regular 4GB RAM 2 VCPU 45.0GB Disk	-	Shutoff	nova	None	Shutdown	7 months	<button>Start Instance</button> <button>More</button>
<input type="checkbox"/>	CL1_R_SERVER1	Ubuntu_New_X2Go	192.164.0.4 10.4.0.4	IITKG_P_xxlarge 32GB RAM 8 VCPU 60.0GB Disk	-	Active	nova	None	Running	9 months, 1 week	<button>Create Snapshot</button> <button>More</button>
<input type="checkbox"/>	Harshit_Utkarsh_LARGE	Ubuntu_14_04_x2go_60G	192.164.0.5 10.4.0.5	IITKG_P_xlarge 16GB RAM 8 VCPU 60.0GB Disk	-	Active	nova	None	Running	1 year, 2 months	<button>Create Snapshot</button> <button>More</button>
<input type="checkbox"/>	cc16_test1	Ubuntu_14_04_x2go_45G	192.164.0.6 10.4.0.6	IITKG_P_regular 4GB RAM 2 VCPU 45.0GB Disk	-	Shutoff	nova	None	Shutdown	1 year, 4 months	<button>Start Instance</button> <button>More</button>
<input type="checkbox"/>	MeghdooPNewMaster	CentOS_6.5_GUI	192.164.0.7 10.4.0.7	IITKG_P_MeghdooP_Bigger 48GB RAM 8 VCPU 600.0GB Disk	-	Active	nova	None	Running	1 year, 4 months	<button>Create Snapshot</button> <button>More</button>
<input type="checkbox"/>	MeghdooP_18	CentOS_6.5_GUI	192.164.0.8 10.4.0.8	MeghdooP_new 8GB RAM 4 VCPU 90.0GB Disk	-	Active	nova	None	Running	1 year, 5 months	<button>Create Snapshot</button> <button>More</button>
<input type="checkbox"/>	MeghdooP_19	CentOS_6.5_GUI	192.164.0.9 10.4.0.9	MeghdooP_new 8GB RAM 4 VCPU 90.0GB Disk	-	Active	nova	None	Running	1 year, 5 months	<button>Create Snapshot</button> <button>More</button>

Thank You!

NPTEL



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

CLOUD COMPUTING

CREATE A PYTHON WEB APP IN MICROSOFT AZURE:

PROF. SOUMYA K. GHOSH
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
IIT KHARAGPUR

Microsoft Azure : An overview

- 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.

Ref: <https://azure.microsoft.com/en-in/>



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

Deploy anywhere with your choice of tools

- Connecting cloud and on-premises with consistent hybrid cloud capabilities and using open source technologies



Ref: <https://azure.microsoft.com/en-in/>

Protect your business with the most trusted cloud

- Azure helps to protect assets through a rigorous methodology and focus on security, privacy, compliance and transparency.



Achieve global scale in local regions



Detect and mitigate threats



Rely on the most trusted cloud

Ref: <https://azure.microsoft.com/en-in/>

Accelerate app innovation

- Build simple to complex projects within a consistent portal experience using deeply-integrated cloud services, so developers can rapidly develop, deploy and manage their apps.



Ref: <https://azure.microsoft.com/en-in/>

Power decisions and apps with insights

- 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

Ref: <https://azure.microsoft.com/en-in/>

In this demo, we are going to present the creation of a python web app in Microsoft Azure.

Ref: <https://azure.microsoft.com/en-in/>



IIT KHARAGPUR



NPTEL
ONLINE
CERTIFICATION COURSES

Azure Web Apps

- Highly scalable, Self-patching web hosting service.
- Prerequisites
 - ✓ To complete this demo:
 - ➔ Install Git
 - ➔ Install Python

Ref: <https://azure.microsoft.com/en-in/>

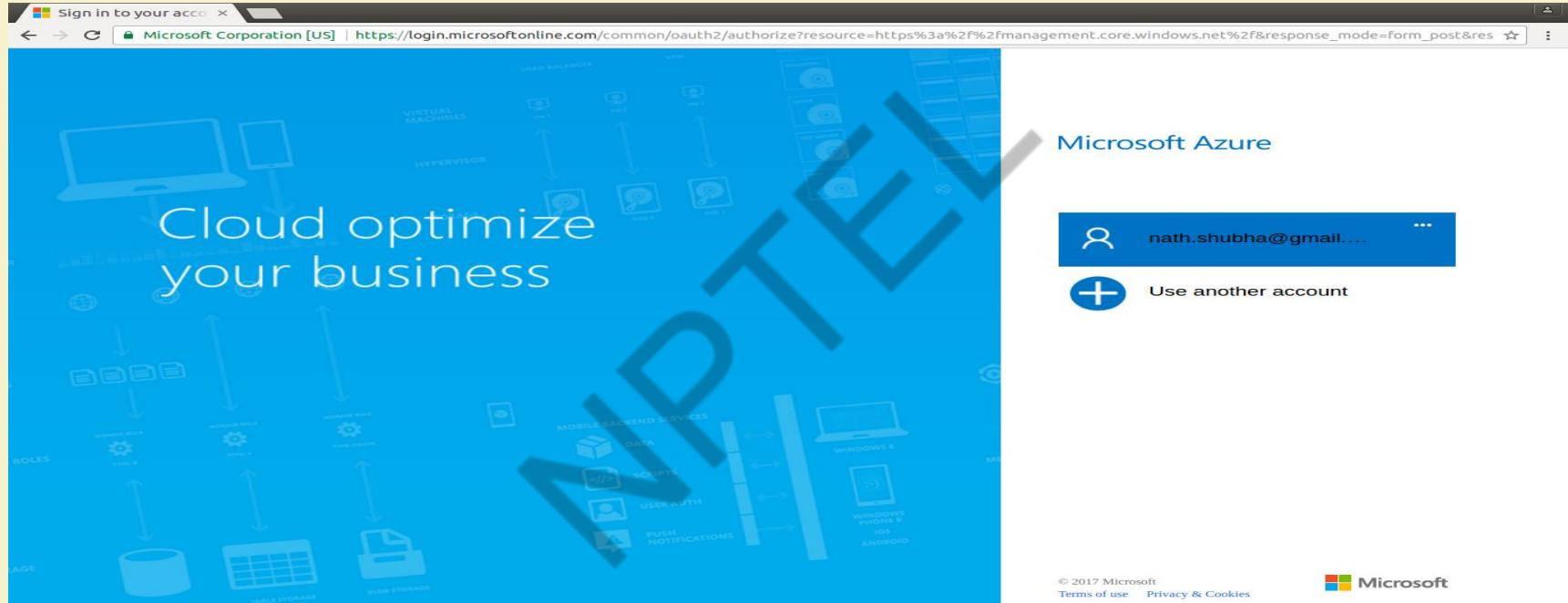


IIT KHARAGPUR



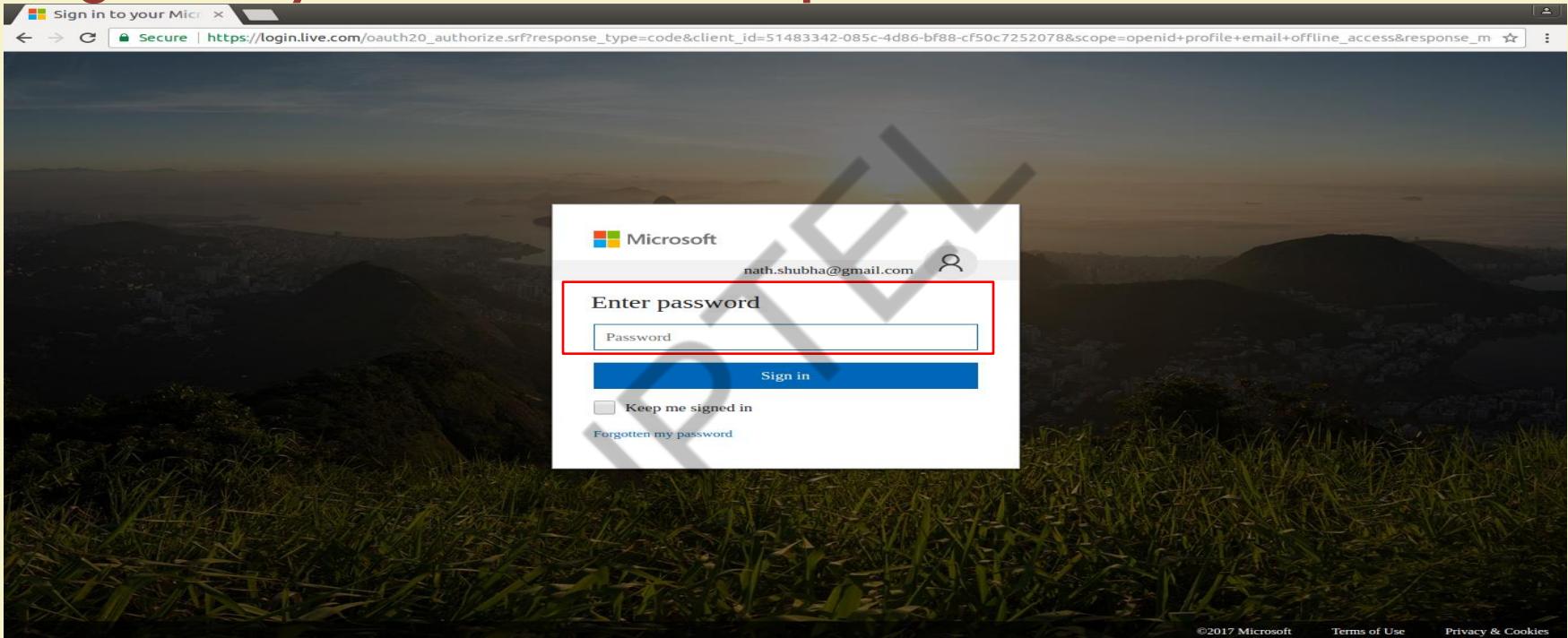
NPTEL
ONLINE
CERTIFICATION COURSES

Go to <https://portal.azure.com/> and login with your username and password



Ref: <https://docs.microsoft.com/en-us/azure/app-service-web/app-service-web-get-started-python>

Login with your username and password



Ref: <https://docs.microsoft.com/en-us/azure/app-service-web/app-service-web-get-started-python>

Launch Azure Cloud Shell : It is a free bash shell that we can directly use within the Azure portal

The screenshot shows the Microsoft Azure portal dashboard. On the left, there's a sidebar with various service icons like New, Dashboard, All resources, Resource groups, App Services, Function Apps, SQL databases, Azure Cosmos DB, Virtual machines, Load balancers, and More services. Below this is a Bash terminal window with the command "nath_shubha@Azure:~\$". At the top right of the dashboard, there's a toolbar with icons for Search resources, Edit dashboard, Share, Fullscreen, Clone, Delete, and a dropdown menu. A red box highlights the dropdown menu icon (three dots) in the toolbar.

Ref: <https://docs.microsoft.com/en-us/azure/app-service-web/app-service-web-get-started-python>

Download the sample

In a terminal window, run the following command to clone the sample app repository to your local machine.

```
root@shubha-OptiPlex-9020:/home/shubha
root@shubha-OptiPlex-9020:/home/shubha# git clone https://github.com/Azure-Samples/python-docs-hello-world
Cloning into 'python-docs-hello-world'...
remote: Counting objects: 18, done.
remote: Total 18 (delta 0), reused 0 (delta 0), pack-reused 18
Unpacking objects: 100% (18/18), done.
Checking connectivity... done.
root@shubha-OptiPlex-9020:/home/shubha#
```

Ref: <https://docs.microsoft.com/en-us/azure/app-service-web/app-service-web-get-started-python>

Change to the directory that contains the sample code

```
root@shubha-OptiPlex-9020: /home/shubha/python-docs-hello-world
root@shubha-OptiPlex-9020: /home/shubha# cd python-docs-hello-world/
root@shubha-OptiPlex-9020: /home/shubha/python-docs-hello-world# █
```

Ref: <https://docs.microsoft.com/en-us/azure/app-service-web/app-service-web-get-started-python>

Install flask

```
root@shubha-OptiPlex-9020:/home/shubha/python-docs-hello-world
root@shubha-OptiPlex-9020:/home/shubha/python-docs-hello-world# pip install flask
Collecting flask
  Downloading Flask-0.12.2-py2.py3-none-any.whl (83kB)
    100% |██████████| 92kB 140kB/s
Collecting itsdangerous>=0.21 (from flask)
  Downloading itsdangerous-0.24.tar.gz (46kB)
    100% |██████████| 51kB 4.3MB/s
Collecting click>=2.0 (from flask)
  Downloading click-6.7-py2.py3-none-any.whl (71kB)
    100% |██████████| 71kB 322kB/s
Collecting Werkzeug>=0.7 (from flask)
  Downloading Werkzeug-0.12.2-py2.py3-none-any.whl (312kB)
    100% |██████████| 317kB 408kB/s
Collecting Jinja2>=2.4 (from flask)
  Downloading Jinja2-2.9.6-py2.py3-none-any.whl (340kB)
    100% |██████████| 348kB 389kB/s
Collecting MarkupSafe>=0.23 (from Jinja2>=2.4->flask)
  Downloading MarkupSafe-1.0.tar.gz
Building wheels for collected packages: itsdangerous, MarkupSafe
  Running setup.py bdist_wheel for itsdangerous ... done
  Stored in directory: /root/.cache/pip/wheels/fc/a8/66/24d655233c757e178d45dea2de22a04c6d92766abfb741129a
  Running setup.py bdist_wheel for MarkupSafe ... done
  Stored in directory: /root/.cache/pip/wheels/88/a7/30/e39a54a87bcbe25308fa3ca64e8ddc75d9b3e5afa21ee32d57
Successfully built itsdangerous MarkupSafe
Installing collected packages: itsdangerous, click, Werkzeug, MarkupSafe, Jinja2, flask
Successfully installed Jinja2-2.9.6 MarkupSafe-1.0 Werkzeug-0.12.2 click-6.7 flask-0.12.2 itsdangerous-0.24
You are using pip version 8.1.1, however version 9.0.1 is available.
You should consider upgrading via the 'pip install --upgrade pip' command.
root@shubha-OptiPlex-9020:/home/shubha/python-docs-hello-world#
```

Ref: <https://docs.microsoft.com/en-us/azure/app-service-web/app-service-web-get-started-python>

Run the app locally

```
root@shubha-OptiPlex-9020: /home/shubha/python-docs-hello-world
root@shubha-OptiPlex-9020: /home/shubha/python-docs-hello-world# python main.py
 * Running on http://127.0.0.1:5000/ (Press CTRL+C to quit)
```

NPTEL

Ref: <https://docs.microsoft.com/en-us/azure/app-service-web/app-service-web-get-started-python>

Open a web browser, and navigate to the sample app at <http://localhost:5000>. You can see the Hello World message from the sample app displayed in the page.



Ref: <https://docs.microsoft.com/en-us/azure/app-service-web/app-service-web-get-started-python>

Configure a deployment user using the command

- A deployment user is required for FTP and local Git deployment to a web app.

```
az webapp deployment user set --user-name <username> --password <password>
```

Ref: <https://docs.microsoft.com/en-us/azure/app-service-web/app-service-web-get-started-python>

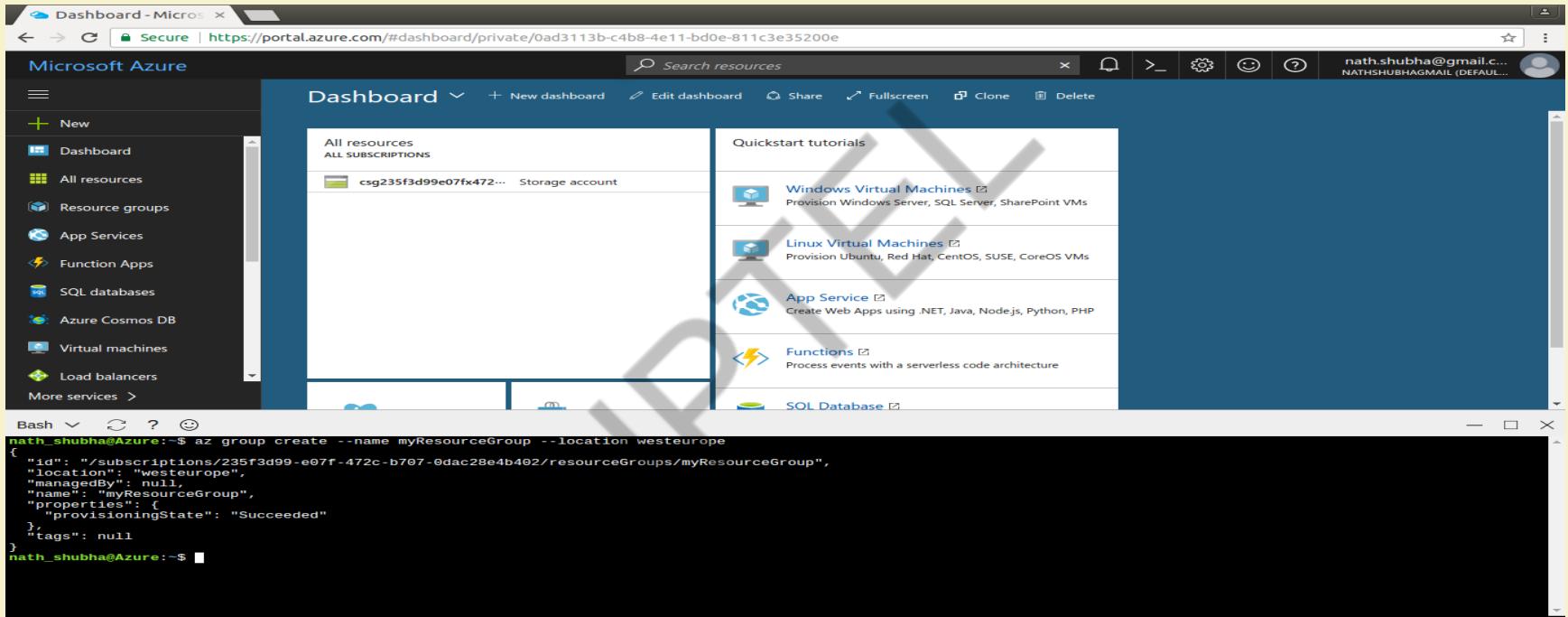


IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

Create a resource group: A resource group is a logical container into which Azure resources like web apps, databases, and storage accounts are deployed and managed.



The screenshot shows the Microsoft Azure portal dashboard. On the left, there's a sidebar with various service icons and a 'Bash' terminal window at the bottom. The main area displays 'All resources ALL SUBSCRIPTIONS' and a 'Quickstart tutorials' section with links for Windows Virtual Machines, Linux Virtual Machines, App Service, Functions, and SQL Database. In the Bash terminal, the command `az group create --name myResourceGroup --location westeurope` is run, followed by its JSON output, and then the command is repeated again.

```
nath_shubha@Azure:~$ az group create --name myResourceGroup --location westeurope
{
  "id": "/subscriptions/235f3d99-e07f-472c-b707-0dac28e4b402/resourceGroups/myResourceGroup",
  "location": "westeurope",
  "managedBy": null,
  "name": "myResourceGroup",
  "properties": {
    "provisioningState": "Succeeded"
  },
  "tags": null
}
nath_shubha@Azure:~$
```

Ref: <https://docs.microsoft.com/en-us/azure/app-service-web/app-service-web-get-started-python>

Create an Azure App Service plan

- An App Service plan specifies the location, size, and features of the web server farm that hosts your app. You can save money when hosting multiple apps by configuring the web apps to share a single App Service plan.
- App Service plans define:
 - Region (for example: North Europe, East US, or Southeast Asia)
 - Instance size (small, medium, or large)
 - Scale count (1 to 20 instances)
 - SKU (Free, Shared, Basic, Standard, or Premium)

Ref: <https://docs.microsoft.com/en-us/azure/app-service-web/app-service-web-get-started-python>



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

Create an Azure App Service plan

The screenshot shows the Microsoft Azure portal dashboard. On the left, there's a sidebar with navigation links like 'Dashboard', 'All resources', 'Resource groups', etc. The main area shows 'All resources ALL SUBSCRIPTIONS' and a 'Storage account'. To the right, there's a 'Quickstart tutorials' section with links to 'Windows Virtual Machines', 'Linux Virtual Machines', 'App Service', 'Functions', and 'SQL Database'. At the bottom, a terminal window is open with the following command:

```
nath.shubha@Azure:~$ az appservice plan create --name myAppServicePlan --resource-group myResourceGroup --sku FREE
{
  "adminSiteName": null,
  "appServicePlanName": "myAppServicePlan",
  "geoRegion": "West Europe",
  "hostingEnvironmentProfile": null,
  "id": "/subscriptions/235f3d99-e07f-472c-b707-0dac28e4b402/resourceGroups/myResourceGroup/providers/Microsoft.Web/serverFarms/myAppServicePlan",
  "kind": "app",
  "location": "West Europe",
  "maximumNumberOfWorkers": 1,
  "name": "myAppServicePlan",
  "numberOfSites": 0,
  "perSiteScaling": false,
  "provisioningState": "Succeeded",
  "reserved": false,
  "resourceGroup": "myResourceGroup",
  "sku": {
```

Ref: <https://docs.microsoft.com/en-us/azure/app-service-web/app-service-web-get-started-python>

Create a web app

- The web app provides a hosting space for your code and provides a URL to view the deployed app.

NPTEL

Ref: <https://docs.microsoft.com/en-us/azure/app-service-web/app-service-web-get-started-python>

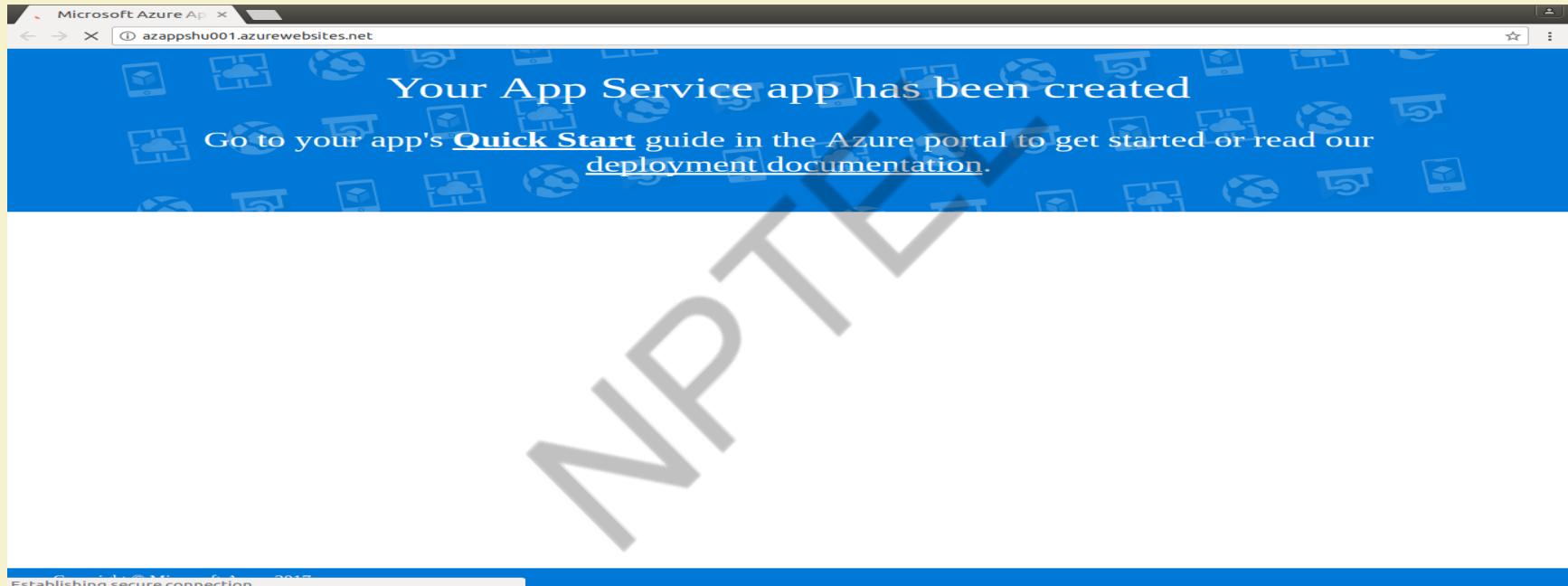
Create a web app

The screenshot shows the Microsoft Azure portal dashboard. On the left, there's a sidebar with a 'New' button and a list of services: Dashboard, All resources, Resource groups, App Services, Function Apps, SQL databases, Azure Cosmos DB, Virtual machines, Load balancers, and More services. The main area displays 'All resources ALL SUBSCRIPTIONS' with one item listed: 'csg235f3d99e07fx472... Storage account'. To the right, there's a 'Quickstart tutorials' section with links to Windows Virtual Machines, Linux Virtual Machines, App Service, Functions, and SQL Database. At the bottom, a terminal window shows the command: `nath.shubha@Azure:~$ az webapp create --name azappshu001 --resource-group myResourceGroup --plan myAppServicePlan`. The output of the command is displayed, showing details about the newly created web app.

```
nath.shubha@Azure:~$ az webapp create --name azappshu001 --resource-group myResourceGroup --plan myAppServicePlan
{
  "availabilityState": "Normal",
  "clientAffinityEnabled": true,
  "clientCertEnabled": false,
  "cloningInfo": null,
  "containerSize": 0,
  "dailyMemoryTimeQuota": 0,
  "defaultHostName": "azappshu001.azurewebsites.net",
  "enabled": true,
  "enabledHostNames": [
    "azappshu001.azurewebsites.net",
    "azappshu001.scm.azurewebsites.net"
  ],
  "ftpPublishingUrl": "ftp://waws-prod-am2-121.ftp.azurewebsites.windows.net/site/wwwroot",
  "gatewaySiteName": null,
  "hostNameSslStates": [
    ...
  ]
}
```

Ref: <https://docs.microsoft.com/en-us/azure/app-service-web/app-service-web-get-started-python>

Browse to the site azappshu001.azurewebsites.net to see your newly created web app.



Ref: <https://docs.microsoft.com/en-us/azure/app-service-web/app-service-web-get-started-python>

Configure to use Python: Setting the Python version this way uses a default container provided by the platform.

The screenshot shows the Microsoft Azure portal interface. On the left, there's a sidebar with various service icons like Dashboard, All resources, Resource groups, App Services, etc. The main area is titled 'Dashboard' and shows 'All resources ALL SUBSCRIPTIONS'. It lists a single item: 'csg235f3d99e07fx472... Storage account'. To the right of the storage account, there's a section titled 'Quickstart tutorials' with links to 'Windows Virtual Machines', 'Linux Virtual Machines', 'App Service', 'Functions', and 'SQL Database'. At the bottom of the portal, there's a terminal window showing a command being run:

```
nath.shubha@Azure:~$ az webapp config set --python-version 3.4 --name azappshu001 --resource-group myResourceGroup
{
  "alwaysOn": false,
  "apiDefinition": null,
  "appCommandLine": "",
  "appSettings": null,
  "autoHealEnabled": false,
  "autoHealRules": [
    {
      "actions": null,
      "triggers": null
    }
  ],
  "autoSwapSlotName": null,
  "connectionStrings": null,
  "cors": null,
  "defaultDocuments": [
    "Default.htm",
    "Default.html"
  ],
  ...
}
```

Ref: <https://docs.microsoft.com/en-us/azure/app-service-web/app-service-web-get-started-python>

Configure local Git deployment

- App Service supports several ways to deploy content to a web app, such as FTP, local Git, GitHub, Visual Studio Team Services, and Bitbucket. For this quickstart, you deploy by using local Git. That means you deploy by using a Git command to push from a local repository to a repository in Azure.

Ref: <https://docs.microsoft.com/en-us/azure/app-service-web/app-service-web-get-started-python>



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

Configure local Git deployment

The screenshot shows the Microsoft Azure portal interface. On the left, there's a sidebar with various service icons like Storage account, App Services, Functions, and SQL databases. The main area is titled 'Dashboard' and shows 'All resources ALL SUBSCRIPTIONS'. It lists a single resource: 'csg235f3d99e07fx472... Storage account'. To the right of the resources is a 'Quickstart tutorials' section with links to 'Windows Virtual Machines', 'Linux Virtual Machines', 'App Service', 'Functions', and 'SQL Database'. At the bottom of the dashboard, there's a terminal window with the following command:

```
nath_shubha@Azure:~$ az webapp deployment source config-local-git --name azappshu001 --resource-group myResourceGroup --query url --output tsv  
https://shudem011@azappshu001.scm.azurewebsites.net/azappshu001.git  
nath_shubha@Azure:~$
```

Ref: <https://docs.microsoft.com/en-us/azure/app-service-web/app-service-web-get-started-python>

Push to Azure from Git: Add an Azure remote to your local Git repository.

```
x root@shubha-OptiPlex-9020: /home/shubha/python-docs-hello-world
root@shubha-OptiPlex-9020:/home/shubha/python-docs-hello-world# git remote add azure https://shudemo11@azap
pshu001.scm.azurewebsites.net/azappshu001.git
root@shubha-OptiPlex-9020:/home/shubha/python-docs-hello-world# █
```

Ref: <https://docs.microsoft.com/en-us/azure/app-service-web/app-service-web-get-started-python>

Push to the Azure remote to deploy your app. You are prompted for the password you created earlier when you created the deployment user. Make sure that you enter the password you created in Configure a deployment user, not the password you use to log in to the Azure portal.

```
root@shubha-OptiPlex-9020: /home/shubha/python-docs-hello-world# git push azure master
Counting objects: 18, done.
Delta compression using up to 4 threads.
Compressing objects: 100% (16/16), done.
Writing objects: 100% (18/18) 4.31 KiB | 0 bytes/s, done.
Total 18 (delta 0)
remote: Updating branch 'master'.
remote: Updating submodules.
remote: Preparing deployment for commit id '44e74fe7dd'.
remote: Generating deployment script...
remote: Creating deployment package 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: Publishing Name: 'teststartstart.html'
remote: Copying file: '.gitignore'
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.
remote: Detecting Python runtime from site configuration
remote: Detected python 3.4
remote: Creating python 3.4 virtual environment.
remote: Pip install requirements...
remote: Downloading/unpacking Flask==0.12.1 (from requirements.txt (line 1))
remote: Downloading/unpacking itsdangerous==0.21 (from Flask==0.12.1-> r requirements.txt (line 1))
remote: Running setup.py (path:D:\home\site\wwwroot\env\build\itsdangerous\setup.py) egg_info for package itsdangerous
remote: 
remote:   warning: no previously-included files matching '*' found under directory 'docs\_build'
remote: Downloading/unpacking Jinja2==2.4 (from Flask==0.12.1-> r requirements.txt (line 1))
remote: Downloading/unpacking click==2.0 (from Flask==0.12.1-> r requirements.txt (line 1))
remote: Downloading/unpacking Werkzeug==0.7 (from Flask==0.12.1-> r requirements.txt (line 1))
remote: Downloading/unpacking MarkupSafe==0.23 (from Jinja2==2.4-> r requirements.txt (line 1))
remote: Downloading MarkupSafe-1.0.tar.gz
remote: Running setup.py (path:D:\home\site\wwwroot\env\build\MarkupSafe\setup.py) egg_info for package MarkupSafe
remote: 
remote: Installing collected packages: Flask, itsdangerous, Jinja2, click, Werkzeug, MarkupSafe
remote: Running setup.py install for itsdangerous
remote:
```

Ref: <https://docs.microsoft.com/en-us/azure/app-service-web/app-service-web-get-started-python>

Browse to the app at azappshu001.azurewebsites.net



Ref: <https://docs.microsoft.com/en-us/azure/app-service-web/app-service-web-get-started-python>

Update and redeploy the code

```
root@shubha-OptiPlex-9020: /home/shubha/python-docs-hello-world
root@shubha-OptiPlex-9020: /home/shubha/python-docs-hello-world# nano main.py
```

NPTEL

Ref: <https://docs.microsoft.com/en-us/azure/app-service-web/app-service-web-get-started-python>



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

Using a local text editor, open the main.py file in the Python app, and make a small change

```
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()

^C Get Help      ^O Write Out     ^W Where Is      ^K Cut Text      ^J Justify      ^C Cur Pos      ^Y Prev Page
^X Exit         ^R Read File     ^\ Replace       ^U Uncut Text    ^T To Linter    ^G Go To Line   ^V Next Page
```

Ref: <https://docs.microsoft.com/en-us/azure/app-service-web/app-service-web-get-started-python>

Commit your 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 file changed, 1 insertion(+), 1 deletion(-)
root@shubha-OptiPlex-9020: /home/shubha/python-docs-hello-world#
```

Ref: <https://docs.microsoft.com/en-us/azure/app-service-web/app-service-web-get-started-python>

Push the code changes to Azure

```
root@shubha-OptiPlex-9020: /home/shubha/python-docs-hello-world
root@shubha-OptiPlex-9020:/home/shubha/python-docs-hello-world# git push azure master
Password for 'https://shudemo11@azappshu001.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), 396 bytes | 0 bytes/s, done.
Total 3 (delta 1), reused 0 (delta 0)
remote: Updating branch 'master'.
remote: Updating submodules.
remote: Preparing deployment for commit id '17a51436e4'.
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: Found compatible 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
```

Ref: <https://docs.microsoft.com/en-us/azure/app-service-web/app-service-web-get-started-python>



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

Once deployment has completed, refresh the page
azappshu001.azurewebsites.net



Ref: <https://docs.microsoft.com/en-us/azure/app-service-web/app-service-web-get-started-python>

References

1. <https://docs.microsoft.com/en-us/azure/app-service-web/app-service-web-get-started-python>



IIT KHARAGPUR



NPTEL
ONLINE
CERTIFICATION COURSES

Thank You!!

NPTEL



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

Google Cloud Platform (GCP)

Prof. Soumya K Ghosh

Department of Computer Science and Engineering

IIT KHARAGPUR

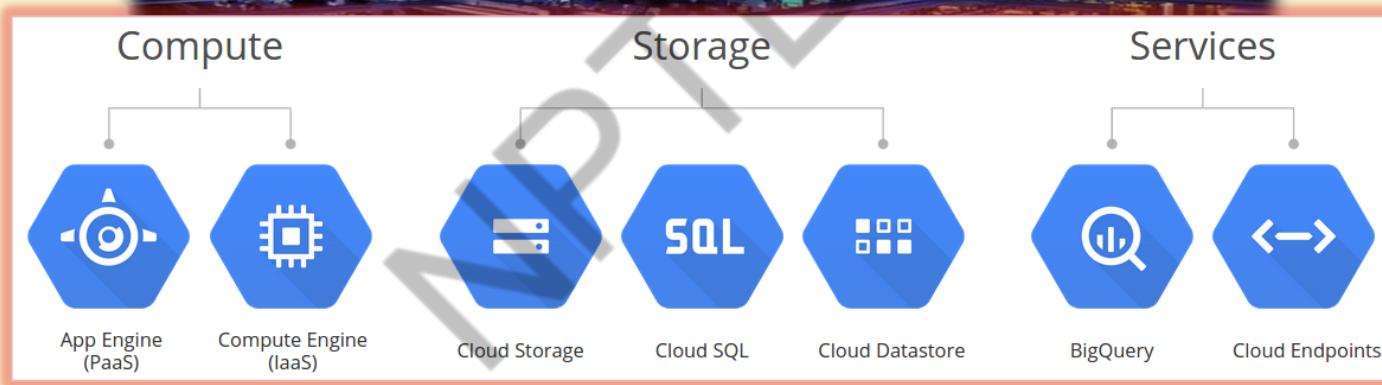
What's Google Cloud Platform?

- **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



Google Cloud Platform

Google Cloud Platform Services!



IIT KHARAGPUR



NPTEL
ONLINE
CERTIFICATION COURSES

Why Google Cloud Platform?

Run on Google's Infrastructure

Build on the same infrastructure that allows Google to return billions of search results in milliseconds, serve 6 billion hours of YouTube video per month and provide storage for 425 million Gmail users.

- ✓ Global Network
- ✓ Redundancy
- ✓ Innovative Infrastructure

Why 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

Why 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
NPTEL ONLINE
CERTIFICATION COURSES

Why 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.

Why Google Cloud Platform? (contd..)

Performance you can count on

Google's compute infrastructure gives you consistent CPU, memory and disk performance. The network and edge cache serve responses rapidly to your users across the world.

- ✓ CPU, Memory and Disk
- ✓ Global Network
- ✓ Transparent maintenance

Why Google Cloud Platform? (contd..)

Get the support you need

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.



IIT KHARAGPUR



NPTEL
ONLINE
CERTIFICATION COURSES

Google Cloud Platform Services

The diagram illustrates the Google Cloud Platform services. On the left, a red box is titled "Compute" and contains two items: "Compute Engine" with a corresponding icon and "App Engine" with a corresponding icon. A blue circle highlights the "Compute Engine" item. To the right of this red box is a white area containing three numbered points:

- I. Cloud Platform offers both a fully managed platform and flexible virtual machines, allowing you to choose a system that meets your needs.
- II. Use App Engine, a Platform-as-a-Service, when you just want to focus on your code and not worry about patching or maintenance.
- III. Get access to raw virtual machines with Compute Engine and have the flexibility to build anything you need.

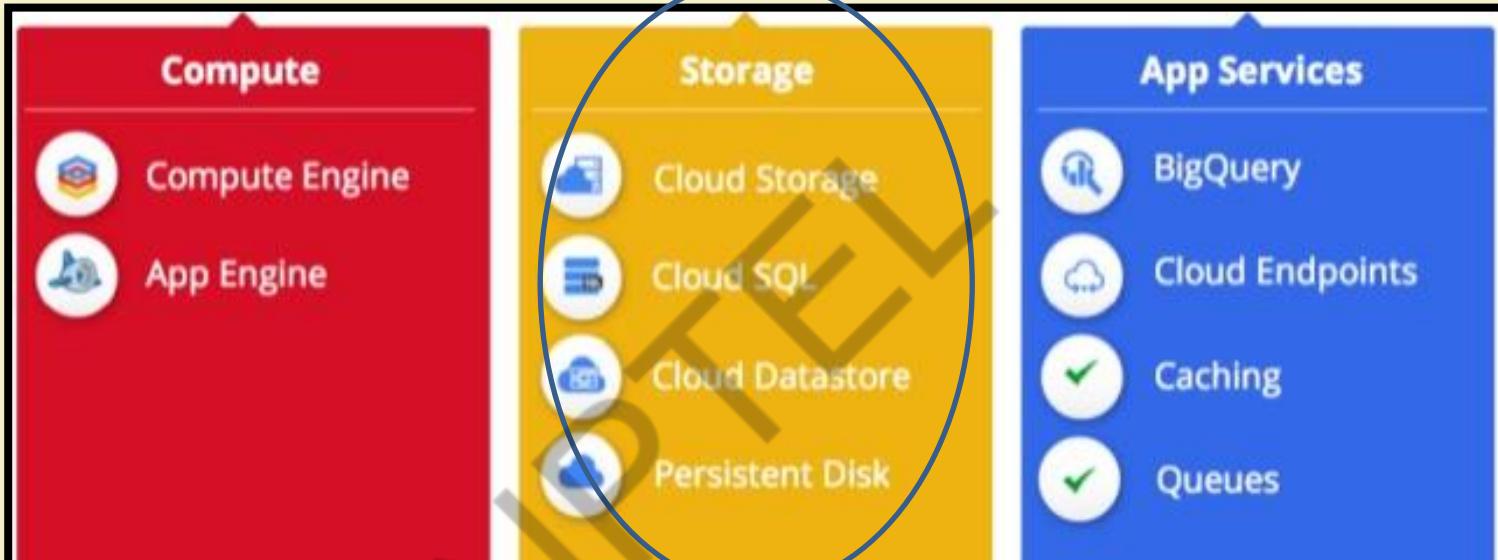


IIT KHARAGPUR



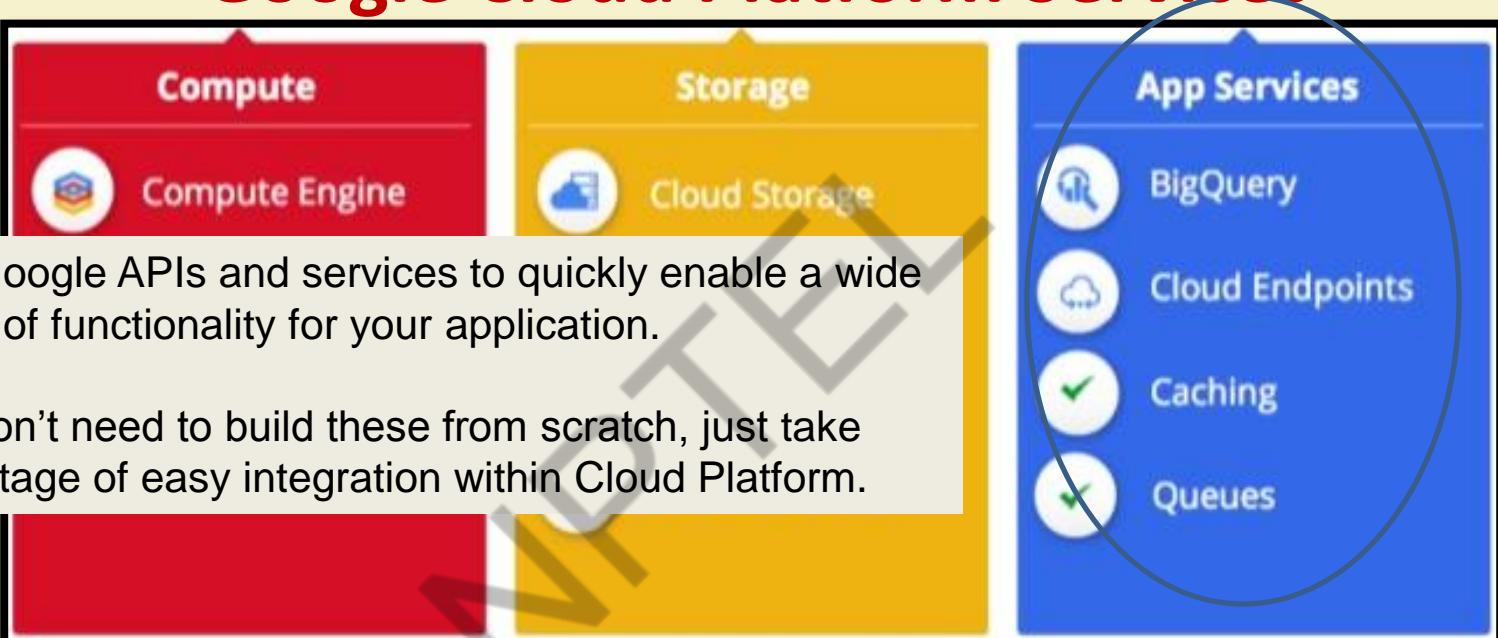
NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Google Cloud Platform Services



- I. Google Cloud Platform provides a range of storage services that allow you to maintain easy and quick access to your data.
- II. With **Cloud SQL** and **Datastore** you get MySQL or NoSQL databases, while **Cloud Storage** provides flexible object storage with global edge caching.

Google Cloud Platform Services



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**.

*Example 1: Host your **web-page** in Google Cloud Platform*

*Example 2: Build your **web-app** using Google App Engine*

*Example 1: Host your **web-page** in Google Cloud Platform*



IIT KHARAGPUR

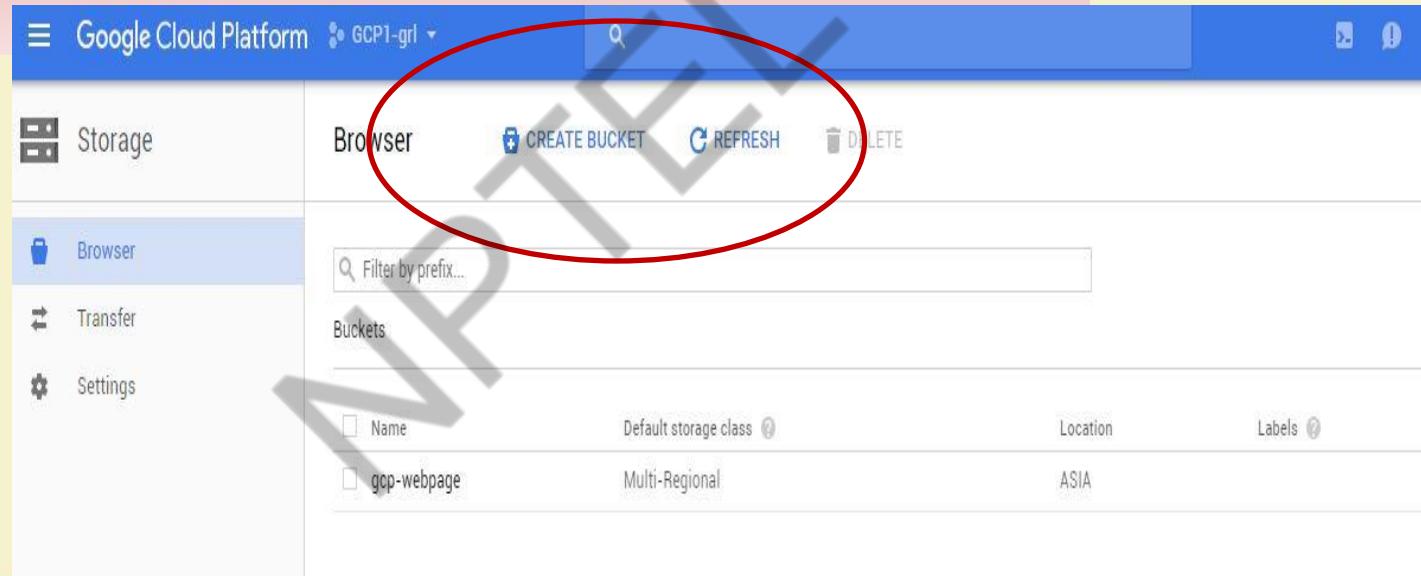


NPTEL

NPTEL ONLINE
CERTIFICATION COURSES

An easy example: Host your *web-page* inside Google Cloud Platform

- i) Open the Cloud Storage browser in the Google Cloud Platform Console & click on Create Bucket



An easy example: Host your *web-page* inside Google Cloud Platform

ii) In the list of buckets, find the bucket you created.
And Click the more actions icon next to the bucket and select **Edit configuration**.

The screenshot shows the Google Cloud Storage Browser interface. On the left, there's a sidebar with 'Storage' selected, followed by 'Browser' (which is highlighted in blue), 'Transfer', and 'Settings'. The main area is titled 'Browser' and contains buttons for 'CREATE BUCKET', 'REFRESH', and 'DELETE'. A 'SHOW INFO PANEL' button is on the far right. Below these are sections for 'Filter by prefix...', 'Buckets', and a table with columns for 'Name', 'Default storage class', 'Location', and 'Labels'. A red oval highlights the row for the bucket 'gcp-webpage', which has 'Multi-Regional' as its storage class and 'ASIA' as its location. To the right of the table, a vertical ellipsis menu is open, showing options: 'Edit bucket permissions', 'Edit labels', and 'Edit default storage class'. A large, semi-transparent 'NPTEL' watermark is overlaid across the entire interface.

Name	Default storage class	Location	Labels
gcp-webpage	Multi-Regional	ASIA	

An easy example: Host your *web-page* inside Google Cloud Platform

iii) In the **Configure website** dialog, specify the **Main Page** and the **404 (Not Found) Page** or even your web-site folder!

The screenshot shows the Google Cloud Platform Storage Browser interface. On the left is a sidebar with 'Storage' selected, followed by 'Browser', 'Transfer', and 'Settings'. The main area is titled 'Browser' and shows a list of files under 'Buckets / gcp-webpage / GCP-Webpage'. A red oval highlights the top navigation bar with 'UPLOAD FILES', 'UPLOAD FOLDER', 'CREATE FOLDER', and 'REFRESH'. A red bracket on the left side groups several files: '404.html', 'bin/', 'cc1.html', 'css/', 'figures/', 'font-awesome/', 'fonts/', 'index.html', 'index1.html', 'js/', 'LICENSE', and 'README.md'. A callout bubble points to this group with the text 'Upload all files/ figures of your web-site!'. Another callout bubble on the right side points to the 'Share publicly' column with the text 'Check whether all are shared publicly!'. The table below lists the files:

Name	Size	Type	Storage class	Last modified	Share publicly
404.html	9.15 KB	text/html	Multi-Regional	7/20/17, 12:37 AM	<input checked="" type="checkbox"/> Public link
bin/	—	Folder	—	—	
cc1.html	6.26 KB	text/html	Multi-Regional	7/20/17, 12:37 AM	<input checked="" type="checkbox"/> Public link
css/	—	Folder	—	—	
figures/	—	Folder	—	—	
font-awesome/	—	Folder	—	—	
fonts/	—	Folder	—	—	
index.html	12.81 KB	text/html	Multi-Regional	7/20/17, 12:37 AM	<input checked="" type="checkbox"/> Public link
index1.html	10.49 KB	text/html	Multi-Regional	7/20/17, 12:37 AM	<input checked="" type="checkbox"/> Public link
js/	—	Folder	—	—	
LICENSE	1.07 KB	application/octet-stream	Multi-Regional	7/20/17, 12:37 AM	<input checked="" type="checkbox"/> Public link
README.md	1.64 KB	application/octet-stream	Multi-Regional	7/20/17, 12:37 AM	<input checked="" type="checkbox"/> Public link

An easy example: Host your *web-page* inside Google Cloud Platform

iv) Get the public link of your html of home-page or *index.html*

The screenshot shows the Google Cloud Platform Storage Browser interface. On the left, there's a sidebar with 'Storage' selected, followed by 'Browser', 'Transfer', and 'Settings'. The main area is titled 'Buckets / gcp-webpage / GCP-Webpage'. It lists several files and folders:

Name	Size	Type	Storage class	Last modified	Share publicly
404.html	9.15 KB	text/html	Multi-Regional	7/20/17, 12:37 AM	<input checked="" type="checkbox"/> Public link
bin/	—	Folder	—	—	
cc1.html	6.26 KB	text/html	Multi-Regional	7/20/17, 12:37 AM	<input checked="" type="checkbox"/> Public link
css/	—	Folder	—	—	
figures/	—	Folder	—	—	
font-awesome/	—	Folder	—	—	
fonts/	—	Folder	—	—	
index.html	12.81 KB	text/html	Multi-Regional	7/20/17, 12:37 AM	<input checked="" type="checkbox"/> Public link
index1.html	10.49 KB	text/html	Multi-Regional	7/20/17, 12:37 AM	<input checked="" type="checkbox"/> Public link
js/	—	Folder	—	—	
LICENSE	1.07 KB	application/octet-stream	Multi-Regional	7/20/17, 12:37 AM	<input checked="" type="checkbox"/> Public link
README.md	1.64 KB	application/octet-stream	Multi-Regional	7/20/17, 12:37 AM	<input checked="" type="checkbox"/> Public link

And you are ready to go! ☺



<https://storage.googleapis.com/gcp-webpage/GCP-Webpage/index1.html>

Hi there!

Home Summary ▾

Data and Computing : Up in the Cloud!

Welcome to Cloud Computing NPTEL Course!

✓ About this Course!

This course will introduce various aspects of cloud computing, including fundamentals, management issues, security challenges and future research trends. This will help students (both UG and PG levels) and researchers to use and explore the cloud computing platforms.

Course PRE-REQUISITES & Suggested Reading

Course Pre-requisites:

- Basics of Computer Architecture and organization
- Networking

Course Instructor & Certification

Taught by: Prof. Soumya K Ghosh, Dept. of CSE, IIT Khargpur

Certification Exam: Exams will be on 22 October 2017. Time: Shift 1: 9am-12 noon; Shift 2: 2pm-5pm. Final score will be calculated as 100% assignment.

*Example 2: Build your *web-app* using *Google App Engine**



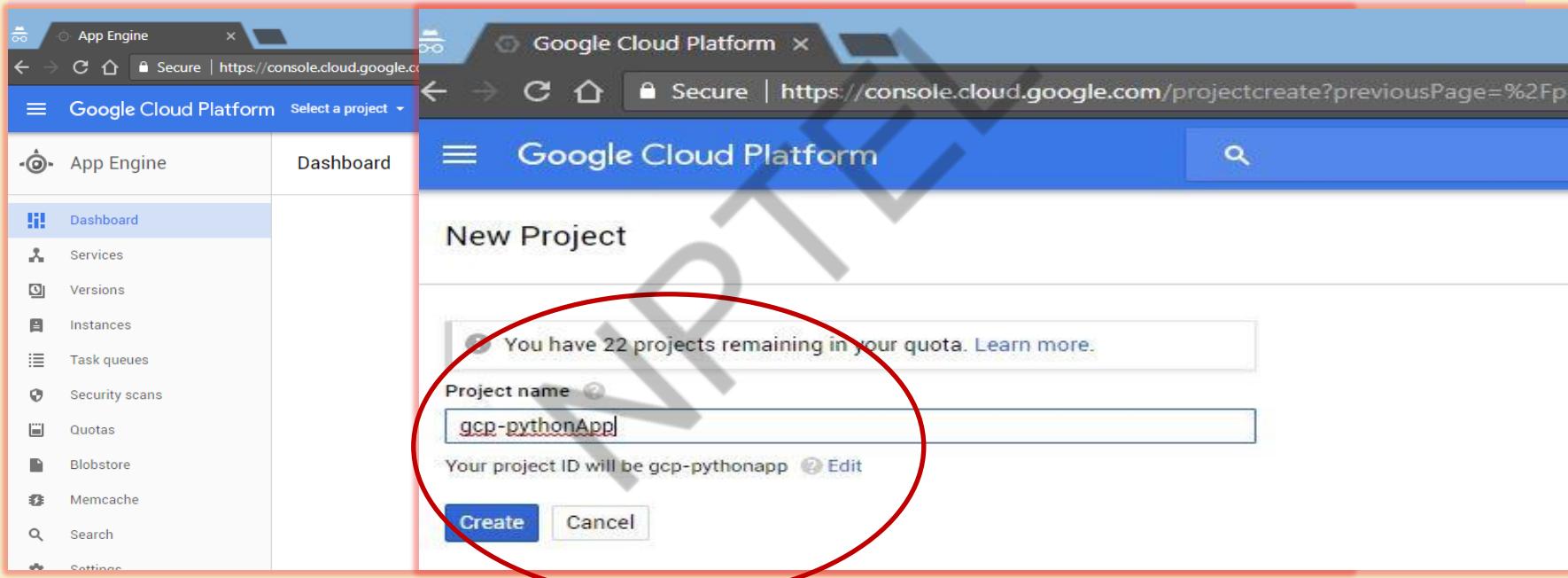
IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

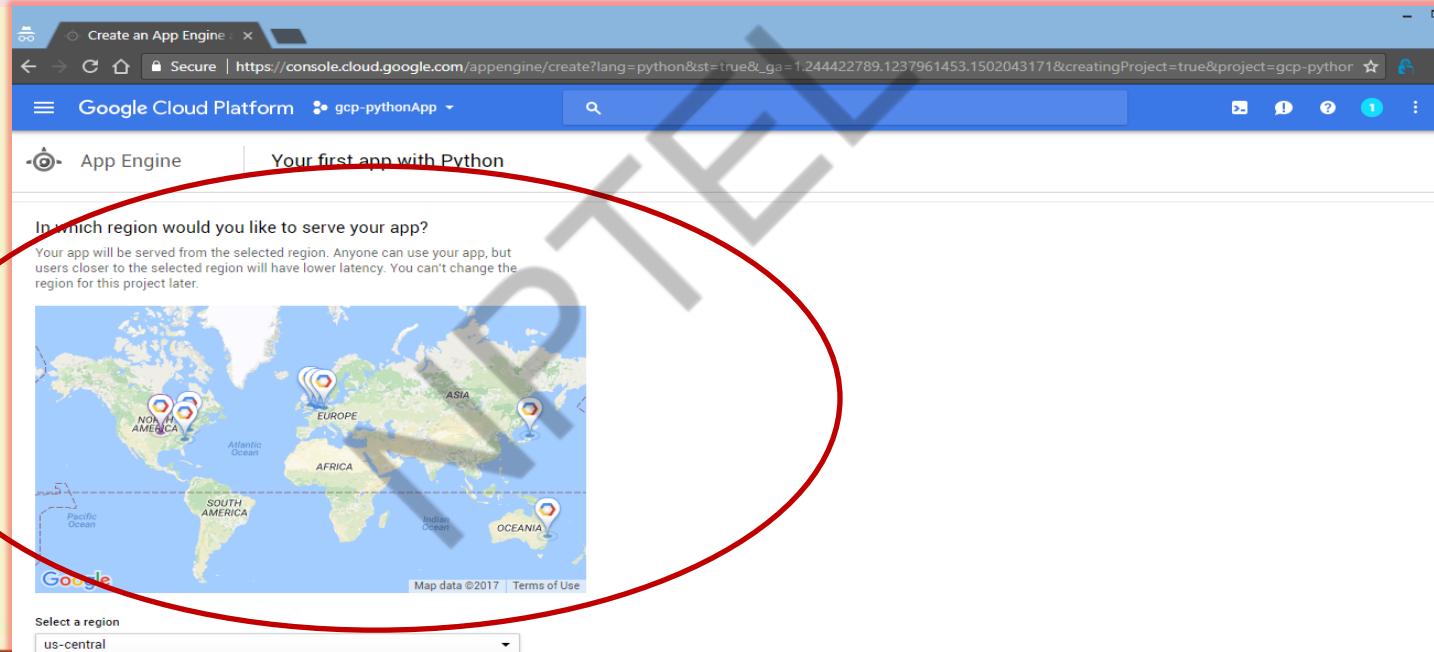
Another example: Host your *web-app* using *Google App Engine*

- i) Open the Google Cloud Platform Console & create a new project using *Cloud Platform project and App Engine application*



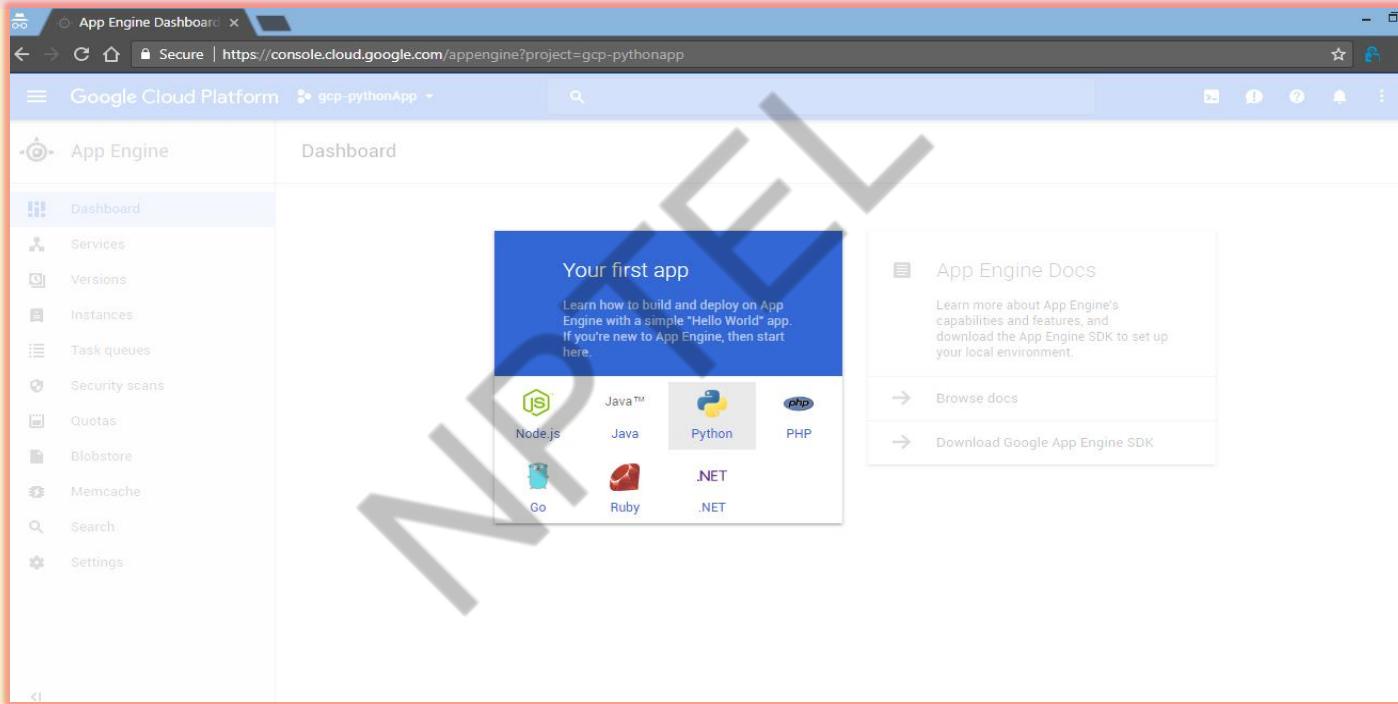
Another example: Host your *web-app* using *Google App Engine*

- ii) When prompted, select the **region** where you want your App Engine application located.



Another example: Host your *web-app* using *Google App Engine*

iii) Select your preferred programming language to build your app.



iv) Activate your ***Google Cloud Shell*** .

The screenshot shows the Google Cloud Platform App Engine Dashboard. A red callout box highlights the "Cloud Shell" tab in the top navigation bar. A red circle is drawn around the "Cloud Shell" tab in the dashboard interface. A green box contains the text "Google Cloud Shell will appear". A message box in the center says "... Connecting: Provisioning your Google Cloud Shell machine...". Below it, a terminal window titled "gcp-pythonapp" shows the welcome message: "Welcome to Cloud Shell! Type \"help\" to get started." and the prompt "sq_researchwork@gcp-pythonapp:~\$". In the bottom right corner of the dashboard, there are "CANCEL TUTORIAL" and "SEND FEEDBACK" buttons.

Cloud Shell x +

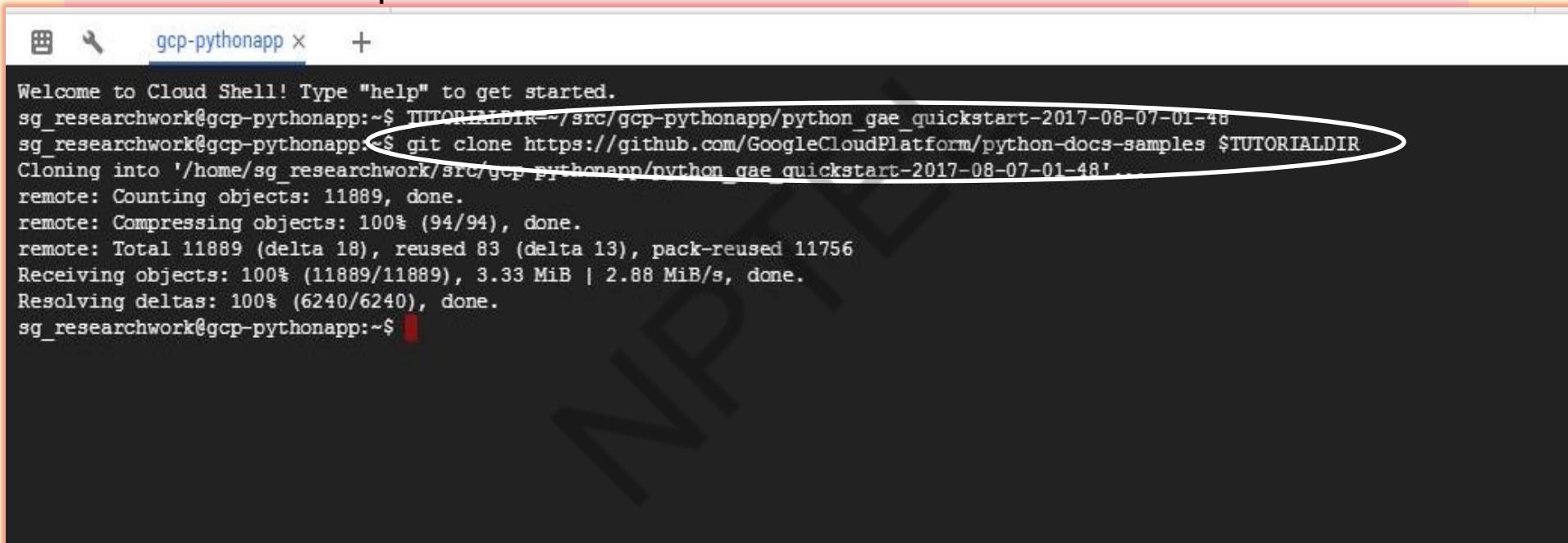
gcp-pythonapp x +

... Connecting: Provisioning your Google Cloud Shell machine...

Welcome to Cloud Shell! Type "help" to get started.
sq_researchwork@gcp-pythonapp:~\$

Cancel Tutorial Send Feedback

v) Clone the Hello World sample app repository and go to the directory that contains the sample code



```
Welcome to Cloud Shell! Type "help" to get started.  
sg_researchwork@gcp-pythonapp:~$ TUTORIALDIR=~/src/gcp-pythonapp/python_gae_quickstart-2017-08-07-01-48  
sg_researchwork@gcp-pythonapp:~$ git clone https://github.com/GoogleCloudPlatform/python-docs-samples $TUTORIALDIR  
Cloning into '/home/sg_researchwork/src/gcp-pythonapp/python_gae_quickstart-2017-08-07-01-48'.  
remote: Counting objects: 11889, done.  
remote: Compressing objects: 100% (94/94), done.  
remote: Total 11889 (delta 18), reused 83 (delta 13), pack-reused 11756  
Receiving objects: 100% (11889/11889), 3.33 MiB | 2.88 MiB/s, done.  
Resolving deltas: 100% (6240/6240), done.  
sg_researchwork@gcp-pythonapp:~$
```

v) Each application must contain ‘app.yaml’ and code base ‘main.py’ [with Flask web app deployment]

```
gcp-pythonapp x +  
sg_research  
sg_research  
runtime: python  
api_version:  
threadsafe:  
  
handlers:  
- url: /.*  
  script: MainPage  
sg_research  
  
# 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)  
sg_researchwork@gcp-pythonapp:~/src/gcp-pythonapp/python_gae_quickstart-2017-08-07-01-48/appengine/standard/hello_world$
```



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

vi) From within the hello_world directory where the app's app.yaml configuration file is located, start the *local development server* :
dev_appserver.py \$PWD



The screenshot shows a terminal window titled "gcp-pythonapp" with the command `dev_appserver.py $PWD` highlighted by a white oval. The terminal output shows the following log messages:

```
sg_researchwork@gcp-pythonapp:~/src/gcp-pythonapp/python_gae_quickstart-2017-08-01-48/appengine/standard/hello_world$ dev_appserver.py $PWD
INFO    2017-08-06 20:26:05,853 devappserver2.py:116] Skipping SDK update check.
WARNING 2017-08-06 20:26:06,452 simple_search_stub.py:116] Could not read search indexes from /tmp/appengine.None.sg_researchwork/search_indexes
INFO    2017-08-06 20:26:06,454 api_server.py:313] Starting API server at: http://0.0.0.0:54678
WARNING 2017-08-06 20:26:06,454 dispatcher.py:287] Your python27 micro version is below 2.7.12, our current production version.
INFO    2017-08-06 20:26:06,457 dispatcher.py:226] Starting module "default" running at: http://0.0.0.0:8080
INFO    2017-08-06 20:26:06,457 admin_server.py:116] Starting admin server at: http://0.0.0.0:8000
```

Visit in your web browser to view the app

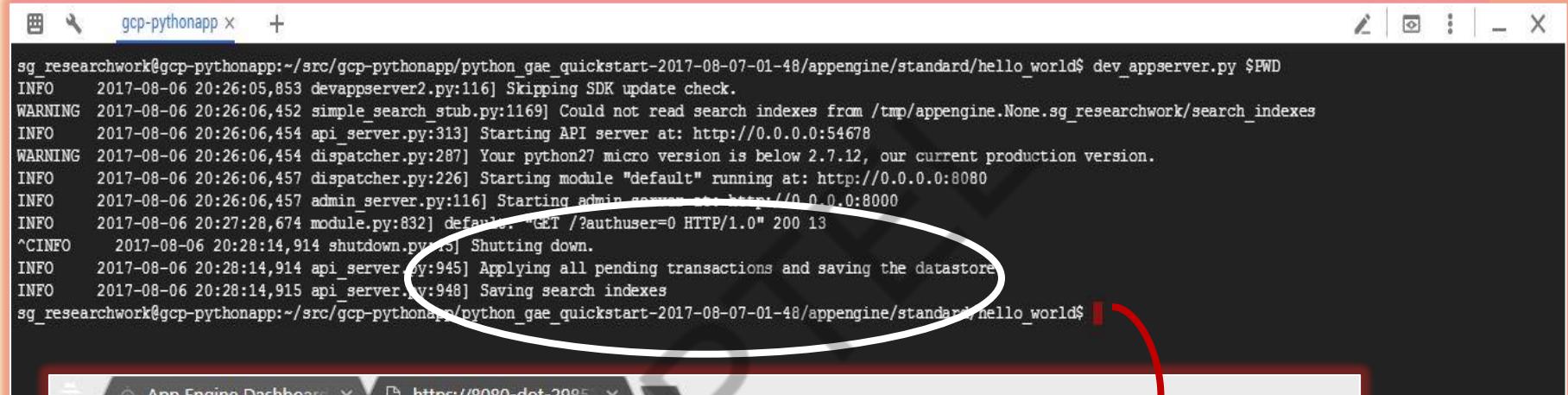
The screenshot shows a terminal window at the top and a browser window below it. A red circle highlights the 'Web preview' icon in the terminal's toolbar, and a red arrow points from the terminal window to the browser window, indicating the connection between them.

sq_researchwork@gcp-pythonapp:~/src/gcp-pythonapp/python_gae_quickstart-2017-08-07-01-48/appengine/standard/hello_world\$ dev_appserver.py \$PWD
INFO 2017-08-06 20:26:05,853 devappserver2.py:116] Skipping SDK update check.
WARNING 2017-08-06 20:26:06,452 simple_search_stub.py:116] Could not read search indexes from /tmp/appengine.None.sg_researchwork/search_indexes
INFO 2017-08-06 20:26:06,454 api_server.py:313] Starting API server at: http://0.0.0.0:54678
WARNING 2017-08-06 20:26:06,454 dispatcher.py:287] Your python27 micro version is below 2.7.12, our current production version.
INFO 2017-08-06 20:26:06,457 dispatcher.py:226] Starting module "default" running at: http://0.0.0.0:8080
INFO 2017-08-06 20:26:06,457 admin_server.py:116] Starting admin server at: http://0.0.0.0:8000

Secure | https://8080-dot-2985339-dot-devshell.appspot.com/?authuser=0

Hello, World!

You can shut-down the development server at any point!

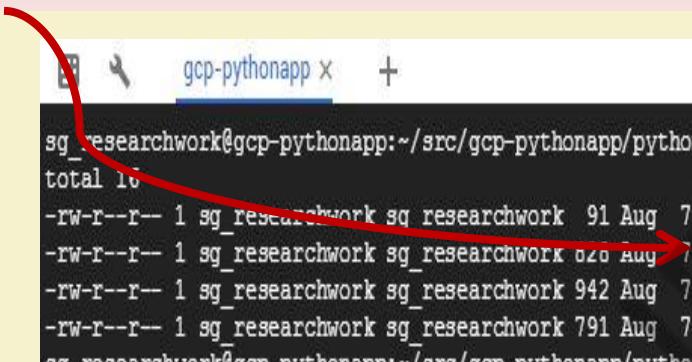


```
sg_researchwork@gcp-pythonapp:~/src/gcp-pythonapp/python_gae_quickstart-2017-08-07-01-48/appengine/standard/hello_world$ dev_appserver.py $PWD
INFO    2017-08-06 20:26:05,853 devappserver2.py:116] Skipping SDK update check.
WARNING 2017-08-06 20:26:06,452 simple_search_stub.py:1169] Could not read search indexes from /tmp/appengine.None.sg_researchwork/search_indexes
INFO    2017-08-06 20:26:06,454 api_server.py:313] Starting API server at: http://0.0.0.0:54678
WARNING 2017-08-06 20:26:06,454 dispatcher.py:287] Your python27 micro version is below 2.7.12, our current production version.
INFO    2017-08-06 20:26:06,457 dispatcher.py:226] Starting module "default" running at: http://0.0.0.0:8080
INFO    2017-08-06 20:26:06,457 admin_server.py:116] Starting admin server at: http://0.0.0.0:8000
INFO    2017-08-06 20:27:28,674 module.py:832] default: "GET /?authuser=0 HTTP/1.0" 200 13
^CINFO    2017-08-06 20:28:14,914 shutdown.py:15] Shutting down.
INFO    2017-08-06 20:28:14,914 api_server.py:945] Applying all pending transactions and saving the datastore
INFO    2017-08-06 20:28:14,915 api_server.py:948] Saving search indexes
sg_researchwork@gcp-pythonapp:~/src/gcp-pythonapp/python_gae_quickstart-2017-08-07-01-48/appengine/standard/hello_world$
```



You can leave the development server running while you develop your application. The development server watches for changes in your source files and reloads them if necessary

Edit main.py



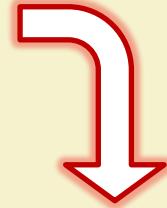
```
gcp-pythonapp x +  
sg_researchwork@gcp-pythonapp:~/src/gcp-pythonapp/python_gae_quickstart-2017-08-07-01-48/appengine/standard/hello_world$ ls -l  
total 16  
-rw-r--r-- 1 sg_researchwork sg_researchwork 91 Aug  7 01:51 app.yaml  
-rw-r--r-- 1 sg_researchwork sg_researchwork 626 Aug  7 01:55 main.py  
-rw-r--r-- 1 sg_researchwork sg_researchwork 942 Aug  7 01:57 main.pyc  
-rw-r--r-- 1 sg_researchwork sg_researchwork 791 Aug  7 01:51 main_test.py  
sg_researchwork@gcp-pythonapp:~/src/gcp-pythonapp/python_gae_quickstart-2017-08-07-01-48/appengine/standard/hello_world$
```

Edit main.py

```
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)
```



```
import webapp2

class MainPage(webapp2.RequestHandler):
    def get(self):
        self.response.headers['Content-Type'] = 'text/plain'
        self.response.write('Hi! Welcome to NPTEL Cloud Computing Course\nHappy Learning!! :)')

app = webapp2.WSGIApplication([
    ('/', MainPage),
], debug=True)
```

Reload the web-page

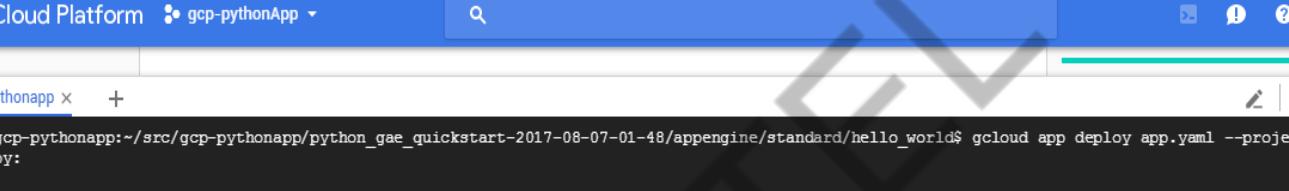


IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

Now deploy your app to App Engine : ***gcloud app deploy app.yaml --project gcp-pythonapp***



```
sg_researchwork@gcp-pythonapp:~/src/gcp-pythonapp/python_gae_quickstart-2017-08-01-48/appengine/standard/hello_world$ gcloud app deploy app.yaml --project gcp-pythonapp
Services to deploy:
descriptor:      [/home/sg_researchwork/src/gcp-pythonapp/python_gae_quickstart-2017-08-01-48/appengine/standard/hello_world/app.yaml]
source:          [/home/sg_researchwork/src/gcp-pythonapp/python_gae_quickstart-2017-08-01-48/appengine/standard/hello_world]
target project: [gcp-pythonapp]
target service: [default]
target version: [20170807t020530]
target url:     [https://gcp-pythonapp.appspot.com]

Do you want to continue (Y/n)? Y

Beginning deployment of service [default]...
Some files were skipped. Pass `--verbosity=info` to see which ones.
You may also view the gcloud log file, found at
[~/tmp/tmp.YLV7NzHY4B/logs/2017.08.07/02.05.25.079263.log].
[Uploading 5 files to Google Cloud Storage] File upload done.
Updating service [default]...
```



Now deploy your app to App Engine : *gcloud app deploy app.yaml --project gcp-pythonapp*

```
sg_researchwork@gcp-pythonapp:~/src/gcp-pythonapp/python_gae_quickstart-2017-08-07-01-48/appengine/standard/hello_world$ gcloud app deploy app.yaml --project gcp-pythonapp
Services to deploy:
descriptor: [/home/sg_researchwork/src/gcp-pythonapp/python_gae_quickstart-2017-08-07-01-48/appengine/standard/hello_world/app.yaml]
source: [/home/sg_researchwork/src/gcp-pythonapp/python_gae_quickstart-2017-08-07-01-48/appengine/standard/hello_world]
target project: [gcp-pythonapp]
target service: [default]
target version: [20170807t020530]
target url: [https://gcp-pythonapp.appspot.com]

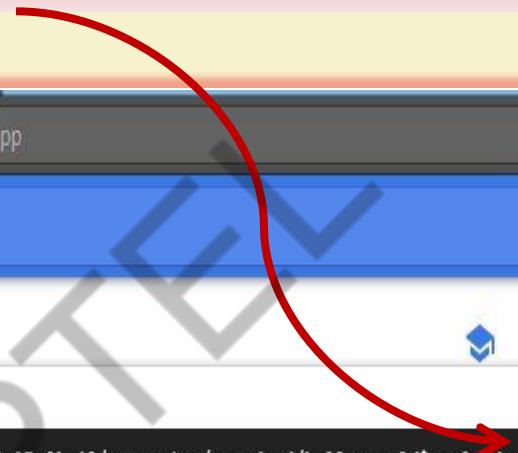
Do you want to continue (Y/n)? Y

Beginning deployment of service [default]...
Some files were skipped. Pass `--verbosity=info` to see which ones.
You may also view the gcloud log file, found at
[~/tmp/tmp.YLV7NzHY4B/logs/2017.08.07/02.05.25.079263.log].
[ Uploading 5 files to Google Cloud Storage ] File upload done.
Updating service [default]...done.
Waiting for operation [apps/gcp-pythonapp/operations/891c8591-ecc1-4ac8-b5a8-a3358c03e16a] to complete...done.
Updating service [default]...done.
Deployed service [default] to [https://gcp-pythonapp.appspot.com]

You can stream logs from the command line by running:
$ gcloud app logs tail -s default

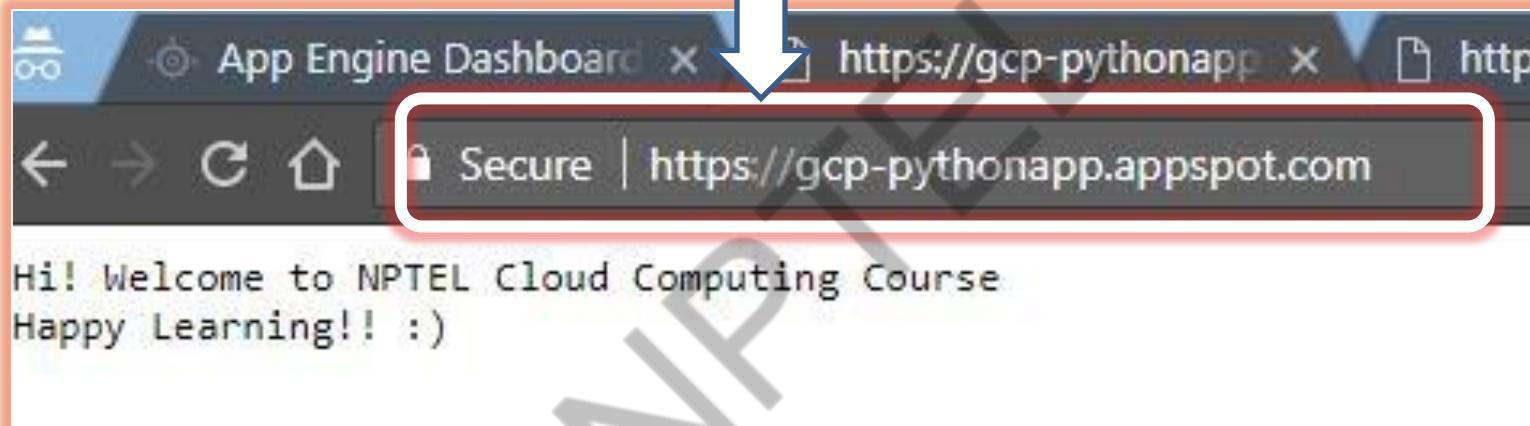
To view your application in the web browser run:
$ gcloud app browse
sg_researchwork@gcp-pythonapp:~/src/gcp-pythonapp/python_gae_quickstart-2017-08-07-01-48/appengine/standard/hello_world$
```

View your application : ***gcloud app browse***



```
Secure | https://console.cloud.google.com/appengine?project=gcp-pythonapp
Google Cloud Platform gcp-pythonApp
App Engine Dashboard
gcp-pythonapp X +
sg_researchwork@gcp-pythonapp:~/src/gcp-pythonapp/python_gae_quickstart-2017-08-07-01-48/appengine/standard/hello_world$ gcloud app browse
Did not detect your browser. Go to this link to view your app:
https://gcp-pythonapp.appspot.com
sg_researchwork@gcp-pythonapp:~/src/gcp-pythonapp/python_gae_quickstart-2017-08-07-01-48/appengine/standard/hello_world$
```

View your application : ***gcloud app browse***



You have successfully deployed an web-app!

The screenshot shows a software window titled "App Engine Quickstart". At the top, there are several icons: a blue square, a speech bubble, a question mark, a bell, and three vertical dots. Below the title, a large green checkmark is displayed over the text. The main content area contains the following text:

App Engine Quickstart

Congratulations

You have successfully deployed an App Engine application! Here are some next steps:

1. Download the Google Cloud SDK and develop locally

Download Cloud SDK for Windows

After it downloads, extract the file \rightarrow and initialize the SDK \rightarrow .

2. Build your next application

Learn how to use App Engine with other Cloud Platform products:



IIT KHARAGPUR



NPTEL
ONLINE
CERTIFICATION COURSES

Some Useful Links!

- Google Cloud Platform Developers Portal: <https://cloud.google.com/developers>
- Google Developers Global Portal: <https://developers.google.com>
- Google Cloud Platform Products list: <https://cloud.google.com/products/compute-engine/>
- Understanding Google APIs: <https://fethidilmi.blogspot.com/2013/01/understanding-google-apis.html>

References

- <https://cloud.google.com/storage/docs/>
- <https://cloud.google.com/why-google/>
- <https://cloud.google.com/products/>
- <http://fethidilmi.blogspot.com>
- <https://www.slideshare.net/delphiexile/google-cloud-platform-overview-28927697>

Thank You!!

NPTEL



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

CLOUD COMPUTING

SLA - Tutorial

PROF. SOUMYA K. GHOSH

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
IIT KHARAGPUR

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

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.

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
<99.95%	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.

Thank You!

NPTEL



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

Cloud Computing : Economics Tutorial

Prof. Soumya K Ghosh

Department of Computer Science and Engineering
IIT KHARAGPUR

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

9/11/2017



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

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

9/11/2017



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

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



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

9/11/2017



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

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
 - I. 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

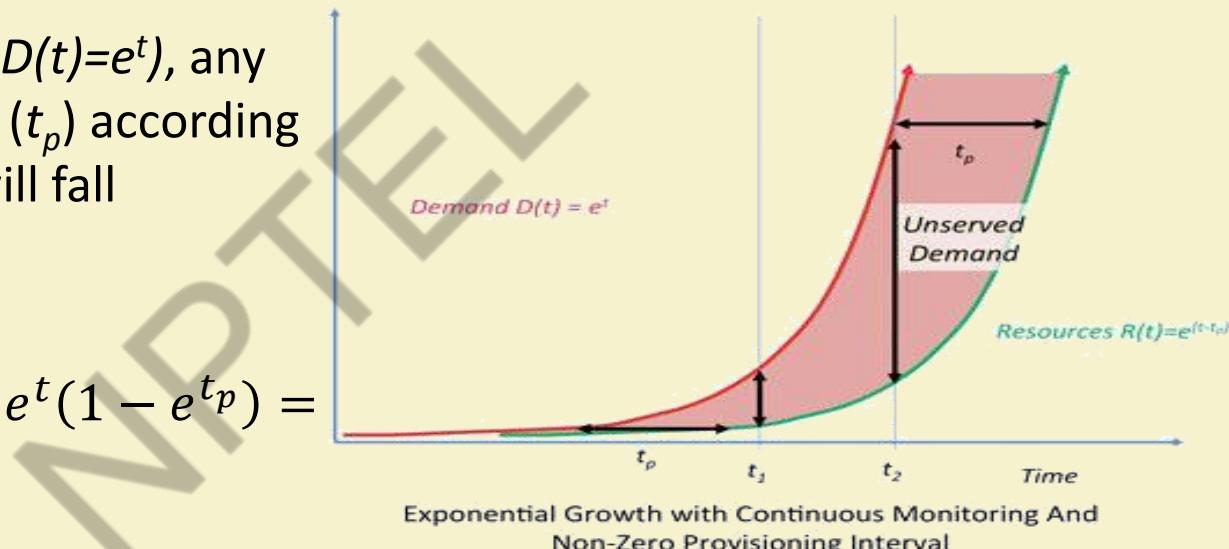
9/11/2017



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

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

9/11/2017



NPTEL
ONLINE
CERTIFICATION COURSES

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, δ 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
NPTEL ONLINE
CERTIFICATION COURSES

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

$$D(t) = 50(1 + e^{-t})$$

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?

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?

Thank You!



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

Cloud Computing : *MapReduce - Tutorial*

Prof. Soumya K Ghosh

Department of Computer Science and Engineering
IIT KHARAGPUR

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

9/11/2017



NPTEL
ONLINE
CERTIFICATION COURSES

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

9/11/2017



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

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:
$$\text{Map: } (k_1, v_1) \rightarrow [(k_2, v_2)].$$
 - 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



IIT KHARAGPUR

9/11/2017



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

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:

Reduce: $(k_2, [v_2]) \rightarrow (k_2, f([v_2])).$

- Final results are written back to the GFS file system



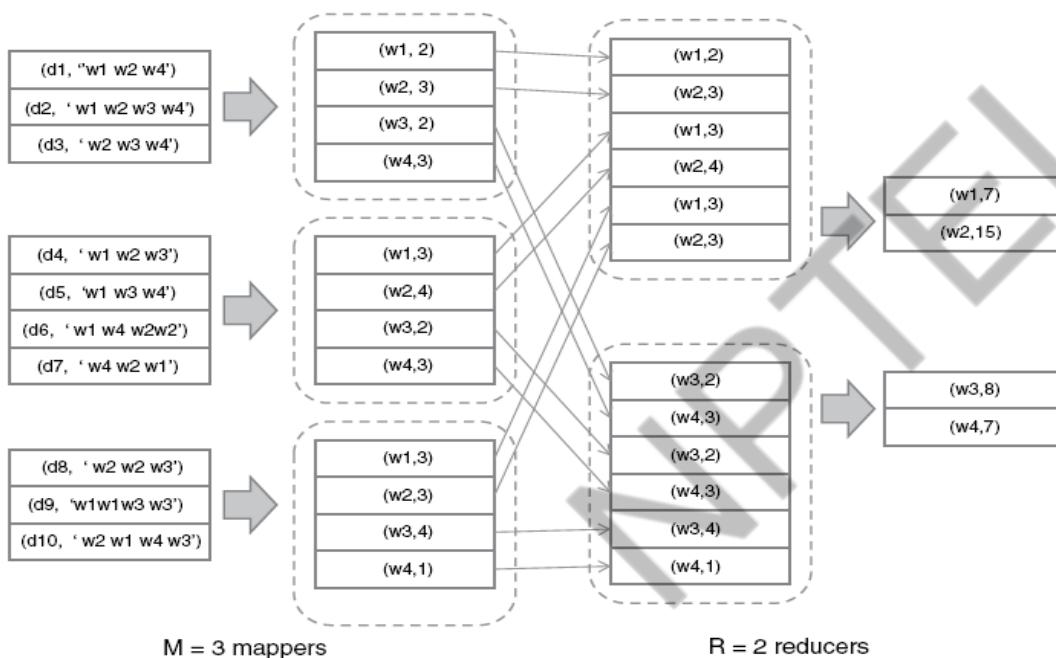
IIT KHARAGPUR

9/11/2017



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

MapReduce: Example



- 3 mappers; 2 reducers
- Map function:

$$(d_k, [w_1 \dots w_n]) \rightarrow [(w_i, c_i)].$$

- Reduce function:

$$(w_i, [c_i]) \rightarrow \left(w_i, \sum_i c_i \right)$$



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

9/11/2017



NPTEL

NPTEL ONLINE
CERTIFICATION COURSES

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

9/11/2017



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Problem-3

Compute total and average salary of organization XYZ and group by based on gender (male or female) using MapReduce. The input is as follows

Name, Gender, Salary

John, M, 10,000

Martha, F, 15,000



IIT KHARAGPUR

9/11/2017



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Problem-4

Write the *Map* and *Reduce* functions (pseudo-codes) for the following ***Word Length Categorization*** problem under *MapReduce* model.

Word Length Categorization: Given a text paragraph (containing only words), categorize each word into following categories. Output the frequency of occurrence of words in each category.

Categories:

tiny: 1-2 letters; **small:** 3-5 letters; **medium:** 6-9 letters; **big:** 10 or more letters



IIT KHARAGPUR

9/11/2017



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Thank You!

NPTEL



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

CLOUD COMPUTING

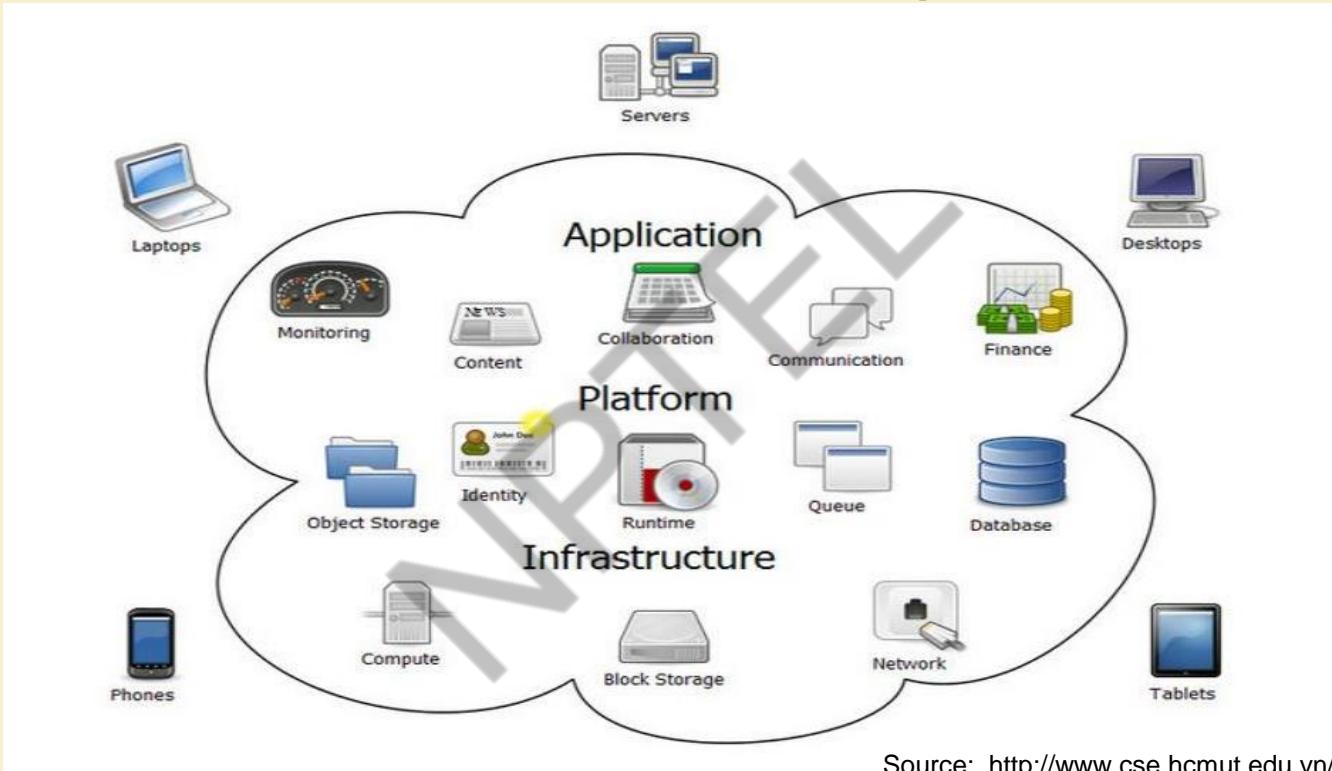
Resource Management - I

PROF. SOUMYA K. GHOSH

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

IIT KHARAGPUR

Different Resources in Computing



Source: <http://www.cse.hcmut.edu.vn/~ptvu/gc/2012/GC-pp.pdf>



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

Resources types

- **Physical resource**
 - Computer, disk, database, network, scientific instruments.
- **Logical resource**
 - Execution, monitoring, communicate application .

Source: <http://www.cse.hcmut.edu.vn/~ptvu/gc/2012/GC-pp.pdf>



IIT KHARAGPUR



NPTEL
ONLINE
CERTIFICATION COURSES

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.hcmut.edu.vn/~ptvu/gc/2012/GC-pp.pdf>



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

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 CO₂ 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,



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

Motivation for Green Data Centers

Economic

- New data centers run on the Megawatt scale, requiring millions of dollars to operate.
- Recently institutions are looking for new ways to reduce costs
- Many facilities are at their peak operating stage, and cannot expand without a new power source.

Environmental

- Majority of energy sources are fossil fuels.
- Huge volume of CO₂ emitted each year from power plants.
- Sustainable energy sources are not ready.
- Need to reduce energy dependence



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

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



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

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.

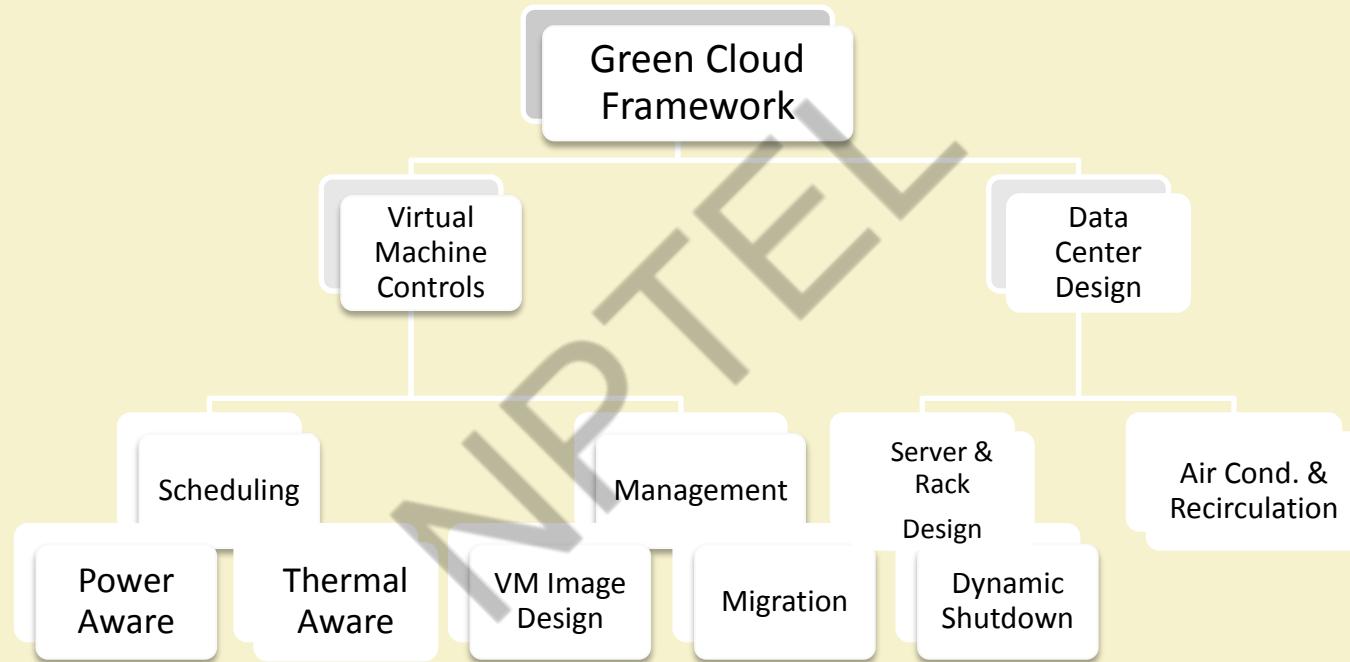


IIT KHARAGPUR



NPTEL
ONLINE
CERTIFICATION COURSES

Steps towards Energy Efficiency



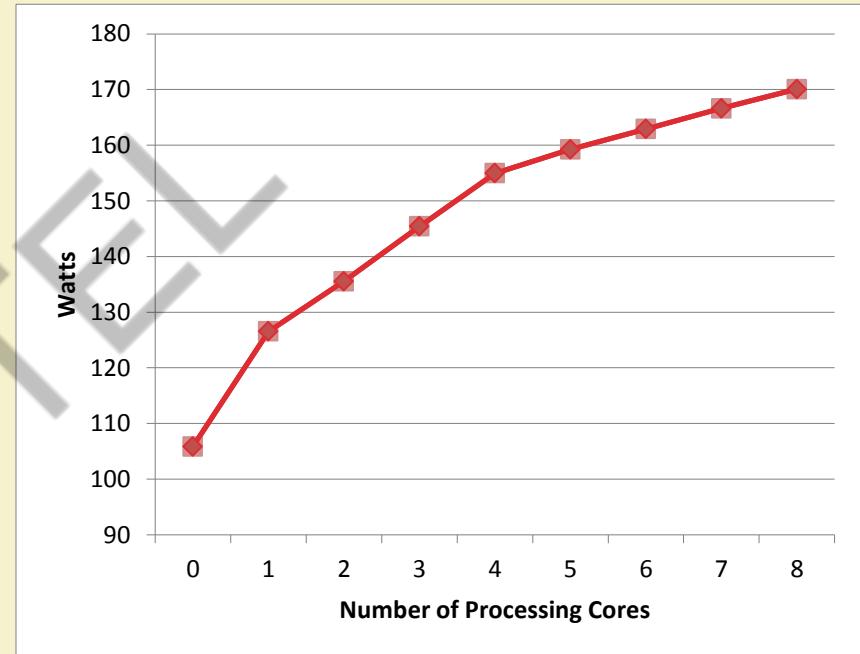
IIT KHARAGPUR



NPTEL
ONLINE
CERTIFICATION COURSES

VM scheduling on Multi-core Systems

- There is a nonlinear relationship between the number of processes used and power consumption
- We can schedule VMs to take advantage of this relationship in order to conserve power



*Power consumption curve on an Intel Core i7 920 Server
(4 cores, 8 virtual cores with Hyperthreading)*

Scheduling



IIT KHARAGPUR



NPTEL
ONLINE
CERTIFICATION COURSES

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

Scheduling

Algorithm 1 Power based scheduling of VMs

```
FOR  $i = 1$  TO  $i \leq |pool|$  DO
     $pe_i$  = num cores in  $pool_i$ 
END FOR

WHILE (true)
    FOR  $i = 1$  TO  $i \leq |queue|$  DO
         $vm = queue_i$ 
        FOR  $j = 1$  TO  $j \leq |pool|$  DO
            IF  $pe_j \geq 1$  THEN
                IF check capacity  $vm$  on  $pe_j$  THEN
                    schedule  $vm$  on  $pe_j$ 
                     $pe_j - 1$ 
                END IF
            END IF
        END FOR
    END FOR
    wait for interval  $t$ 
END WHILE
```

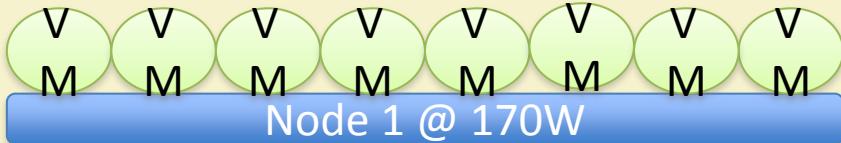


IIT KHARAGPUR

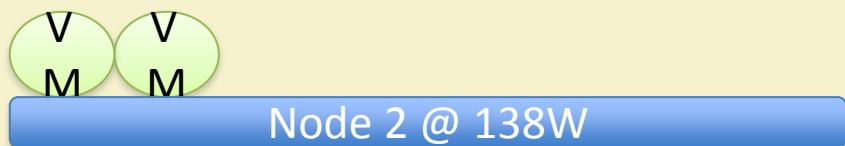


NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

485 Watts vs. 552 Watts !



VS.



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

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.

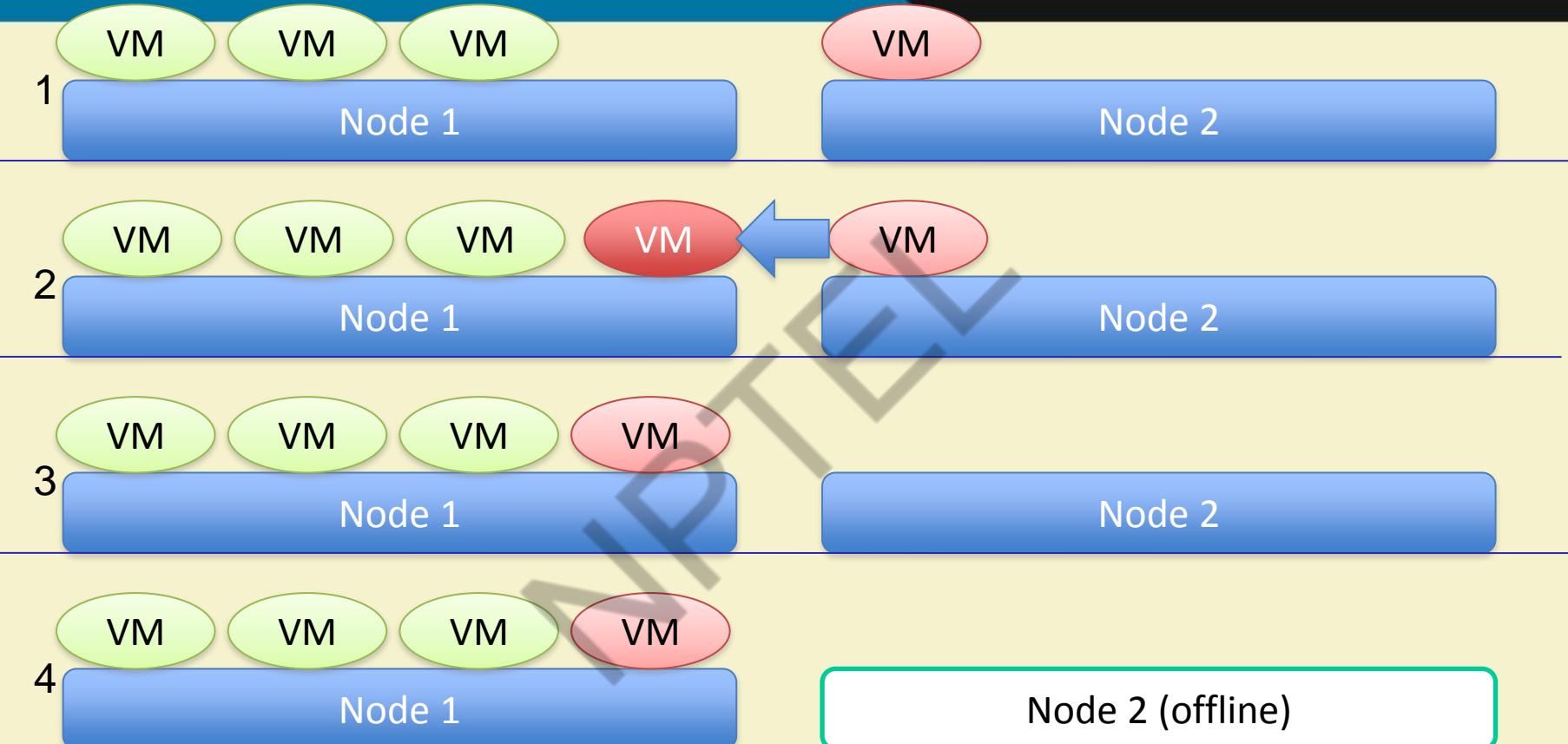
Management



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES



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



IIT KHARAGPUR

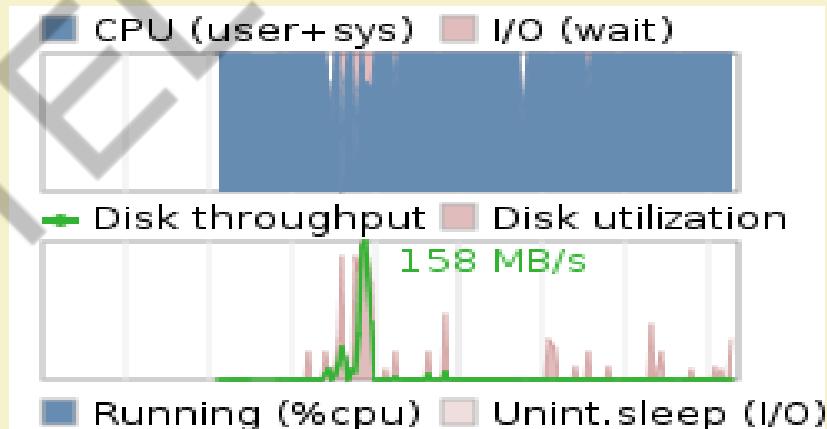


NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Typical Cloud Linux Image

- Start with Ubuntu 9.04.
- Remove all packages not required for base image.
 - No X11
 - No Window Manager
 - Minimalistic server install
 - Can load language support on demand (via package manager)
- Readahead profiling utility.
 - Reorder boot sequence
 - Pre-fetch boot files on disk
 - Minimize CPU idle time due to I/O delay
- Optimize Linux kernel.
 - Built for Xen DomU
 - No 3d graphics, no sound, minimalistic kernel
 - Build modules within kernel directly

Boot chart for ubuntu-minimal (Fri May 8 15:01:26 EDT 2009)
uname: Linux 2.6.28-11-generic #42-Ubuntu SMP Fri Apr 17 01:58:03 UTC 2009 x86_64
release: Ubuntu 9.04
CPU: Intel(R) Core(TM)2 Duo CPU T9300 @ 2.50GHz (1)
kernel options: root=UUID=042a98cc-dab1-4c5d-a45f-9088b7067ad9 ro quiet splash quiet
time: 0:08



VM Image
Design

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.

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.

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
NPTEL ONLINE
CERTIFICATION COURSES

Thank you!

NPTEL



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

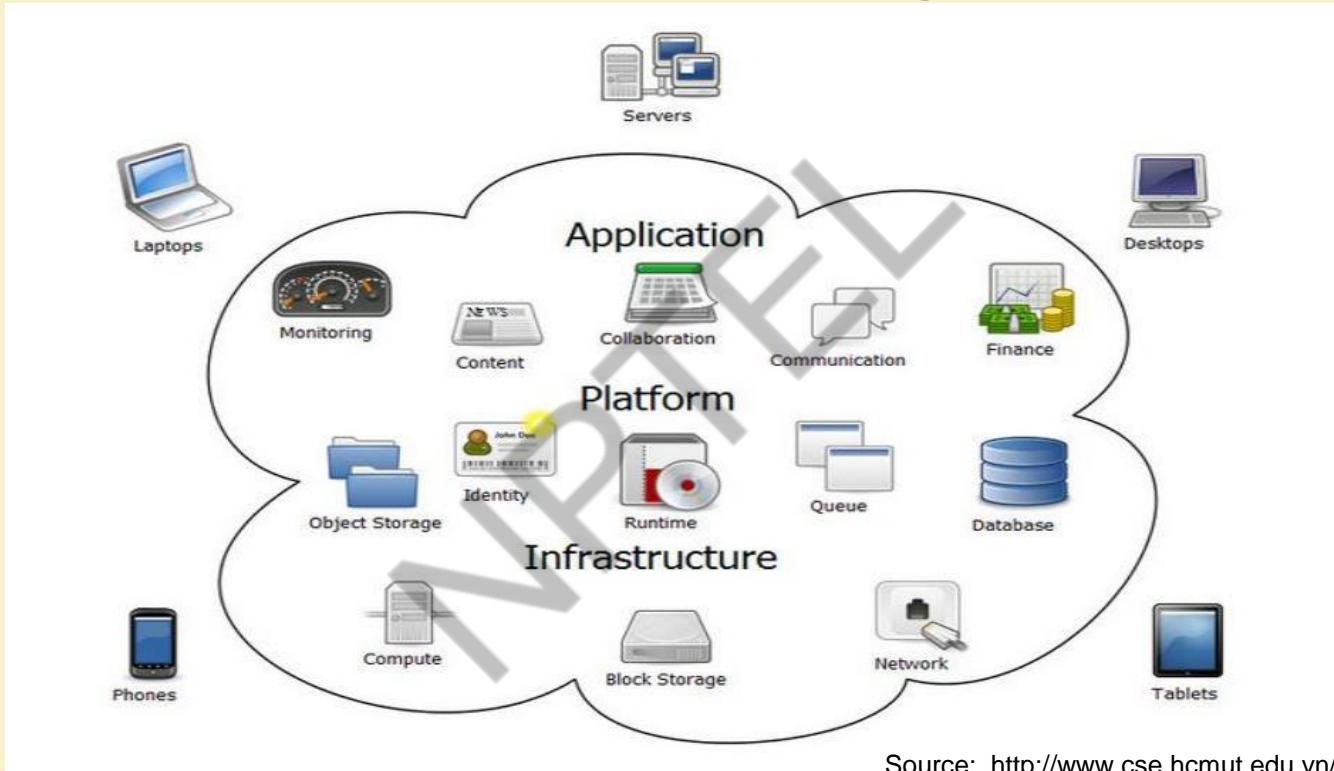
CLOUD COMPUTING

Resource Management - II

PROF. SOUMYA K. GHOSH

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
IIT KHARAGPUR

Different Resources in Computing



Source: <http://www.cse.hcmut.edu.vn/~ptvu/gc/2012/GC-pp.pdf>

Resources types

- **Physical resource**
 - Computer, disk, database, network, scientific instruments.
- **Logical resource**
 - Execution, monitoring, communicate application .

Source: <http://www.cse.hcmut.edu.vn/~ptvu/gc/2012/GC-pp.pdf>



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

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.hcmut.edu.vn/~ptvu/gc/2012/GC-pp.pdf>



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

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
NPTEL ONLINE
CERTIFICATION COURSES

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

Resource Management – Challenges (Hardware)

- CPU (central processing unit)
- Memory
- Storage
- Workstations
- Network elements
- Sensors/actuators

NPTEL



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

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
NPTEL ONLINE
CERTIFICATION COURSES

Resource Management Aspects

- Resource provisioning
- Resource allocation
- Resource requirement mapping
- Resource adaptation
- Resource discovery
- Resource brokering
- Resource estimation
- Resource modeling

NPTEL



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

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	<p>Resource modeling is based on detailed information of transmission network elements, resources and entities participating in the network.</p> <p>Attributes of resource management: states, transitions, inputs and outputs within a given environment.</p> <p>Resource modeling helps to predict the resource requirements in subsequent time intervals</p>
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

Resource Provisioning Approaches

Approach	Description
Nash 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 successive 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 idiosyncrasies, 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 utilization.
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.

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 satisfactions 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 demand 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

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
SOA 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 (DEVAs)	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.



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

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

Performance Metrics for Resource Management

- Reliability
- Ease of deployment
- QoS
- Delay
- Control overhead

NPTEL



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

Thank you!

NPTEL



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

CLOUD COMPUTING

CLOUD SECURITY I

PROF. SOUMYA K. GHOSH

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
IIT KHARAGPUR

Security - Basic Components

- Confidentiality
 - Keeping data and resources hidden
- Integrity
 - Data integrity (integrity)
 - Origin integrity (authentication)
- Availability
 - Enabling access to data and resources



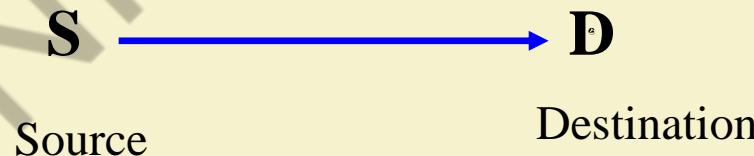
IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

Security Attacks

- Any action that compromises the security of information.
- Four types of attack:
 1. Interruption
 2. Interception
 3. Modification
 4. Fabrication
- Basic model:



IIT KHARAGPUR

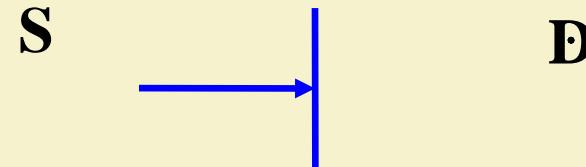


NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Security Attacks (contd.)

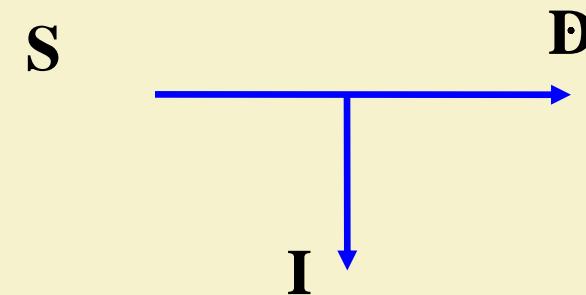
□ Interruption:

- Attack on availability



□ Interception:

- Attack on confidentiality



IIT KHARAGPUR

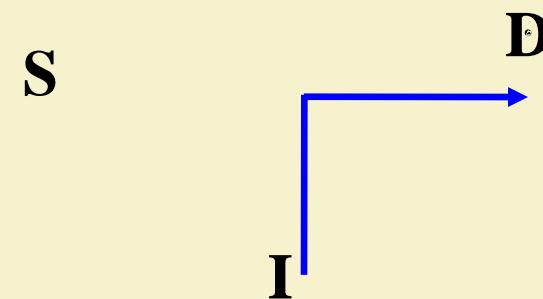
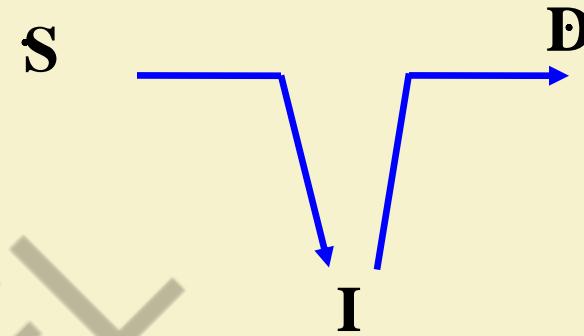


NPTEL
ONLINE
CERTIFICATION COURSES

Security Attacks

- Modification:
 - Attack on integrity

- Fabrication:
 - Attack on authenticity



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

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

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
NPTEL ONLINE
CERTIFICATION COURSES

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
NPTEL ONLINE
CERTIFICATION COURSES

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

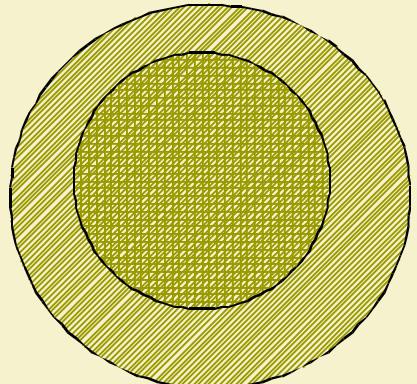


IIT KHARAGPUR

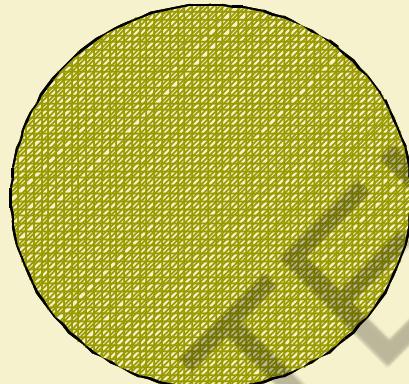


NPTEL ONLINE
CERTIFICATION COURSES

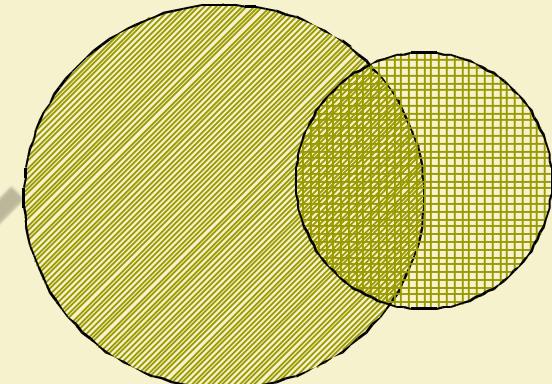
Types of Mechanisms



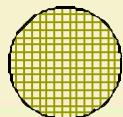
secure



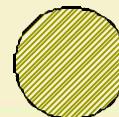
precise



broad



set of reachable states



set of secure states



IIT KHARAGPUR



NPTEL
ONLINE
CERTIFICATION COURSES

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

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
NPTEL ONLINE
CERTIFICATION COURSES

Human Issues

- Organizational Problems
 - Power and responsibility
 - Financial benefits
- People problems
 - Outsiders and insiders
 - Social engineering

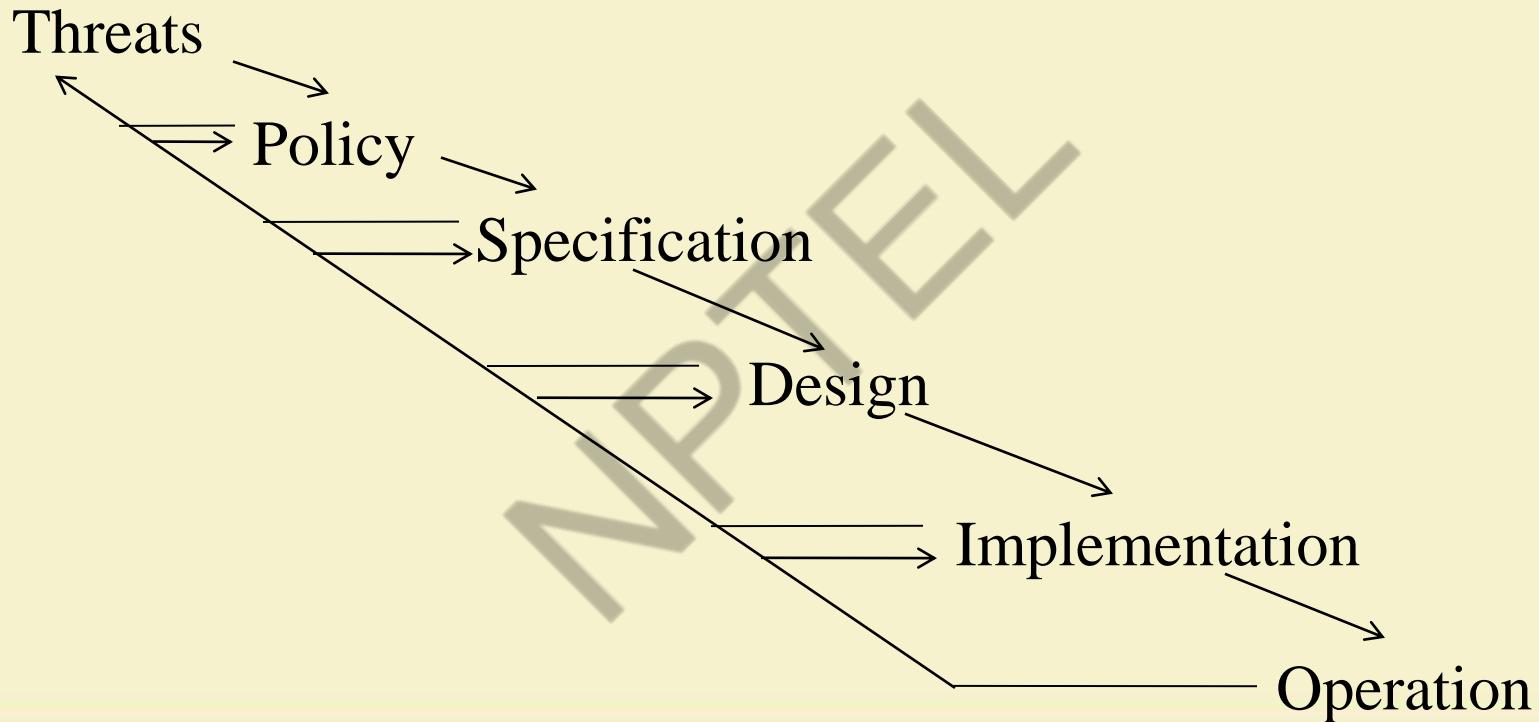


IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Tying Together



IIT KHARAGPUR



NPTEL
ONLINE
CERTIFICATION COURSES

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
NPTEL ONLINE
CERTIFICATION COURSES

❑ 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.



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

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
NPTEL ONLINE
CERTIFICATION COURSES

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

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

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
NPTEL ONLINE
CERTIFICATION COURSES

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 IDSes
 - *snort* is a free and upgradeable IDS for several platforms
 - Most IDSs send alerts to log files regularly
 - Serious events can trigger paging, E-Mail, telephone

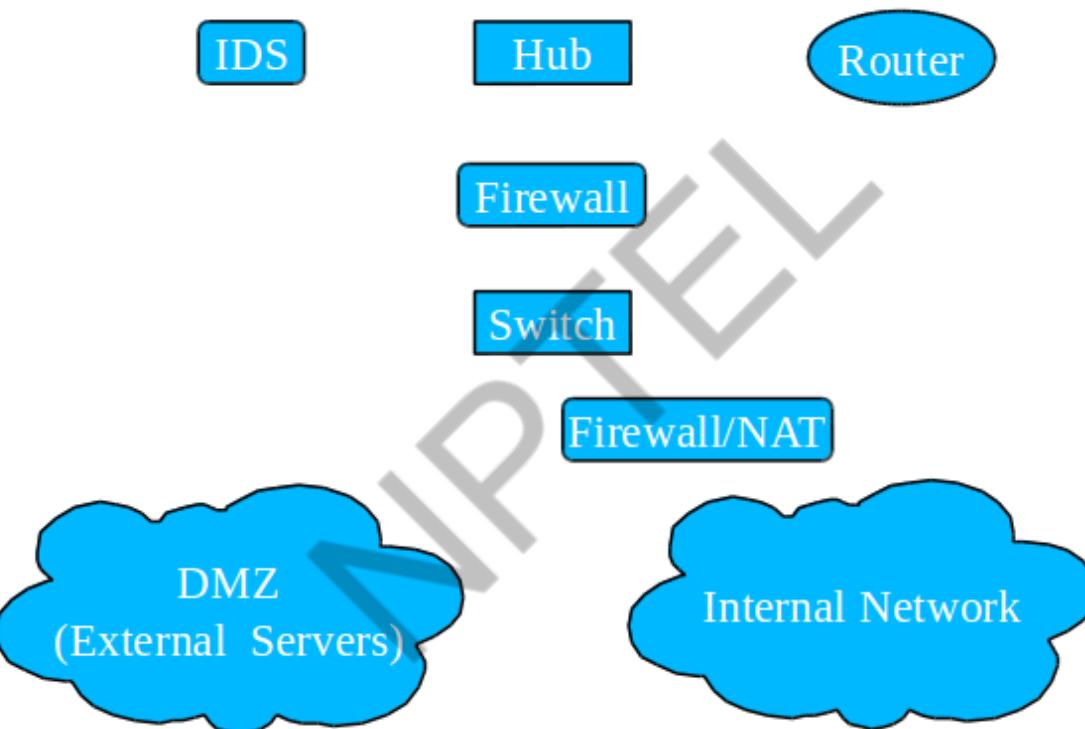


IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Step 2: Implement Security Policy



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

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
NPTEL ONLINE
CERTIFICATION COURSES

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
NPTEL ONLINE
CERTIFICATION COURSES

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
NPTEL ONLINE
CERTIFICATION COURSES

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
NPTEL ONLINE
CERTIFICATION COURSES

Thank You!

NPTEL



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

CLOUD COMPUTING

CLOUD SECURITY II

PROF. SOUMYA K. GHOSH

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
IIT KHARAGPUR

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
NPTEL ONLINE
CERTIFICATION COURSES

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
NPTEL ONLINE
CERTIFICATION COURSES

Why aren't Everyone using Cloud?

Clouds are **still** subject to traditional data confidentiality, integrity, availability, and privacy issues, plus some additional attacks



IIT KHARAGPUR

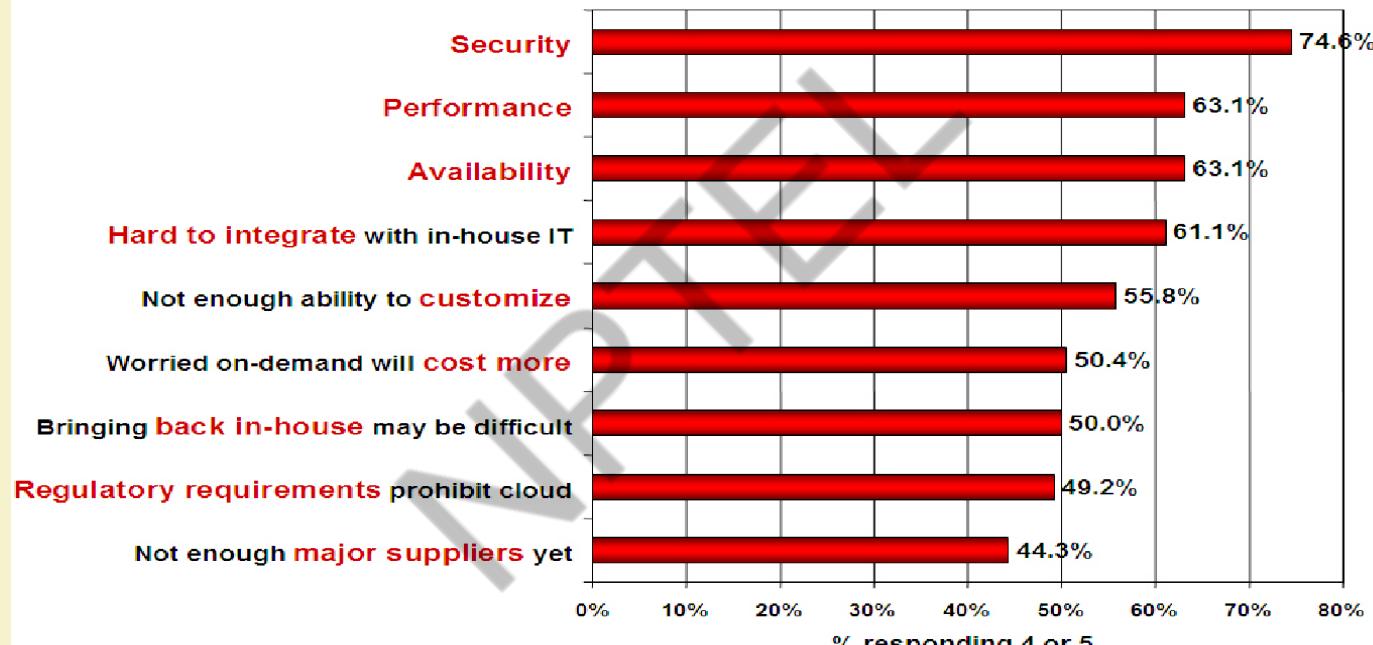


NPTEL ONLINE
CERTIFICATION COURSES

Concern...

Q: Rate the challenges/issues ascribed to the 'cloud'/on-demand model

(1=not significant, 5=very significant)



Source: IDC Enterprise Panel, August 2008 n=244

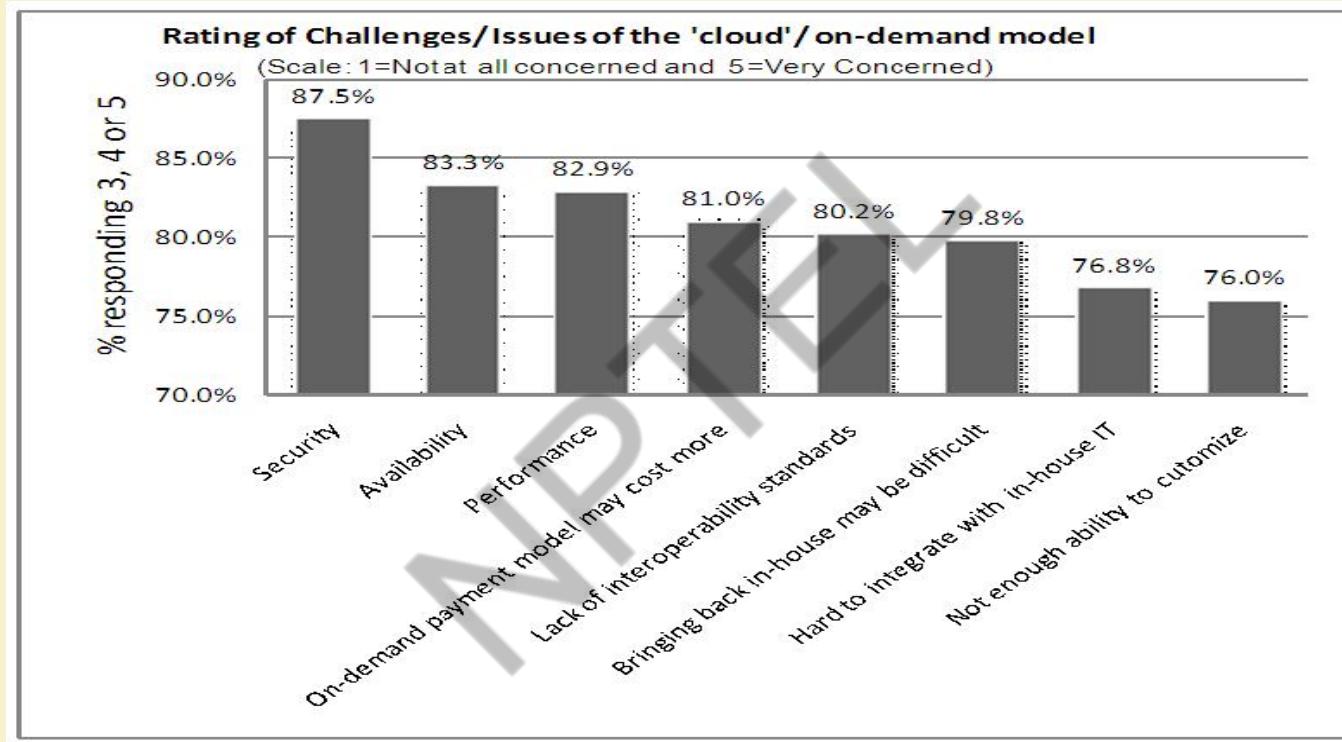


IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Survey on Potential Cloud Barriers



Source: IDC Ranking Security Challenges

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**



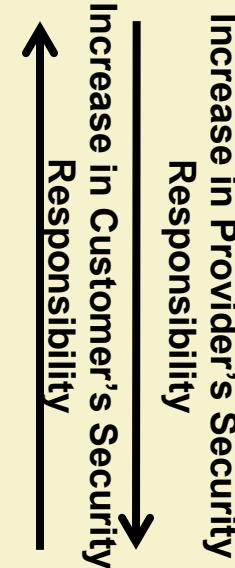
IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

Security Stack

- **IaaS:** entire infrastructure from facilities to hardware
- **PaaS:** application, middleware, database, messaging supported by IaaS
 - Customer-side system administrator manages the same with provider handling platform, infrastructure security
- **SaaS:** self contained operating environment: content, presentation, apps, management
 - Service levels, security, governance, compliance, liability, expectations of the customer & provider are contractually defined

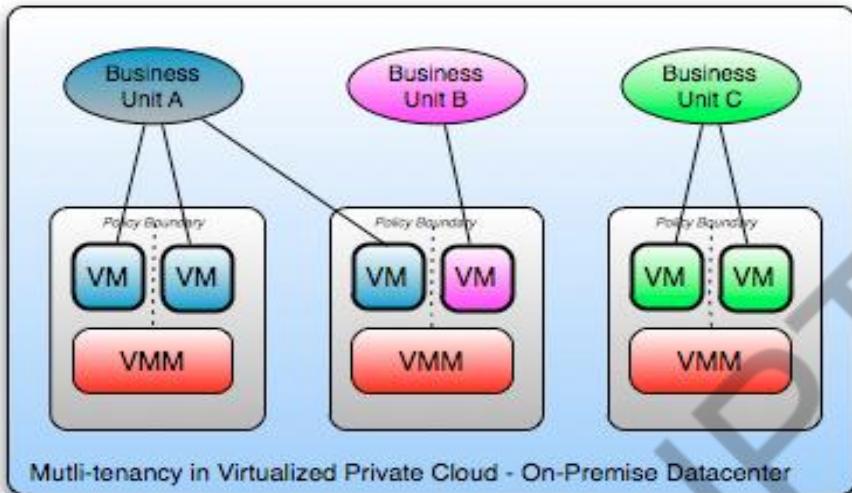


IIT KHARAGPUR

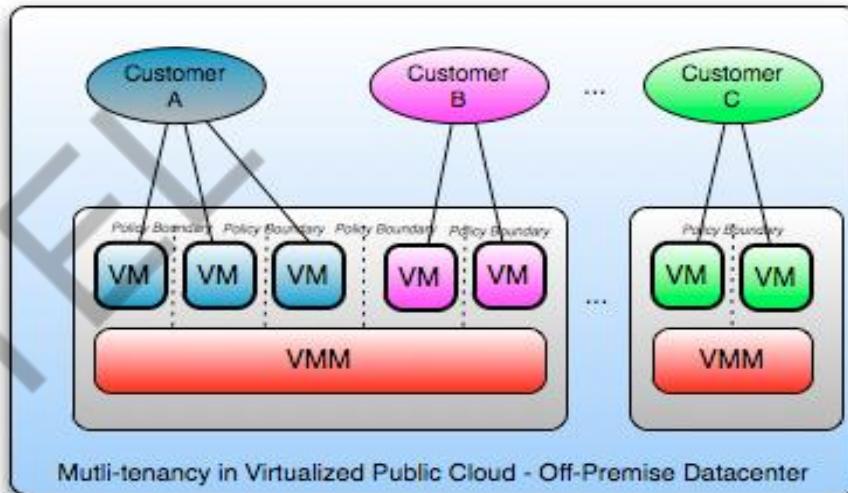


NPTEL ONLINE
CERTIFICATION COURSES

Sample Clouds



Private Cloud of Company XYZ with 3 business units, each with different security, SLA, governance and chargeback policies on shared infrastructure



Public Cloud Provider with 3 business customers, each with different security, SLA, governance and billing policies on shared infrastructure

Source: "Security Guidance for Critical Areas of Focus in Cloud Computing" v2.1, p.18



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

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
NPTEL ONLINE
CERTIFICATION COURSES

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

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
NPTEL ONLINE
CERTIFICATION COURSES

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 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



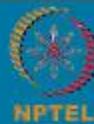
NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

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

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

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
NPTEL ONLINE
CERTIFICATION COURSES

Other Cloud Security Issues...

- Virtualization
- Access Control & Identity Management
- Application Security
- Data Life Cycle Management



IIT KHARAGPUR



NPTEL
ONLINE
CERTIFICATION COURSES

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
NPTEL ONLINE
CERTIFICATION COURSES

Virtualization (contd...)

- Hypervisor vulnerabilities:
 - Keystroke logging: Some VM technologies enable the logging of keystrokes and screen updates to be passed across virtual terminals in the virtual machine, writing to host files and permitting the monitoring of encrypted terminal connections inside the VM.
 - Virtual machine backdoors: covert communication channel
 - ARP Poisoning: redirect packets going to or from the other VM.
- Hypervisor Risks
 - Rogue hypervisor rootkits
 - Initiate a 'rogue' hypervisor
 - Hide itself from normal malware detection systems
 - Create a covert channel to dump unauthorized code



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

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
NPTEL ONLINE
CERTIFICATION COURSES

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

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

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 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
NPTEL ONLINE
CERTIFICATION COURSES

Thank You!

NPTEL



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

CLOUD COMPUTING

CLOUD SECURITY III

PROF. SOUMYA K. GHOSH

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
IIT KHARAGPUR

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

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

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
NPTEL ONLINE
CERTIFICATION COURSES

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

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

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

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?

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
NPTEL ONLINE
CERTIFICATION COURSES

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
NPTEL ONLINE
CERTIFICATION COURSES

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
NPTEL ONLINE
CERTIFICATION COURSES

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

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



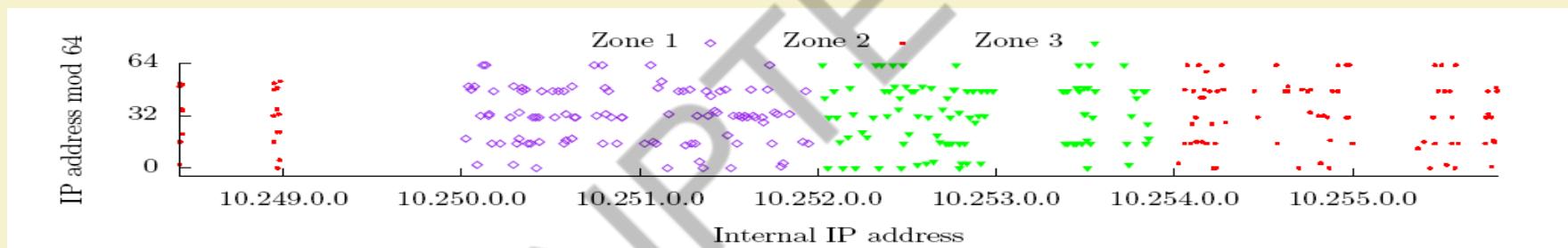
IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Instance Placement Parameters

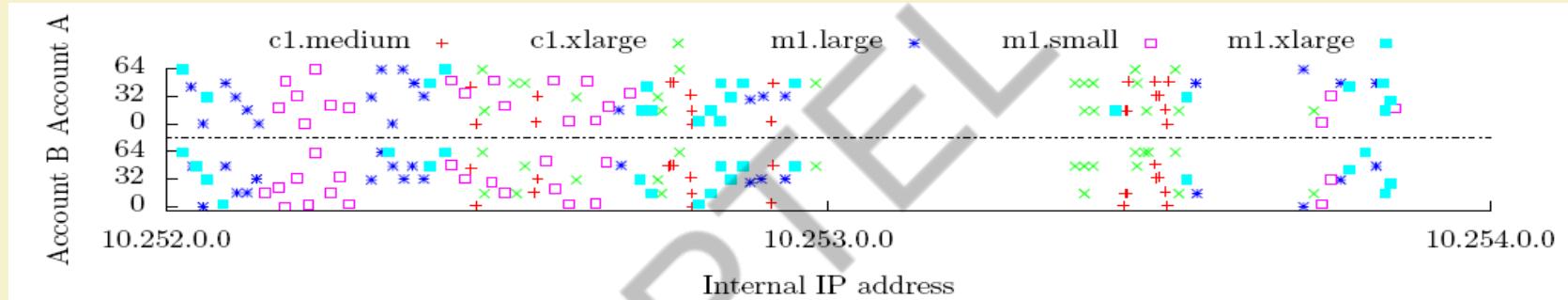
- EC2's internal address space is cleanly partitioned between availability zones
 - Three availability zone; five instance-type/zone
 - 20 instances launched for each of the 15 availability zone/instance type pairs from a particular account (Say, Account A)



- Samples from each zone are assigned IP addresses from disjoint portions of the observed internal address space
- **Assumption:** internal IP addresses are statically assigned to physical machines
 - To ease out IP routing
- Availability zones use separate physical infrastructure

Instance Placement Parameters (contd...)

- 100 instances have been launched in Zone 3 using two different accounts: A & B (39 hours after terminating the Account A instances)



- Of 100 Account A Zone 3 instances
 - 92 had unique /24 prefixes
 - Four /24 prefixes had two instances each
- Of 100 Account B Zone 3 instances
 - 88 had unique /24 prefixes
 - Six of the /24 prefixes had two instances each
- A single /24 had both an m1.large and m1.xlarge instance
- Of 100 Account B IP's, 55 were repeats of IP addresses assigned to instances for Account A

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

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
NPTEL ONLINE
CERTIFICATION COURSES

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)



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

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 naïve 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



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

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
NPTEL ONLINE
CERTIFICATION COURSES

Leveraging Placement Locality (contd...)

- Experiment
 - Single victim instance is launched
 - Attacker launches 20 instances within 5 minutes
 - Perform co-residence check
 - 40% of the time the attacker launching just 20 probes achieves co-residence against a specific target instance



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

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

Exploiting Co-residence (contd...)

- Measuring cache usage
 - Time-shared cache allows an attacker to measure when other instances are experiencing computational load
 - Load measurement: allocate a contiguous buffer B of b bytes, s is cache line size (in bytes)
 - *Prime*: read B at s -byte offsets in order to ensure that it is cached.
 - *Trigger*: busy-loop until CPU's cycle counter jumps by a large value
 - *Probe*: measure the time it takes to again read B at s -byte offset
 - Cache-based covert channel:
 - Sender idles to transmit a 0 and frantically accesses memory to transmit a 1
 - Receiver accesses a memory block and observes the access latencies
 - High latencies are indicative that "1" is transmitted



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

Exploiting Co-residence (contd...)

- Load-based co-residence check
 - Co-residence check can be done without network- base technique
 - Adversary can actively cause load variation due to a publicly-accessible service running on the target
 - Use a priori knowledge about load variation
 - Induce computational load (lots of HTTP requests) and observe the differences in load samples

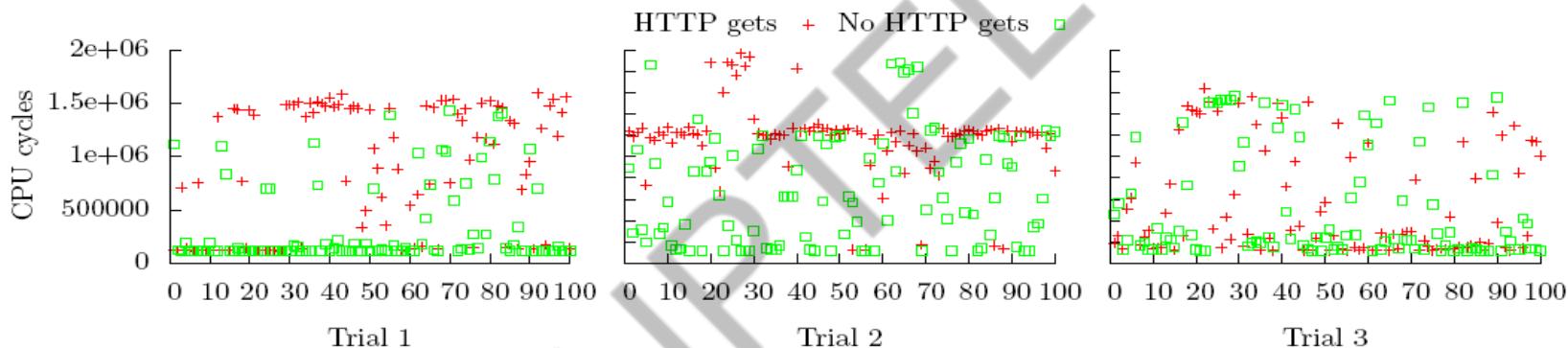


Figure 5: Results of executing 100 Prime+Trigger+Probe cache timing measurements for three pairs of m1.small instances, both when concurrently making HTTP get requests and when not. Instances in Trial 1 and Trial 2 were co-resident on distinct physical machines. Instances in Trial 3 were not co-resident.

- Instances in Trial 1 and Trial 2 were co-resident on distinct physical machines; instances in Trial 3 were not co-resident

Exploiting Co-residence (contd...)

- Estimating traffic rates
 - Load measurement might provide a method for estimating the number of visitors to a co-resident web server
 - It might not be a public information and could be damaging
 - Perform 1000 cache load measurements in which
 - no HTTP requests are sent
 - HTTP requests sent at a rate of (i) 50 per minute, (ii) 100 per minute, (iii) 200 per minutes

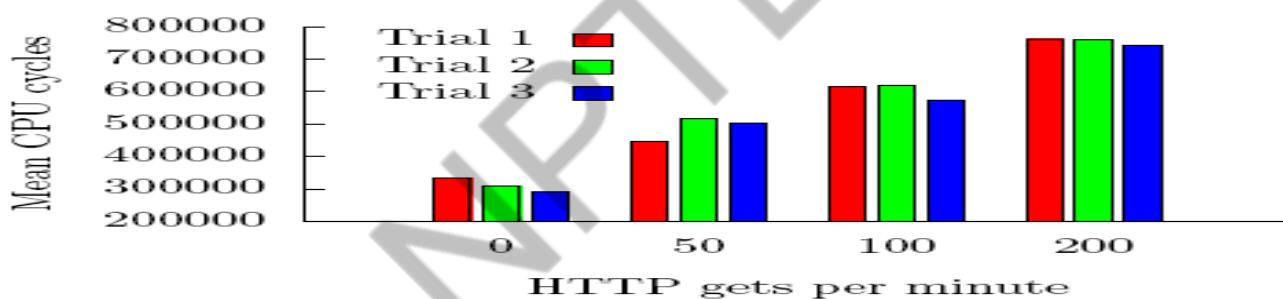


Figure 6: Mean cache load measurement timings (over 1 000 samples) taken while differing rates of web requests were made to a 3 megabyte text file hosted by a co-resident web server.



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Exploiting Co-residence (contd...)

- Keystroke timing attack
 - The goal is to measure the time between keystrokes made by a victim typing a password (or other sensitive information)
 - Malicious VM can observe keystroke timing in real time via cache-based load measurements
 - Inter-keystroke times if properly measured can be used to perform recovery of the password
 - In an otherwise idle machine, a spike in load corresponds to a letter being typed into the co-resident VM's terminal
 - Attacker does not directly learn exactly which keys are pressed, the attained timing resolution suffices to conduct the password-recovery attacks on SSH sessions



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

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



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

Summary

- New risks from cloud computing
- Shared physical infrastructure may and most likely will cause problems
 - Exploiting software vulnerabilities not addressed here
- Practical attack performed
- Some countermeasures proposed



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Thank You!

NPTEL



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

CLOUD COMPUTING

CLUSTER SECURITY IV

Security Issues in Collaborative SaaS Cloud

PROF. SOUMYA K. GHOSH

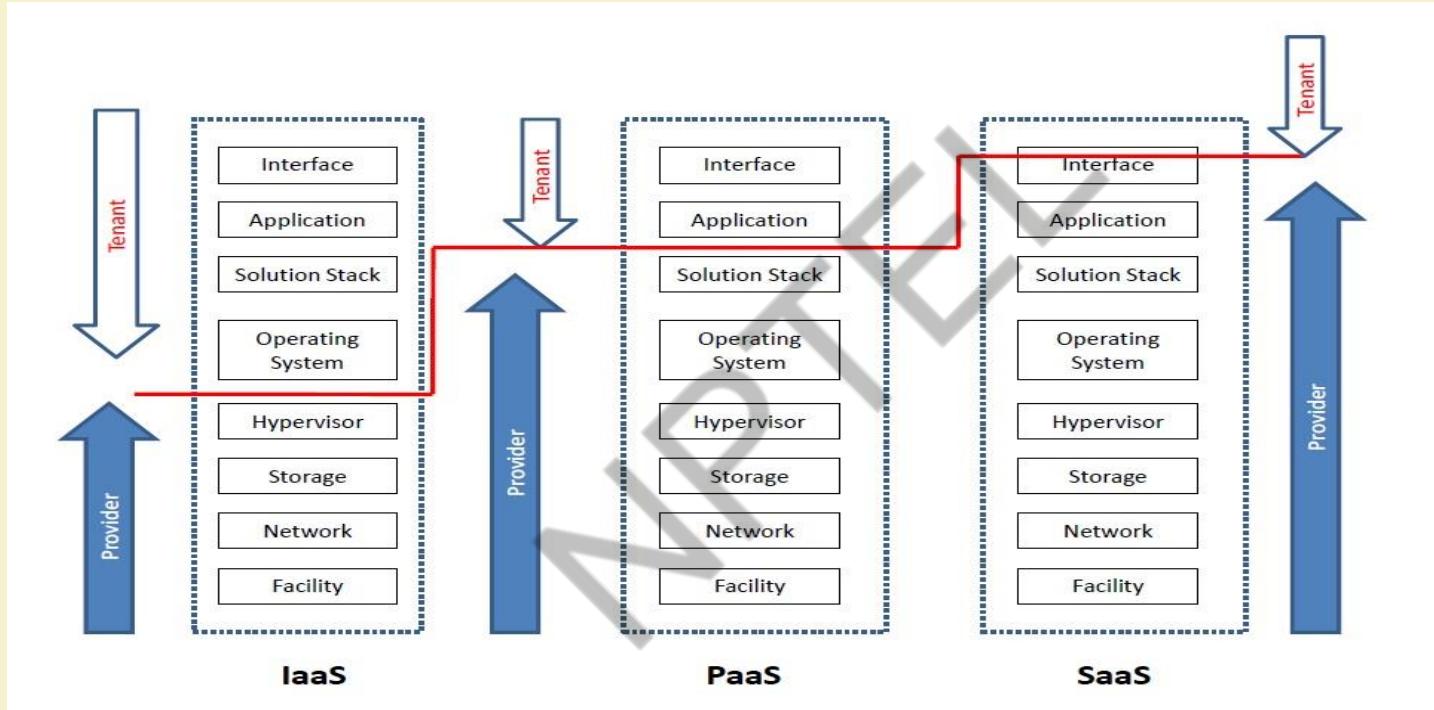
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

IIT KHARAGPUR

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

Security Responsibilities



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

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

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
NPTEL ONLINE
CERTIFICATION COURSES

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

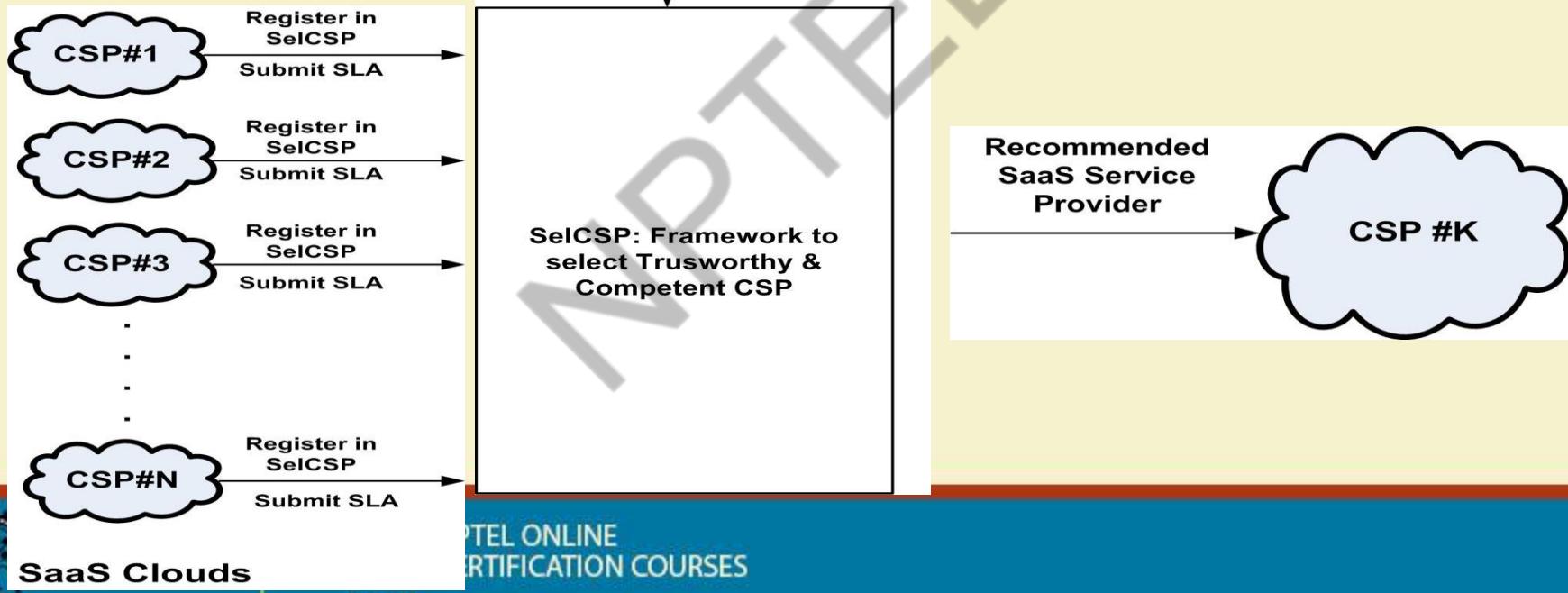


NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Objective - I

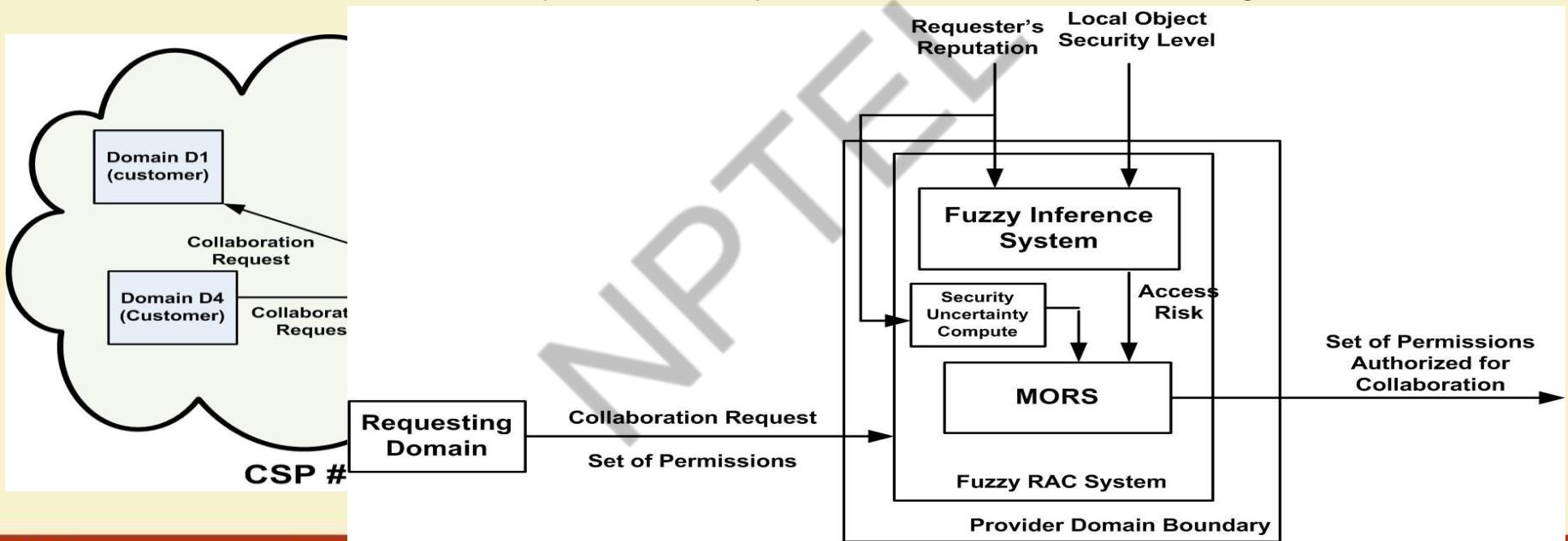
A framework (SelCSP) for selecting a trustworthy and competent SaaS collaboration service provider.

trustworthy and competent



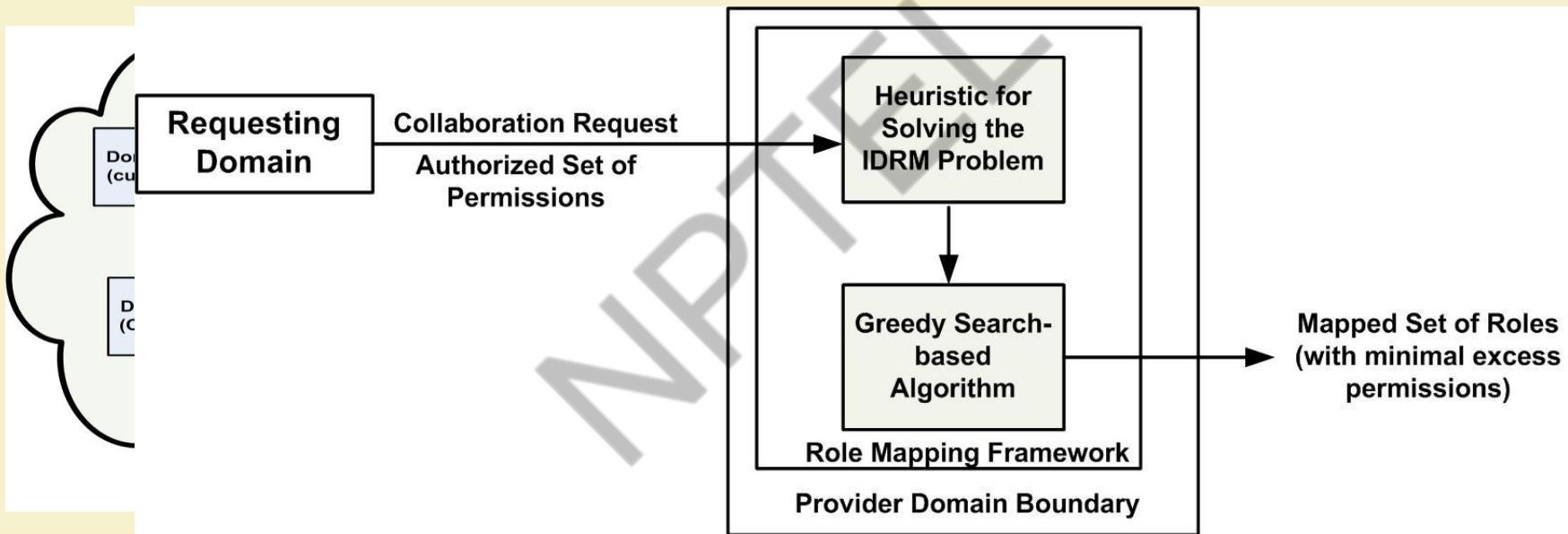
Objective - II

Select requests (for accessing local resources) from anonymous users, such that both access risk and security uncertainty due to information sharing are kept low.

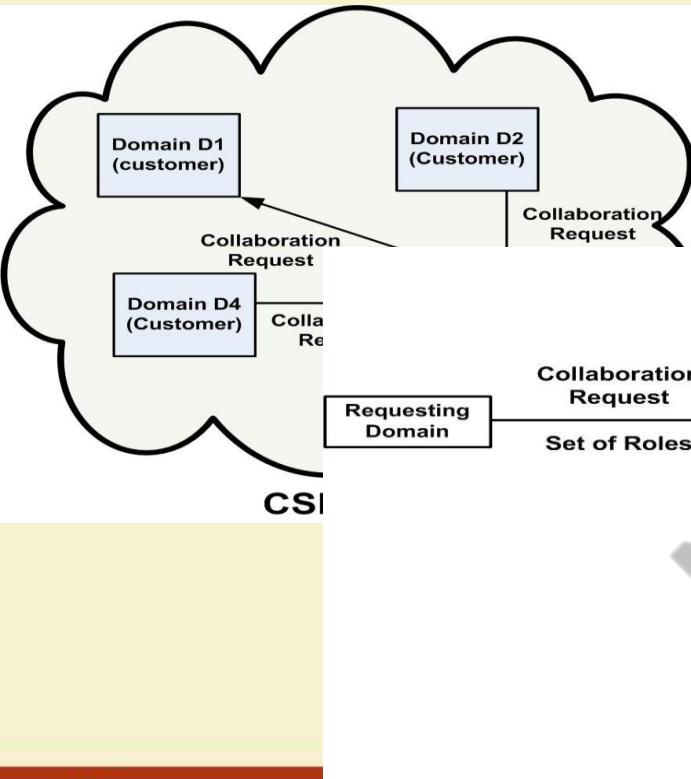


Objective - III

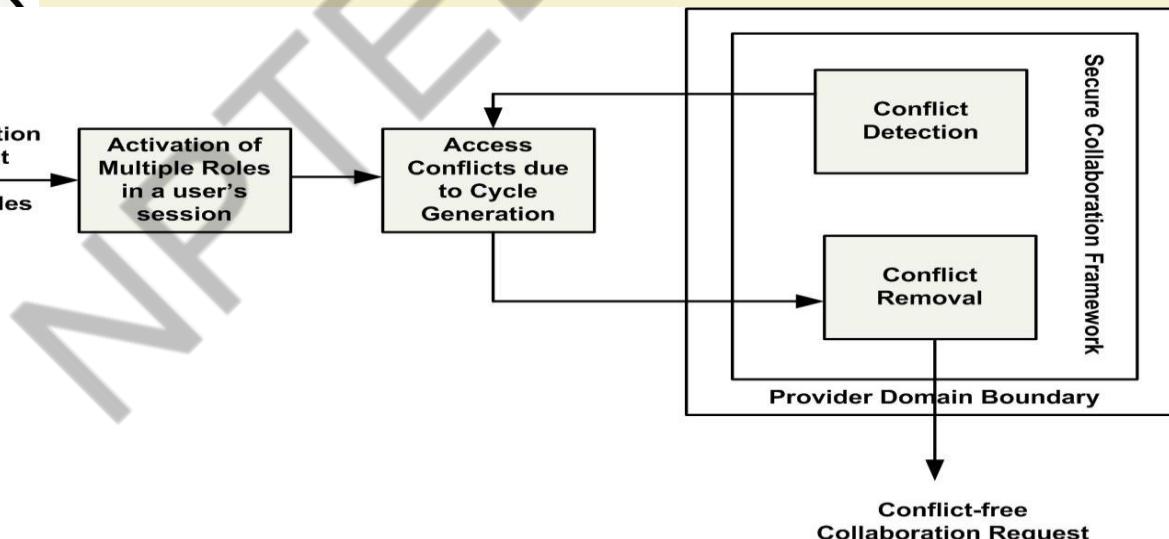
Formulate a heuristic for solving the IDRM problem, such that minimal excess privilege is granted



Objective - IV



A distributed secure collaboration framework, which uses only local information to dynamically detect and remove access conflicts.



Selection of Trustworthy and Competent SaaS Cloud Provider for Collaboration



IIT KHARAGPUR

9/20/2017



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

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

9/20/2017

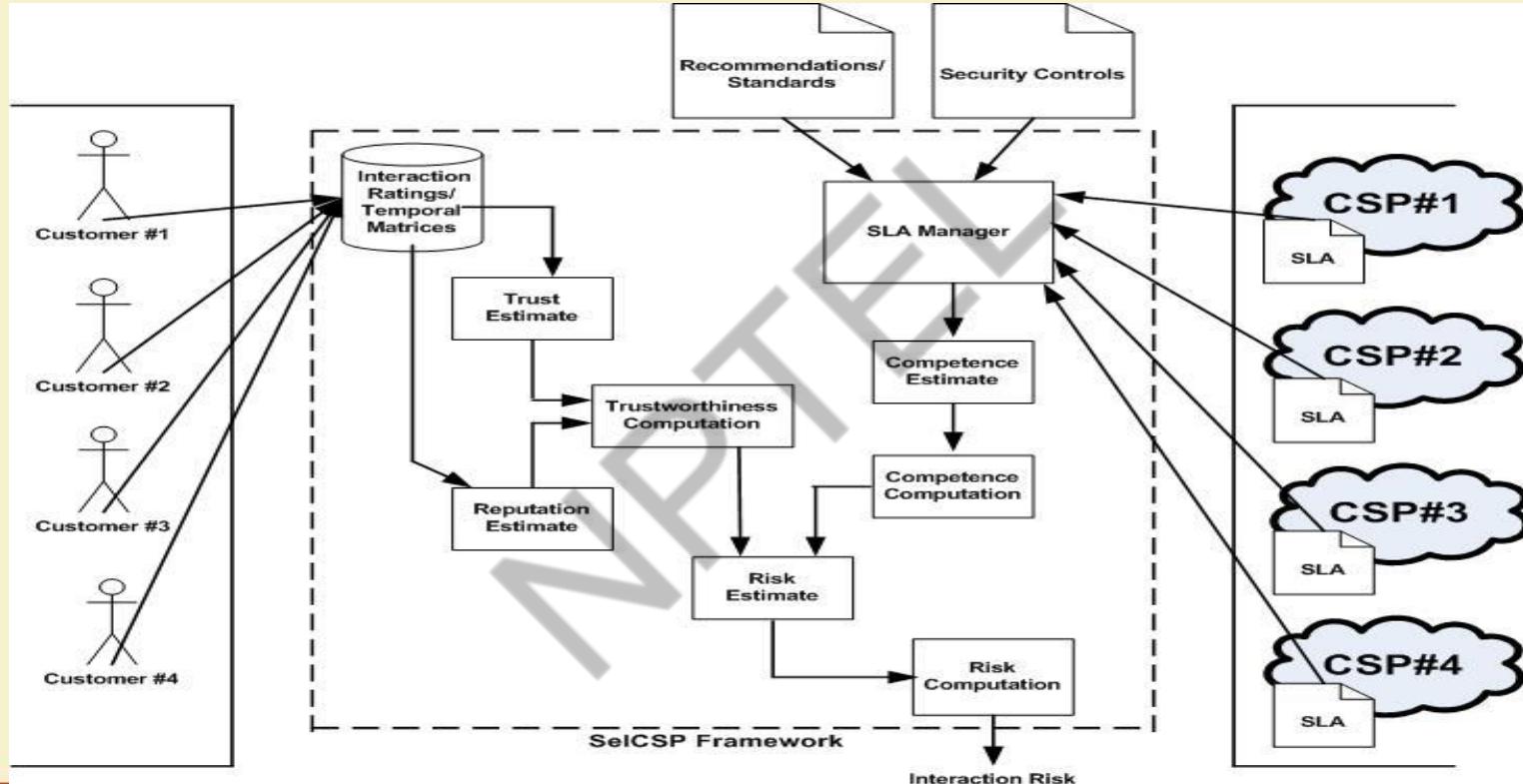


NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

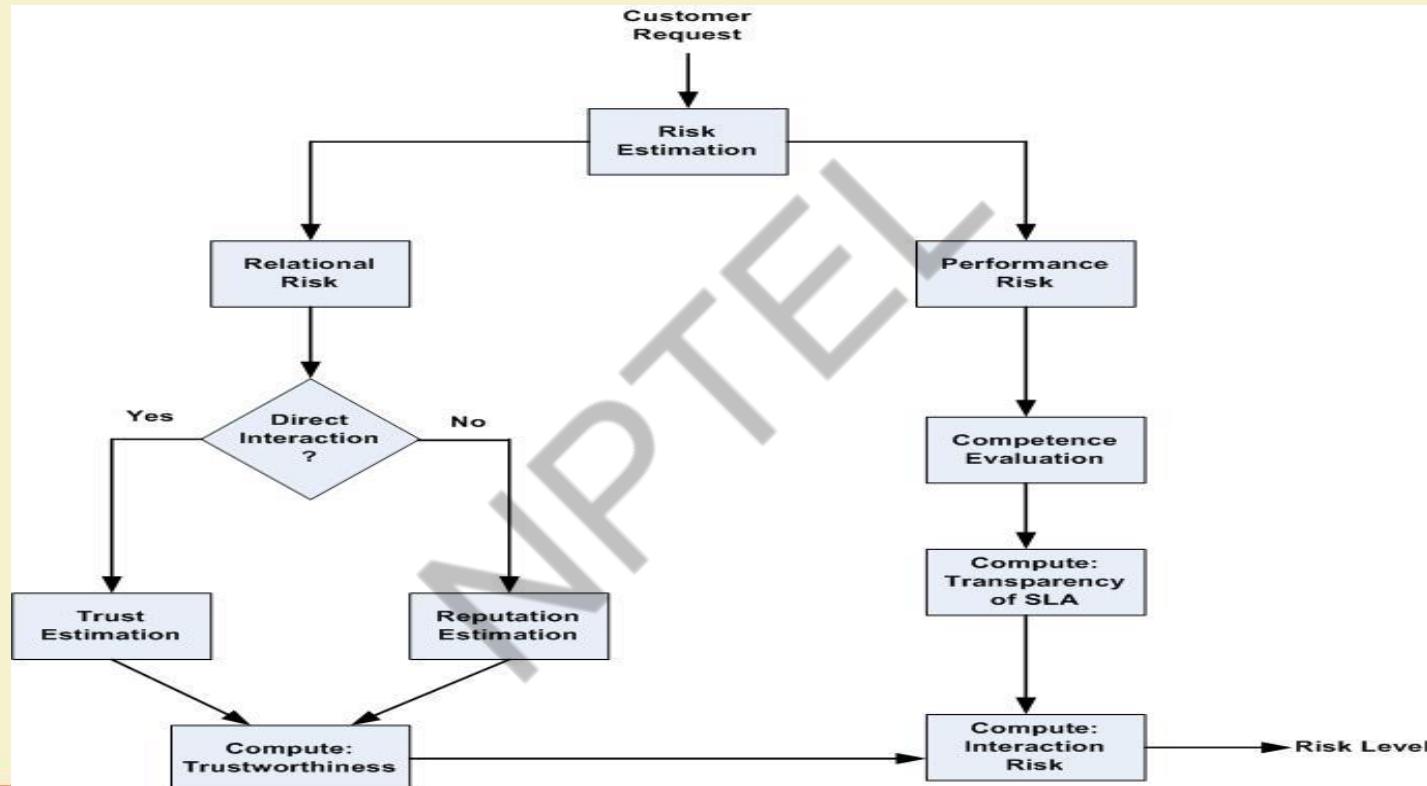
Service Level Agreement (SLA) for Clouds

- 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**

SelCSP Framework



SelCSP Framework - Overview



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

Recommending Access Requests from Anonymous Users for Authorization



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

Risk-based Access Control (RAC)

- RAC: Gives access to subjects even though they lack proper permissions
 - Goal: balance between *access risk* and *security uncertainty due to information sharing*
 - Flexible compared to binary MLS
- Challenges
 - Computing security uncertainty: not addressed
 - Authorization in existing RAC system: based on risk threshold and operational need.
 - Operational need: not quantified.
 - Discards many requests which potentially maximizes information sharing



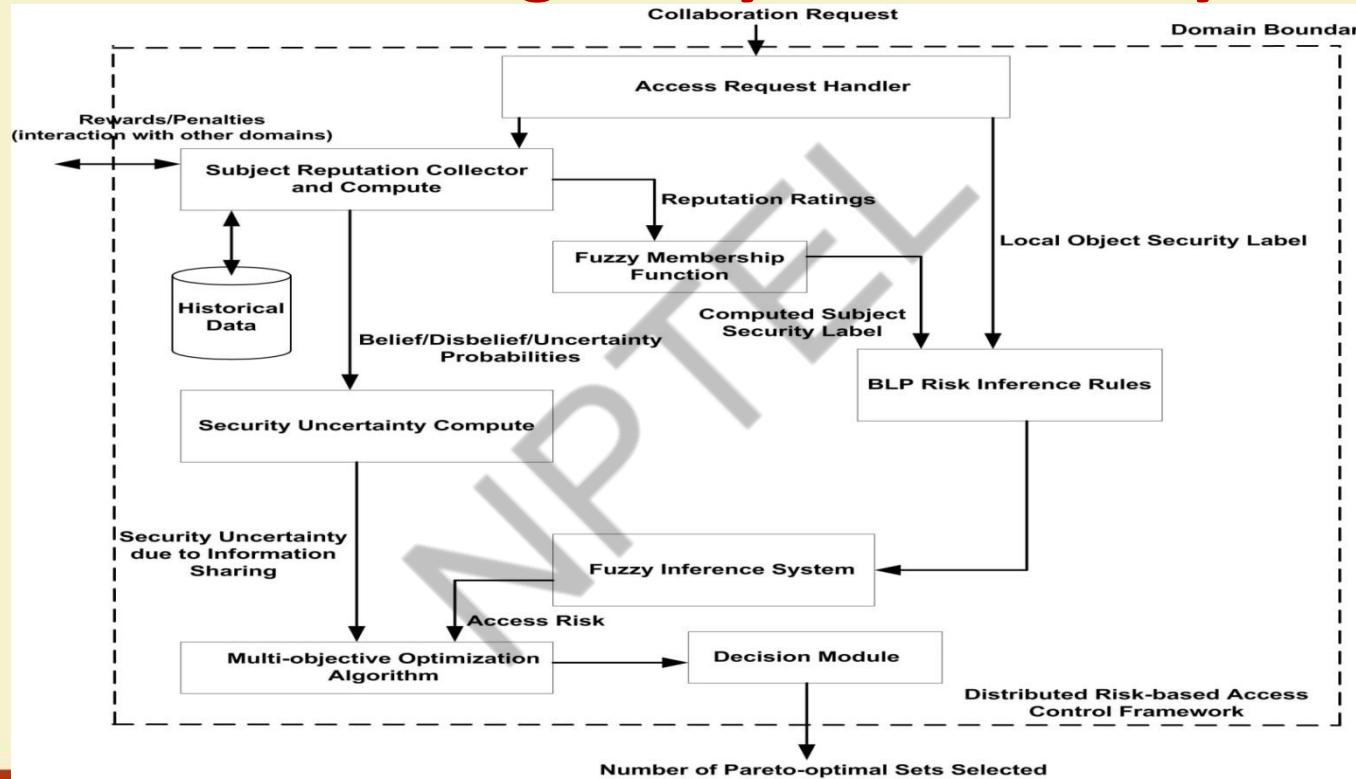
IIT KHARAGPUR

9/20/2017



NPTEL
ONLINE
CERTIFICATION COURSES

Distributed RAC using Fuzzy Inference System



Mapping of Authorized Permissions into Local Roles



JIT KHARAGPUR

9/20/2017



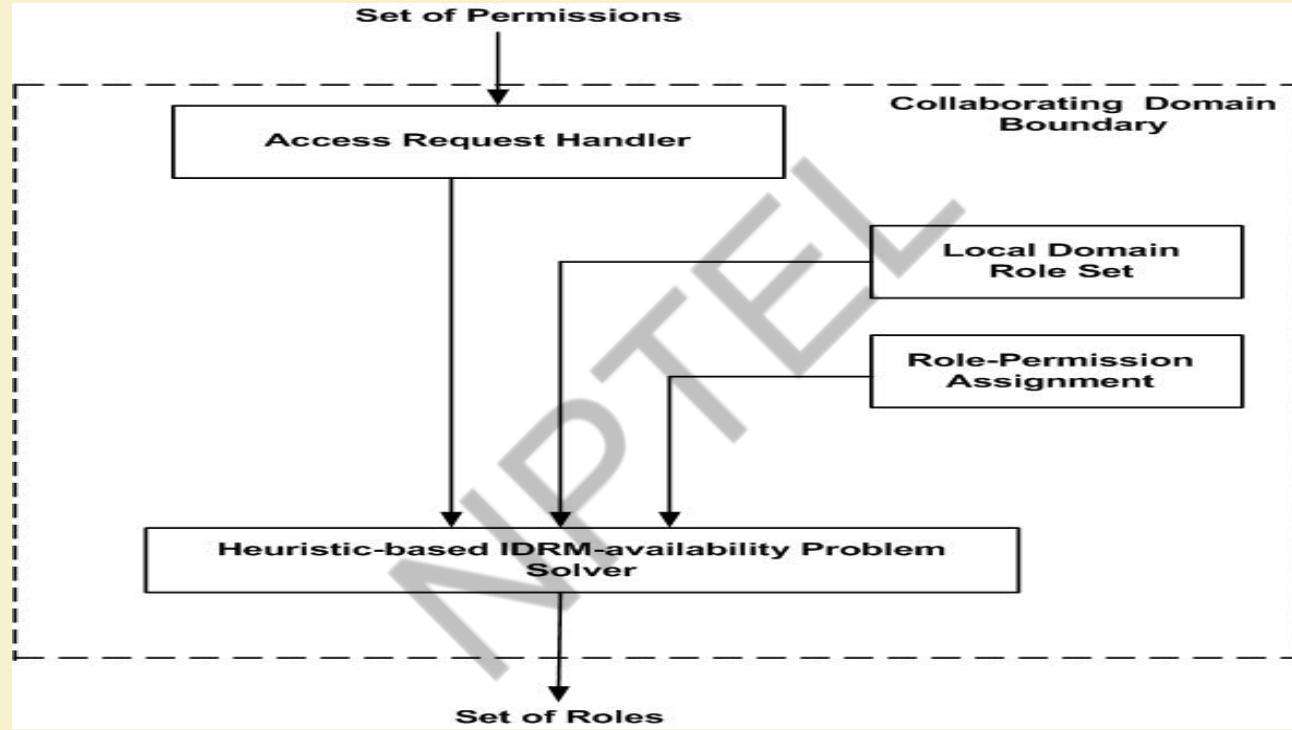
NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Inter-Domain Role Mapping (IDRM)

- Finds a minimal set of role which encompasses the requested permission set.
 - No polynomial time solution
 - Greedy search-based heuristics: suboptimal solutions
- Challenges:
 - There may exist multiple minimal role sets
 - There may not exist any role set which exactly maps all permissions
- Two variants of IDRM proposed: *IDRM-safety*, *IDRM-availability*
- Objective: formulate a novel heuristic to generate better solution for the IDRM-availability problem.
- Minimize the number of additional permissions



Distributed Role Mapping Framework



Dynamic Detection and Removal of Access Policy Conflicts

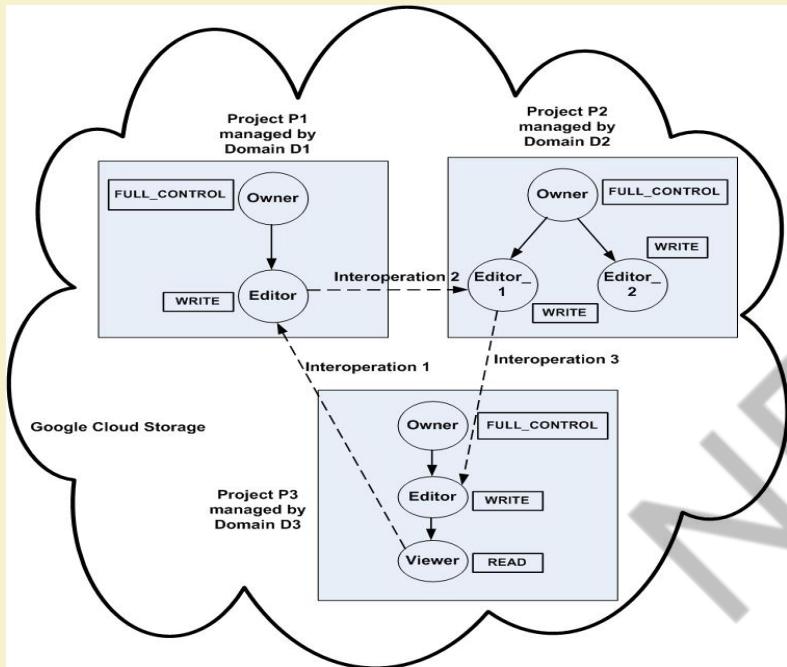


IIT KHARAGPUR

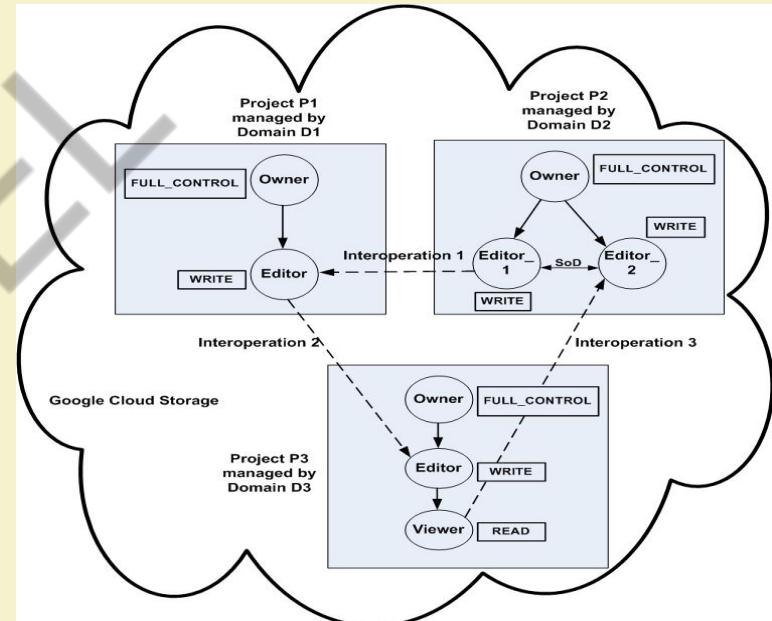


NPTEL
ONLINE
CERTIFICATION COURSES

Access Conflicts



Cyclic Inheritance Conflict



Violation of SoD Constraint



JIT KHARAGPUR

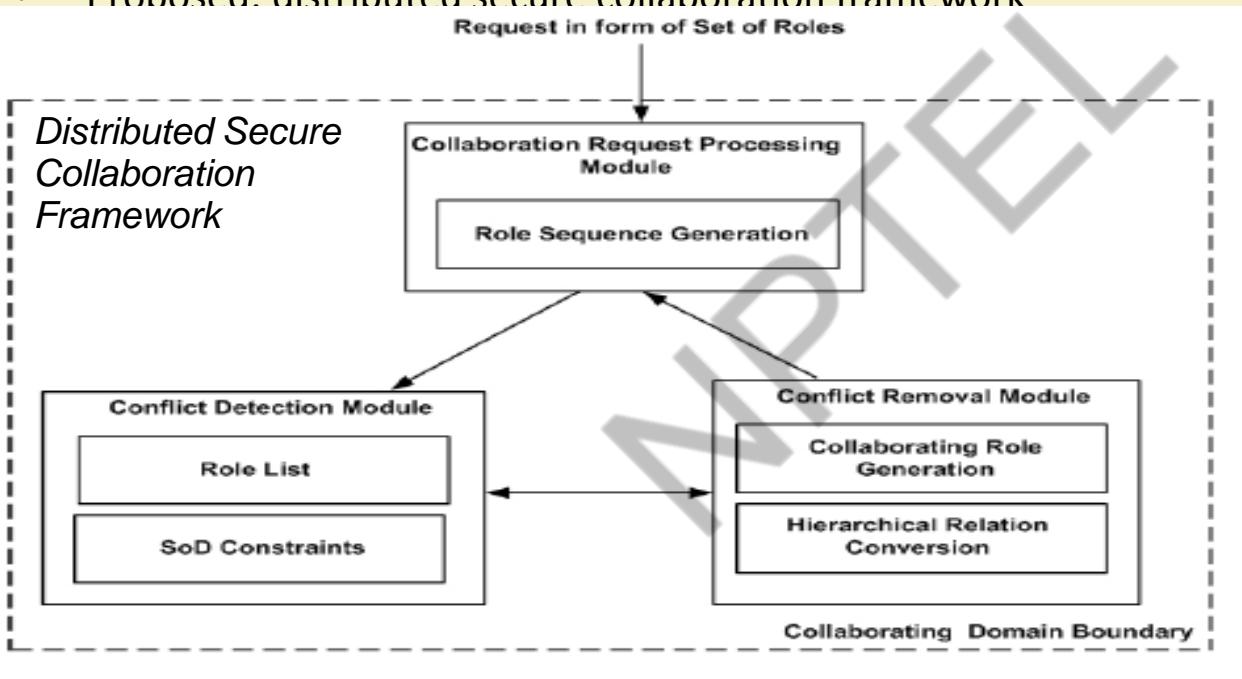
9/20/2017



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Objective

- Dynamic detection of conflicts to address **security** issue
- Removal of conflicts to address **availability** issue
- Proposed: distributed secure collaboration framework



- Role Sequence Generation
 - Interoperation request: pair of *entry* (from requesting domain), *exit* (from providing domain) roles
 - Role sequence: ordered succession of entry and exit roles
 - Role cycle:
 - Safe role cycle
 - Unsafe role cycle



Conflict Detection

- Detection of inheritance conflict
 - Necessary condition: at least one exit role
 - Sufficient condition: current entry role is senior to at least one exit role
- Detection of SoD constraint violation
 - Necessary condition: at least one exit role
 - Sufficient condition: current entry role and at least one exit role forms *conflicting pair*



IIT KHARAGPUR

9/20/2017



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Conflict Detection Algorithm

Conflict Removal

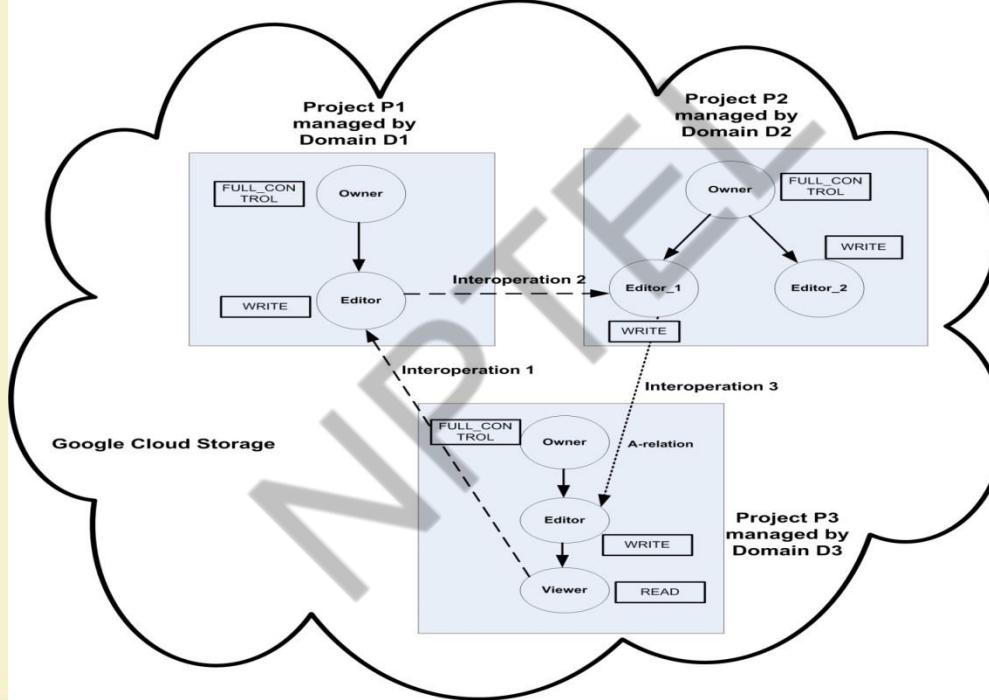
Cyclic Inheritance

- 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



Conflict Removal

Cyclic Inheritance: Inheritance Conflict Removal Rule for Exactly Matched Role



IIT KHARAGPUR

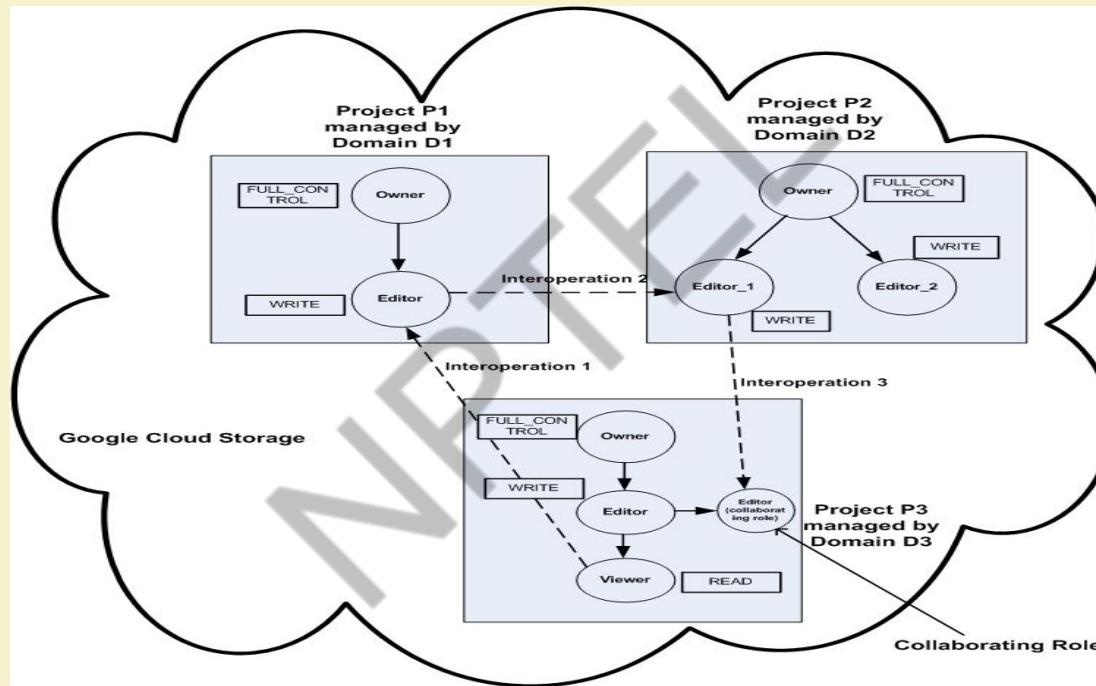
9/20/2017



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Conflict Removal

Cyclic Inheritance: Inheritance Conflict Removal Rule for No-Exactly Matched Role



IIT KHARAGPUR

9/20/2017



NPTEL
ONLINE
CERTIFICATION COURSES

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

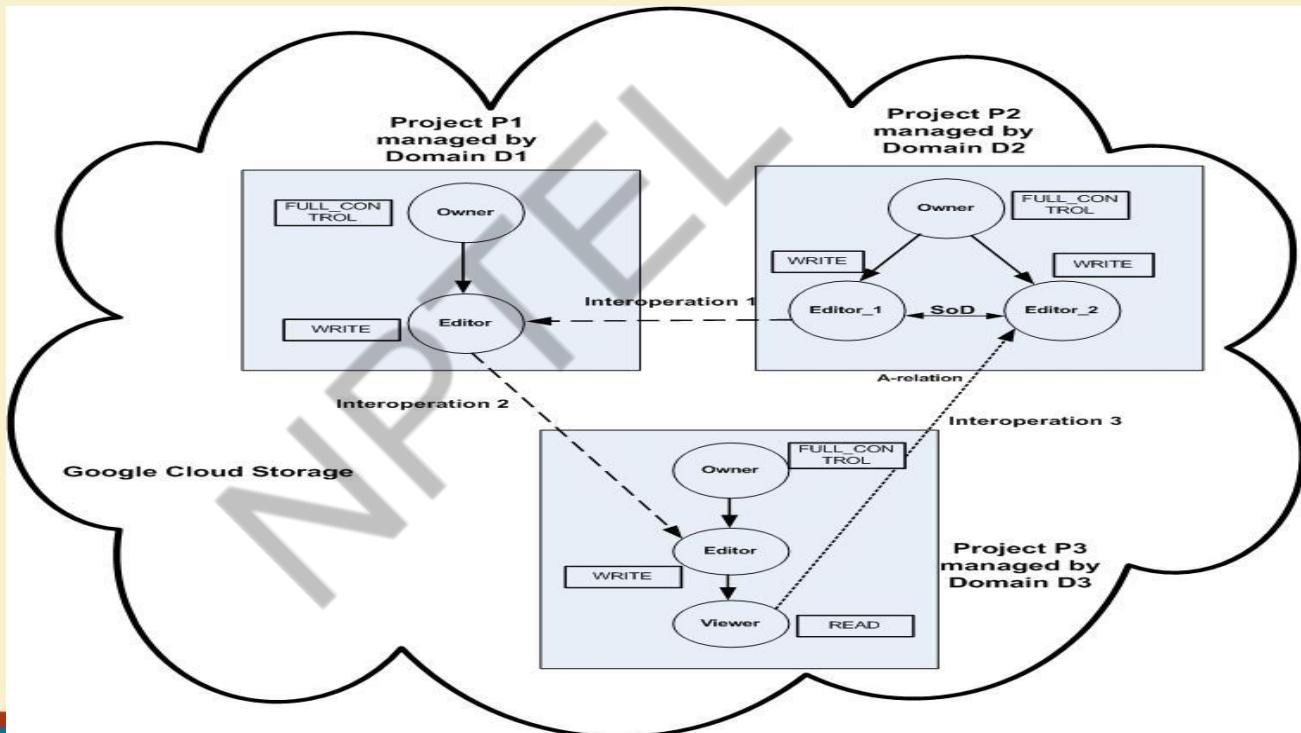
9/20/2017



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Conflict Removal

SoD Constraint Violation: SoD Conflict Removal Rule for Exactly Matched Role



IIT KHARAGPUR

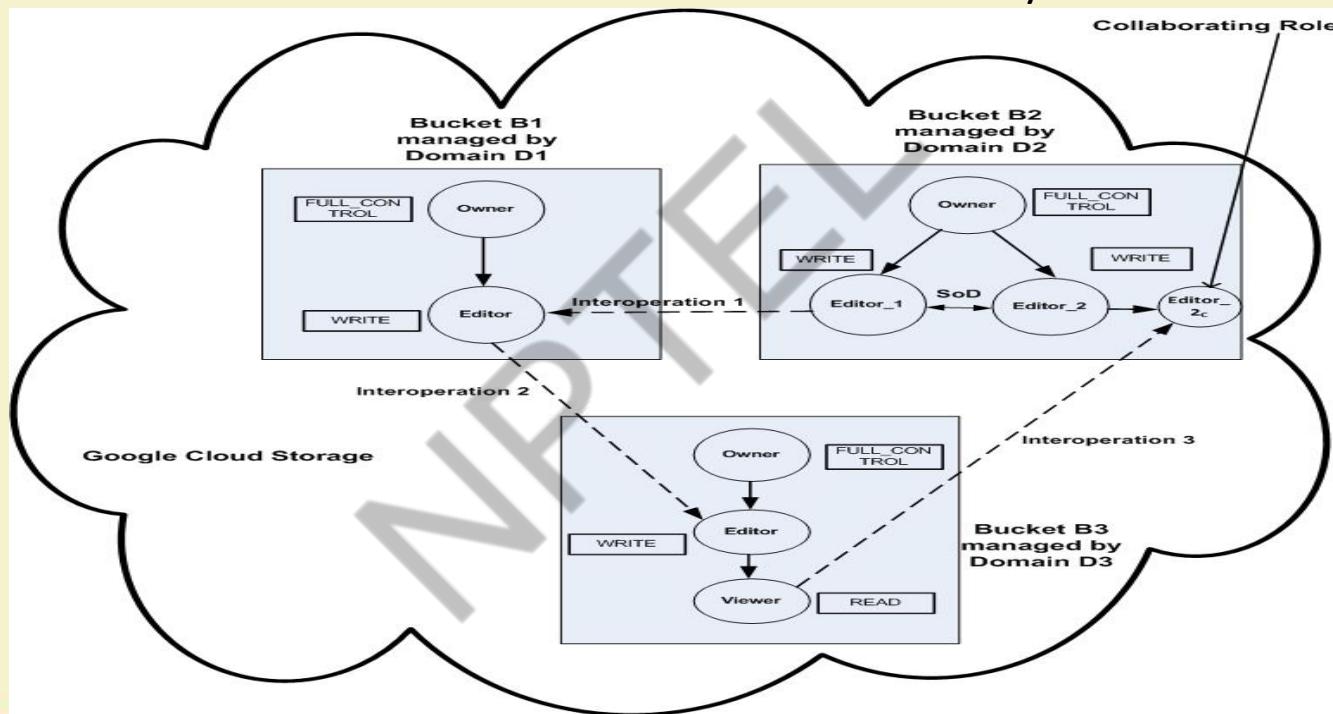
9/20/2017



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Conflict Removal

SoD Constraint Violation: SoD Conflict Removal Rule for No-Exactly Matched Role



IIT KHARAGPUR

9/20/2017



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Summary

Secure Collaboration SaaS Clouds: A Typical Approach

- Selection of Trustworthy and Competent SaaS Cloud Provider for Collaboration
- Recommending Access Requests from Anonymous Users for Authorization
- Mapping of Authorized Permissions into Local Roles
- Dynamic Detection and Removal of Access Policy Conflicts

Thank You!

NPTEL



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

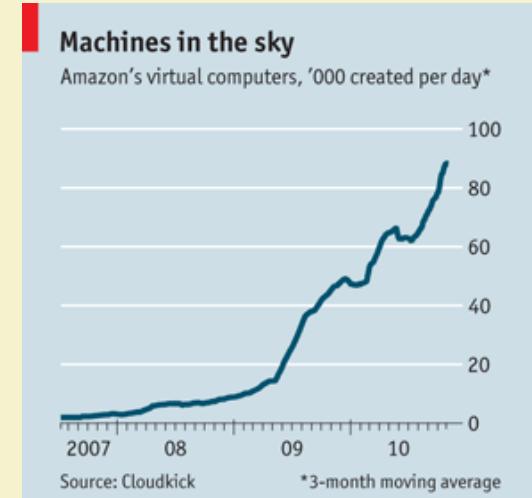
Cloud Computing : *Broker for Cloud Marketplace*

Prof. Soumya K Ghosh

Department of Computer Science and Engineering
IIT KHARAGPUR

INTRODUCTION

- Rapid growth of available cloud services
- Huge number of providers with varying QoS
- Different types of customer use cases – each with different requirements



IIT KHARAGPUR

9/20/2017



NPTEL
ONLINE
CERTIFICATION COURSES

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



MOTIVATION

- Flexible selection of cloud provider
- Trustworthiness of provider
- Monitoring of services
- Avoiding vendor lock-in



IIT KHARAGPUR

9/20/2017



NPTEL ONLINE
CERTIFICATION COURSES

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

9/20/2017



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

9/20/2017



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

CUSTOMER QoS PARAMETERS

Infrastructure-as-a-Service



Software-as-a-Service



- *More QoS parameter can be added easily.*



IIT KHARAGPUR

9/20/2017



NPTEL
ONLINE
CERTIFICATION COURSES

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



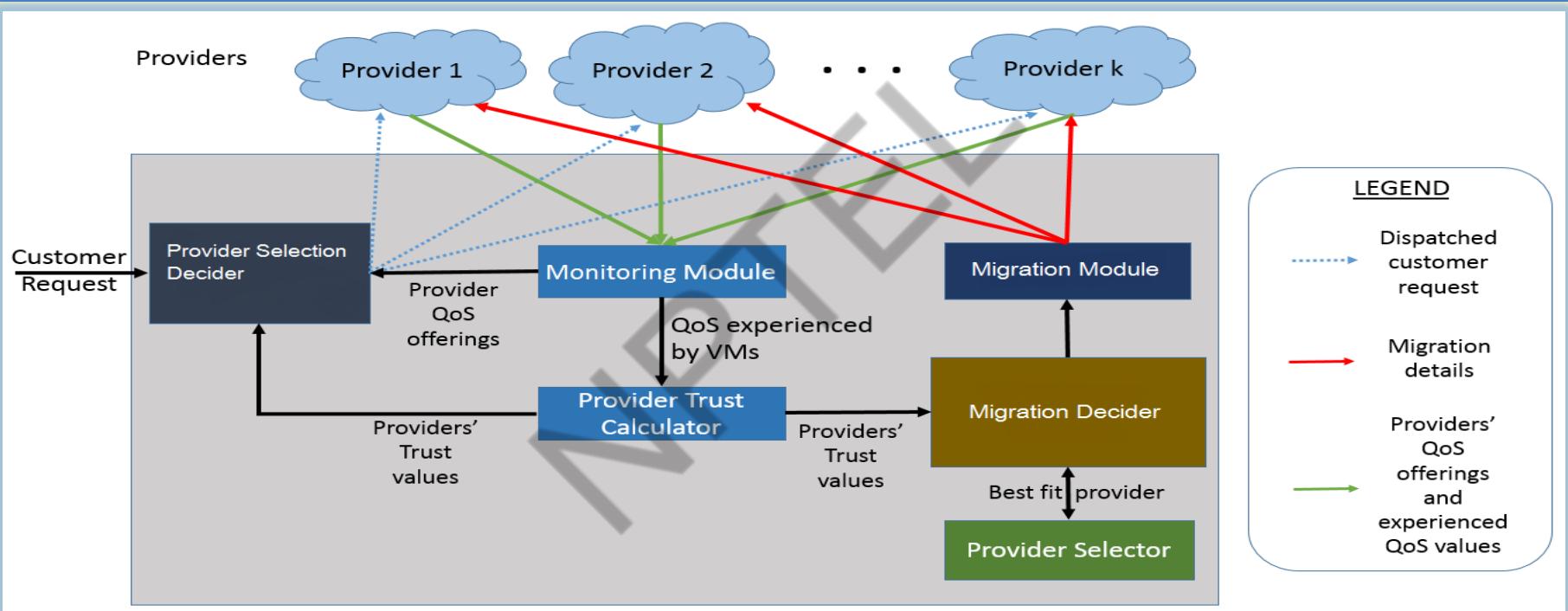
IIT KHARAGPUR

9/20/2017



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Typical MARKETPLACE Architecture



PROVIDER SELECTION

- 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



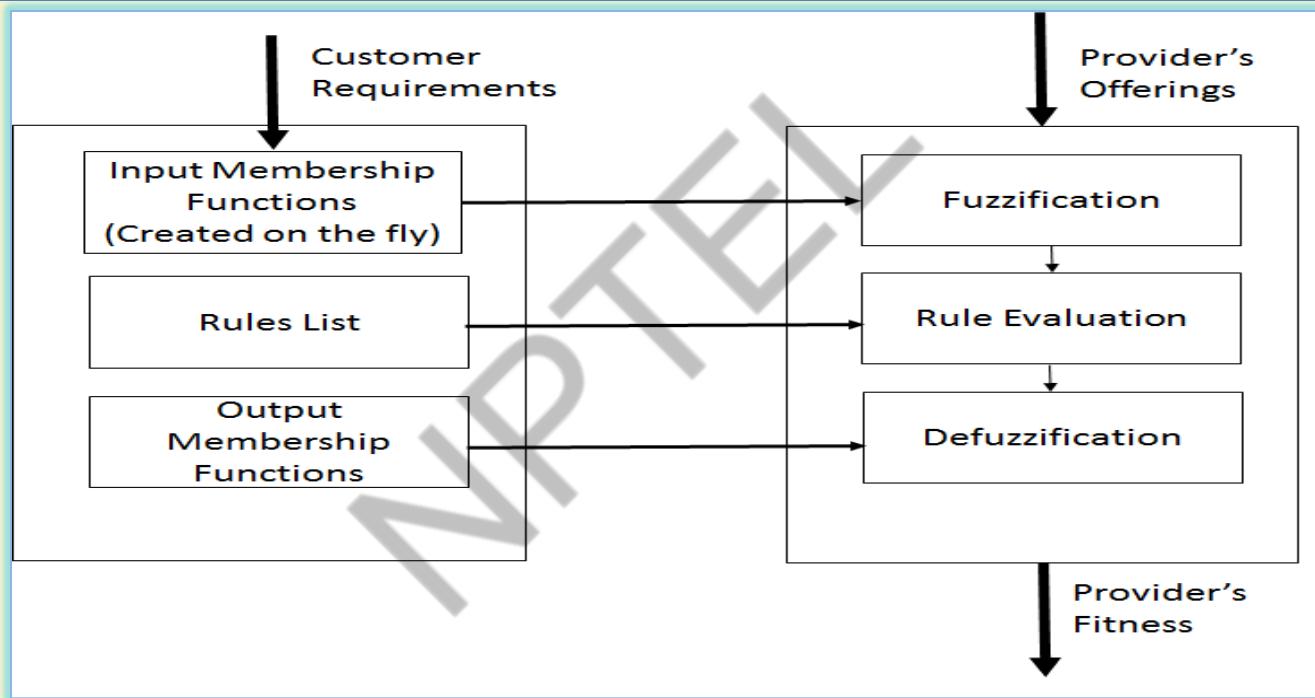
IIT KHARAGPUR

9/20/2017

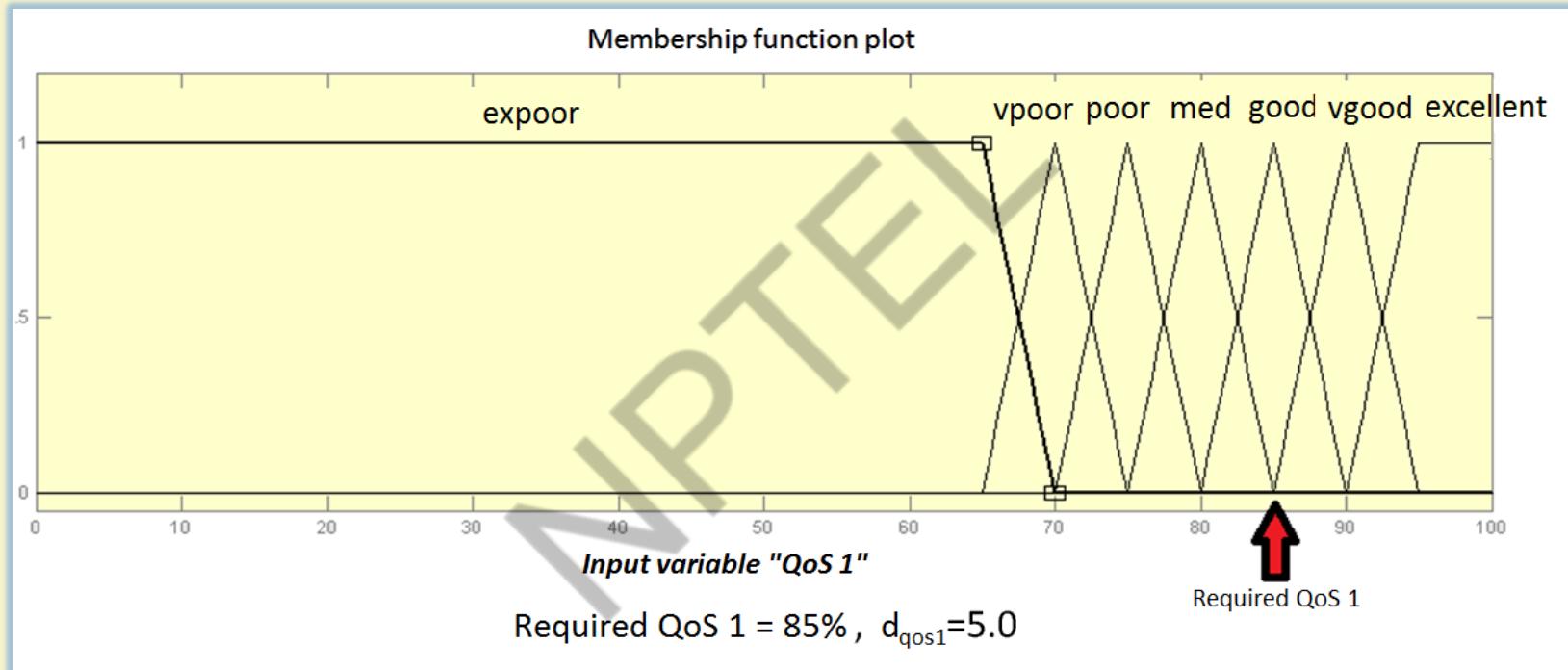


NPTEL ONLINE
CERTIFICATION COURSES

PROVIDER SELECTION



PROVIDER SELECTION – INPUT MEMBERSHIP FUNCTION



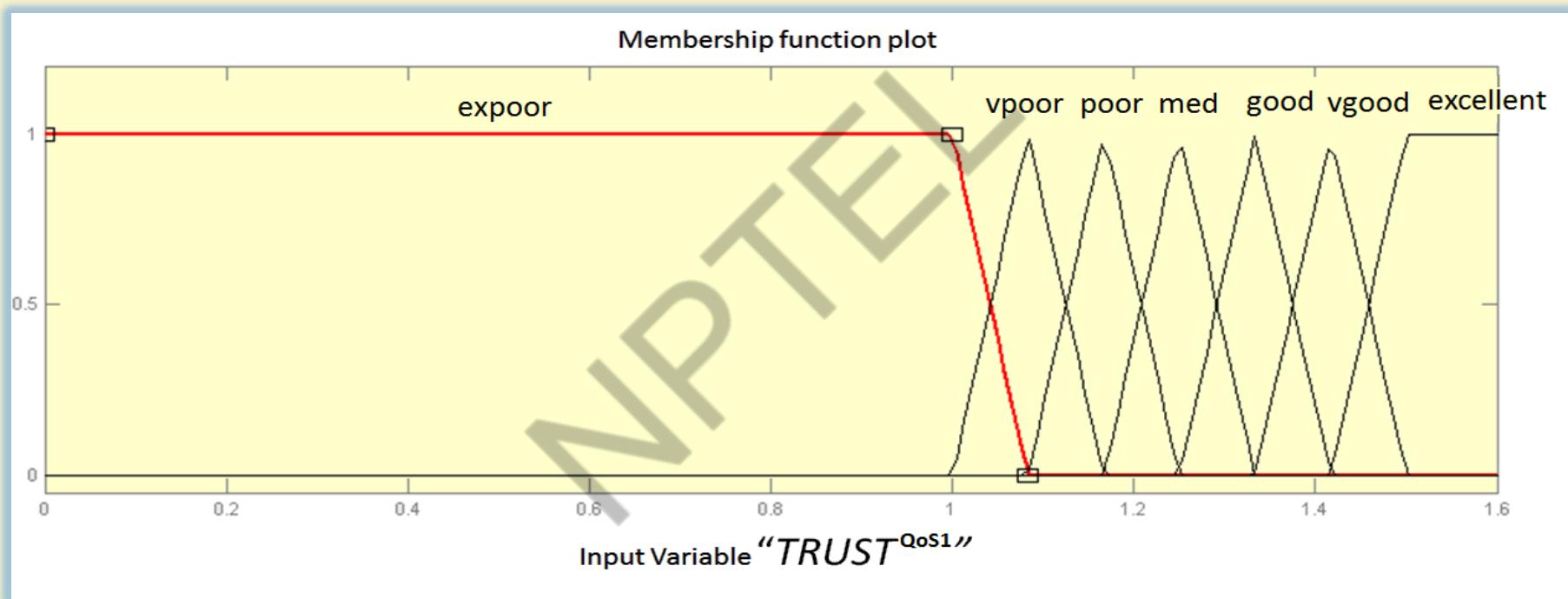
IIT KHARAGPUR

9/20/2017



NPTEL ONLINE
CERTIFICATION COURSES

PROVIDER SELECTION – INPUT MEMBERSHIP FUNCTION



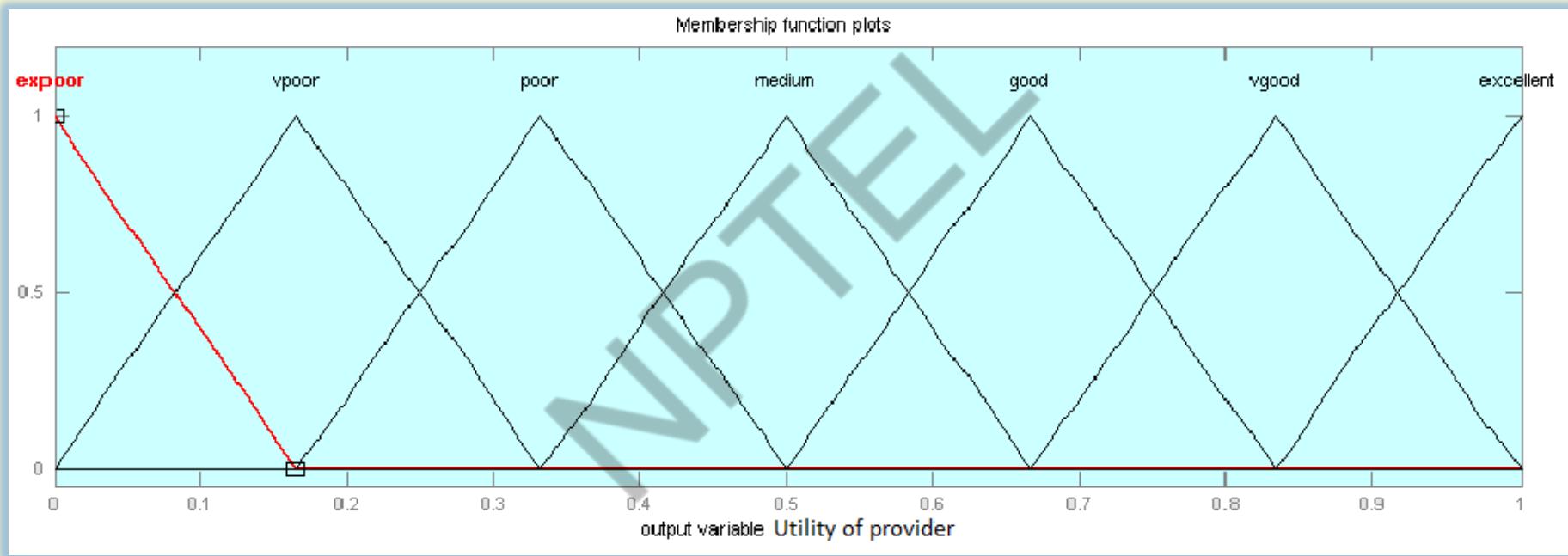
IIT KHARAGPUR

9/20/2017



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

PROVIDER SELECTION – OUTPUT MEMBERSHIP FUNCTION



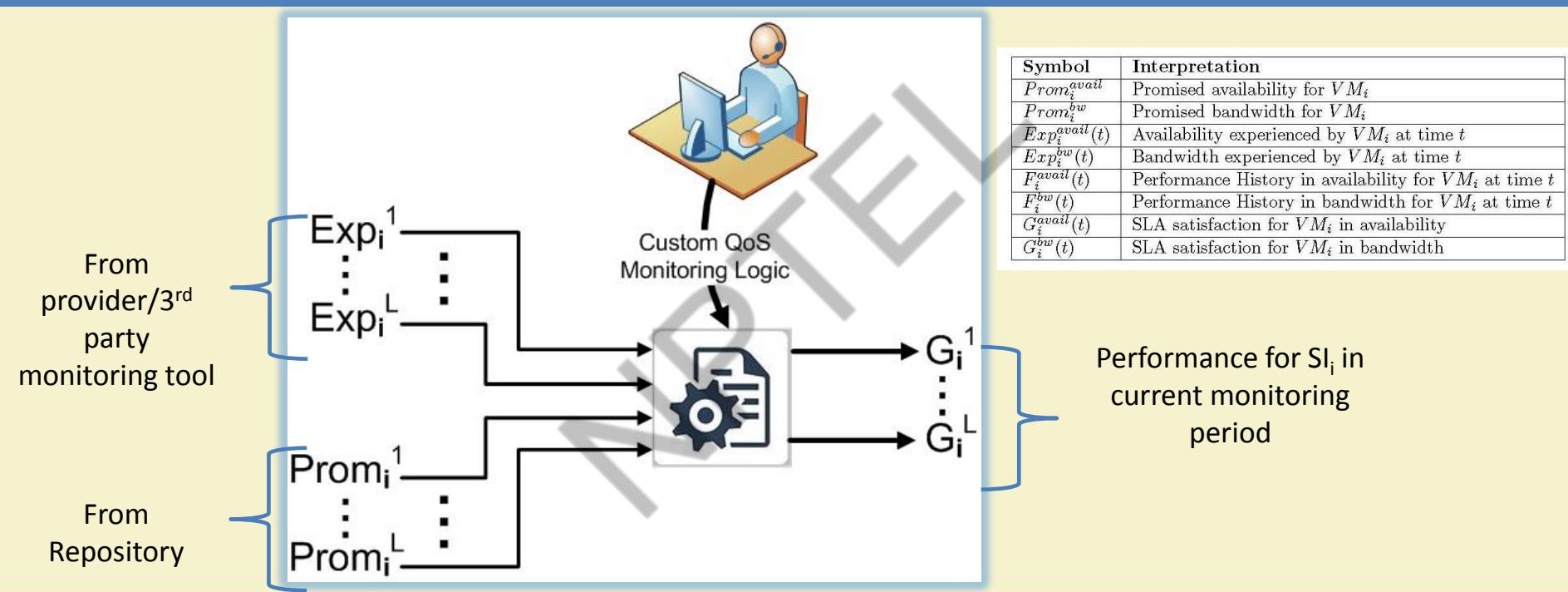
IIT KHARAGPUR

9/20/2017



NPTEL ONLINE
CERTIFICATION COURSES

MONITORING MODULE



Migration Decider

- 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



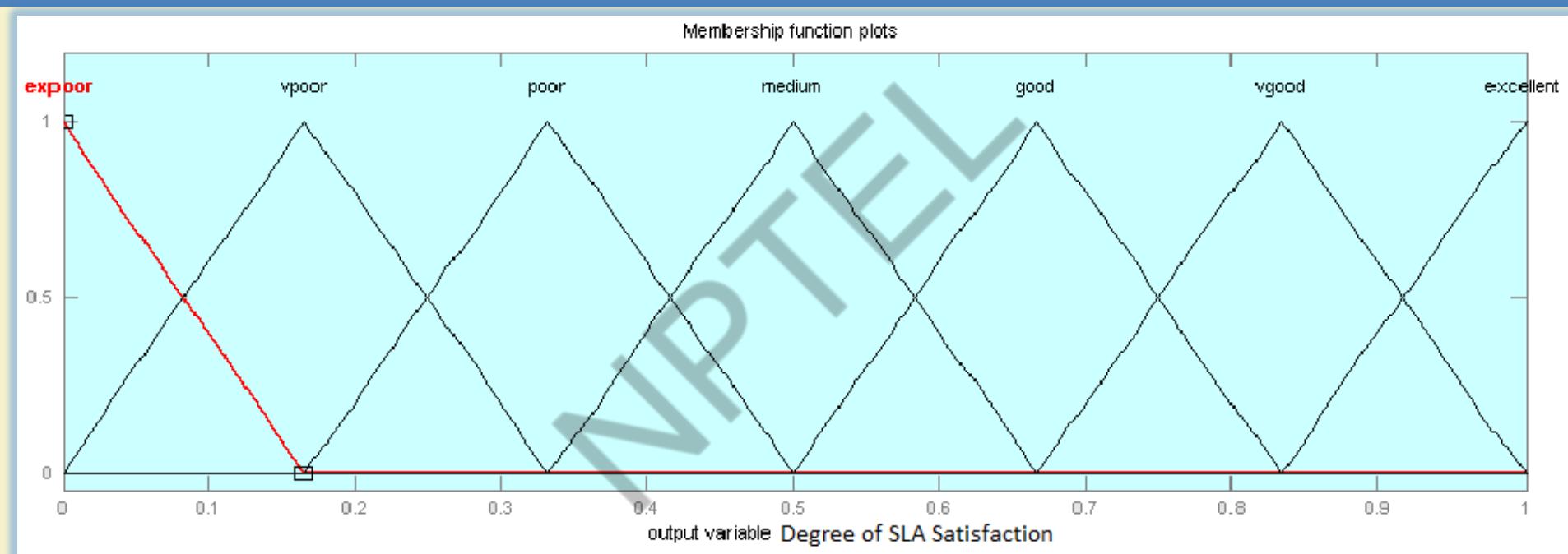
IIT KHARAGPUR

9/20/2017



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

MIGRATION DECIDER – OUTPUT MEMBERSHIP FUNCTION



IIT KHARAGPUR

9/20/2017



NPTEL ONLINE
CERTIFICATION COURSES

MIGRATION MODULE - SELECTION OF TARGET PROVIDER

- Similar to provider selection
- Selection done using a fuzzy inference engine



IIT KHARAGPUR

9/20/2017



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Case study on IaaS Marketplace

- 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



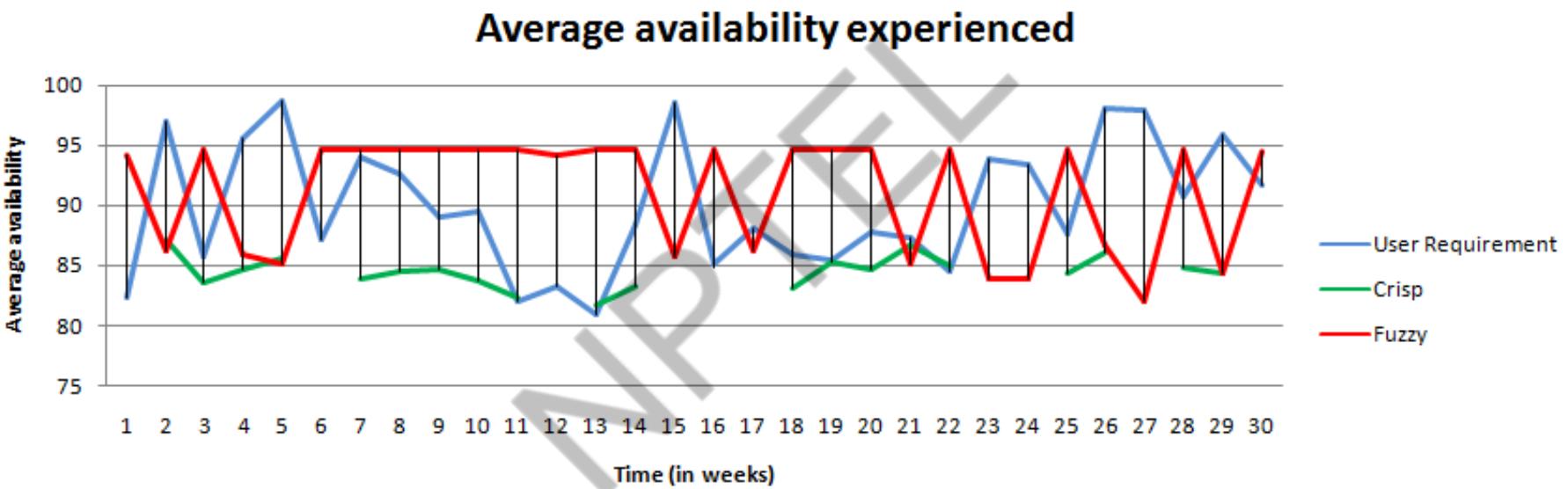
IIT KHARAGPUR

9/20/2017



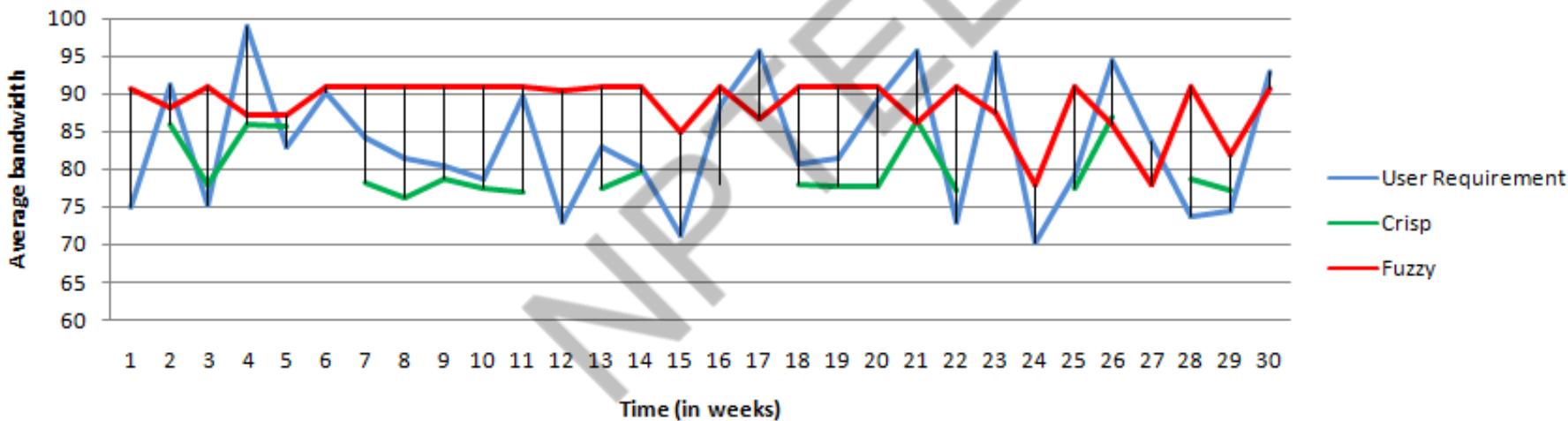
NPTEL ONLINE
CERTIFICATION COURSES

EXPERIMENTS AND RESULTS



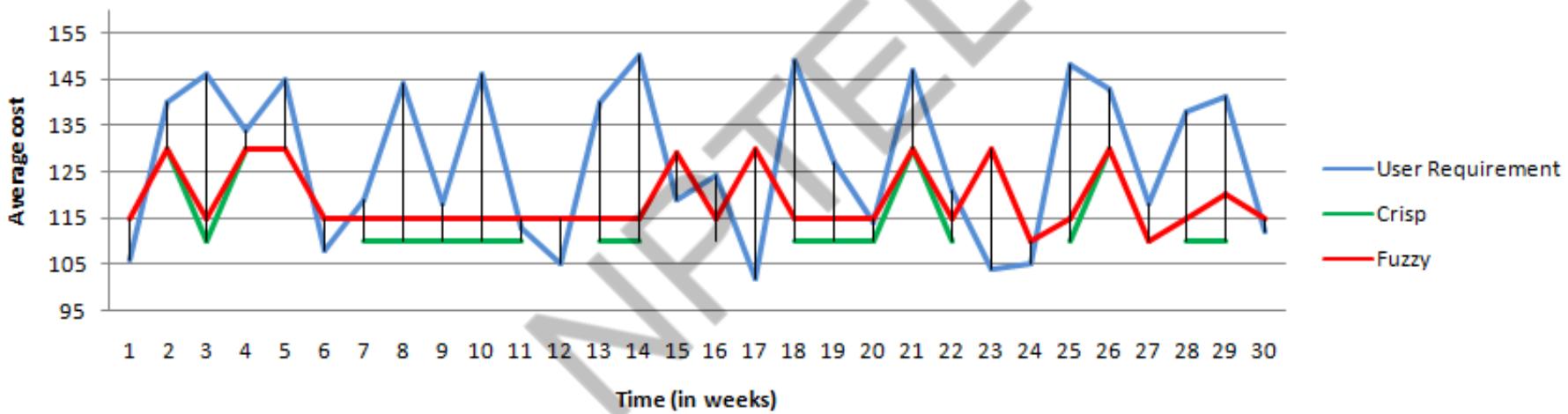
EXPERIMENTS AND RESULTS

Average bandwidth experienced



EXPERIMENTS AND RESULTS

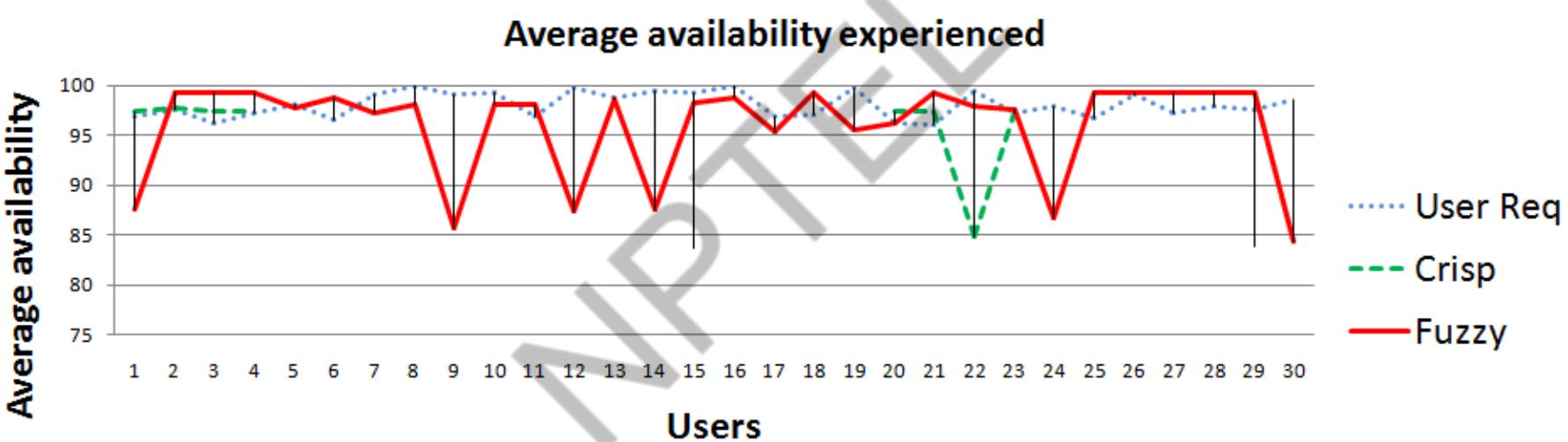
Average cost per VM per hour



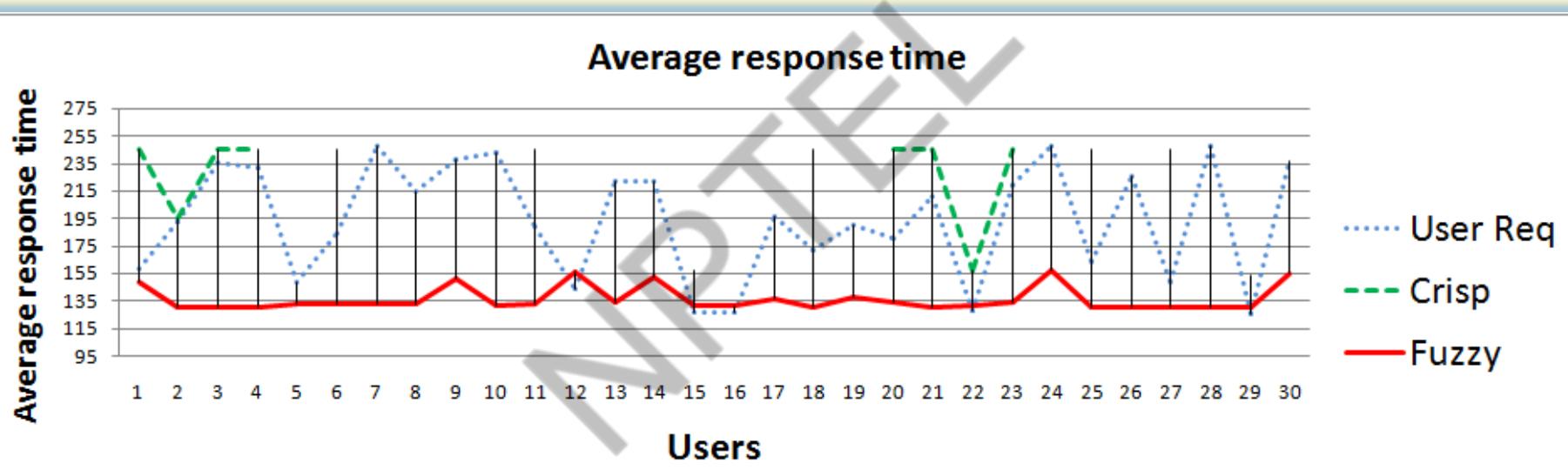
Case study on SaaS Marketplace

- 10 providers with varying offered QoS
- 500 service requests
- Year long simulation
- Few providers exhibit performance degradation. Degraded QoS parameters follow a Gaussian distribution
- Comparison made with conventional (minimum cost) crisp broker

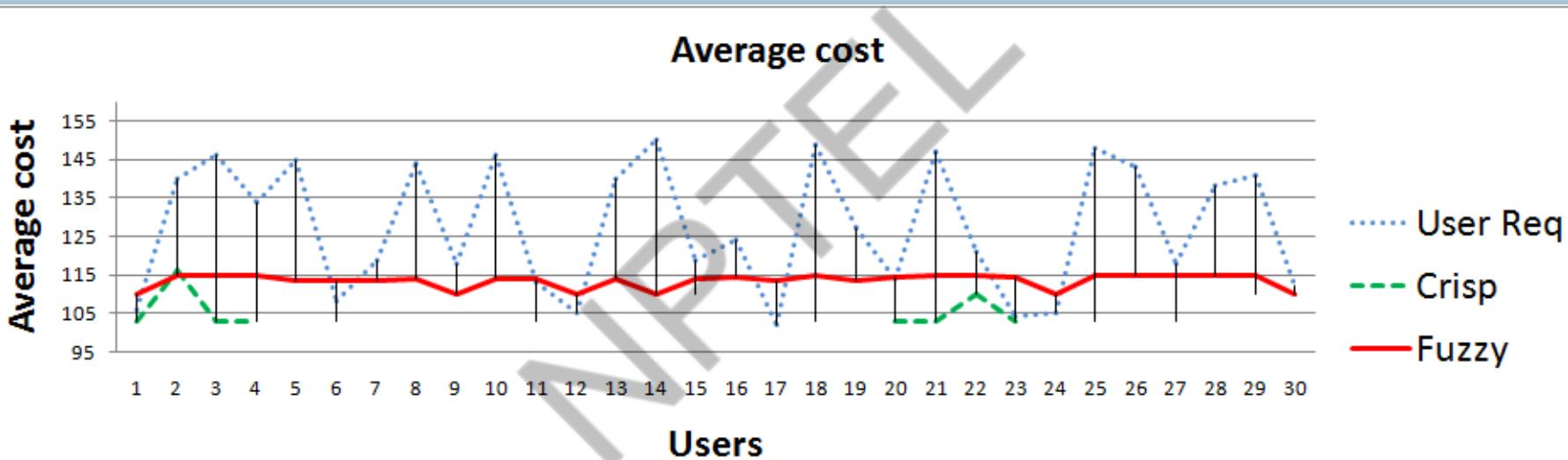
EXPERIMENTS AND RESULTS



Experiments and Results



EXPERIMENTS AND RESULTS



Future Scope

- Specification of flexibility in QoS requirements
- Comparison against existing approaches on production workload
- Service classes for customers



IIT KHARAGPUR

9/20/2017



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Thank You!!

NPTEL



IIT KHARAGPUR

9/20/2017



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

CLOUD COMPUTING

Mobile Cloud Computing - I

Prof. Soumya K Ghosh

Department of Computer Science and Engineering
IIT KHARAGPUR



Motivation

- *Growth in the use of Smart phones, apps*
- *Increased capabilities of mobile devices*
- *Access of internet using Mobile devices than PCs!*



- *Resource challenges (battery life, storage, bandwidth etc.) in mobile devices??*
- *Cloud computing offers advantages to users by allowing them to use infrastructure, platforms and software by cloud providers at low cost and elastically in an on-demand fashion*

“Information at your fingertips anywhere anytime..”

MobileBackend-as-a-service

What	<ul style="list-style-type: none">Provides mobile application developers a way to connect their application to backend cloud storage and processing
Why	<ul style="list-style-type: none">Abstract away complexities of launching and managing own infrastructureFocus more on front-end development instead of backend functions
When	<ul style="list-style-type: none">Multiple Apps, Multiple Backends, Multiple DevelopersMultiple Mobile Platforms, Multiple Integration, Multiple 3rd Party Systems & Tools
How	<ul style="list-style-type: none">Meaningful resources for app development acceleration – 3rd party API, Device SDK's, Enterprise Connectors, Social integration, Cloud storage

<http://www.rapidvaluesolutions.com/whitepapers/How-MBaaS-is-Shaping-up-Enterprise-Mobility-Space.html>

Augmenting Mobiles with Cloud Computing

- 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

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 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
 - ✓ PCs and servers
 - ✓ Accessed over the wireless connection based on a thin native client

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
NPTEL ONLINE
CERTIFICATION COURSES

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

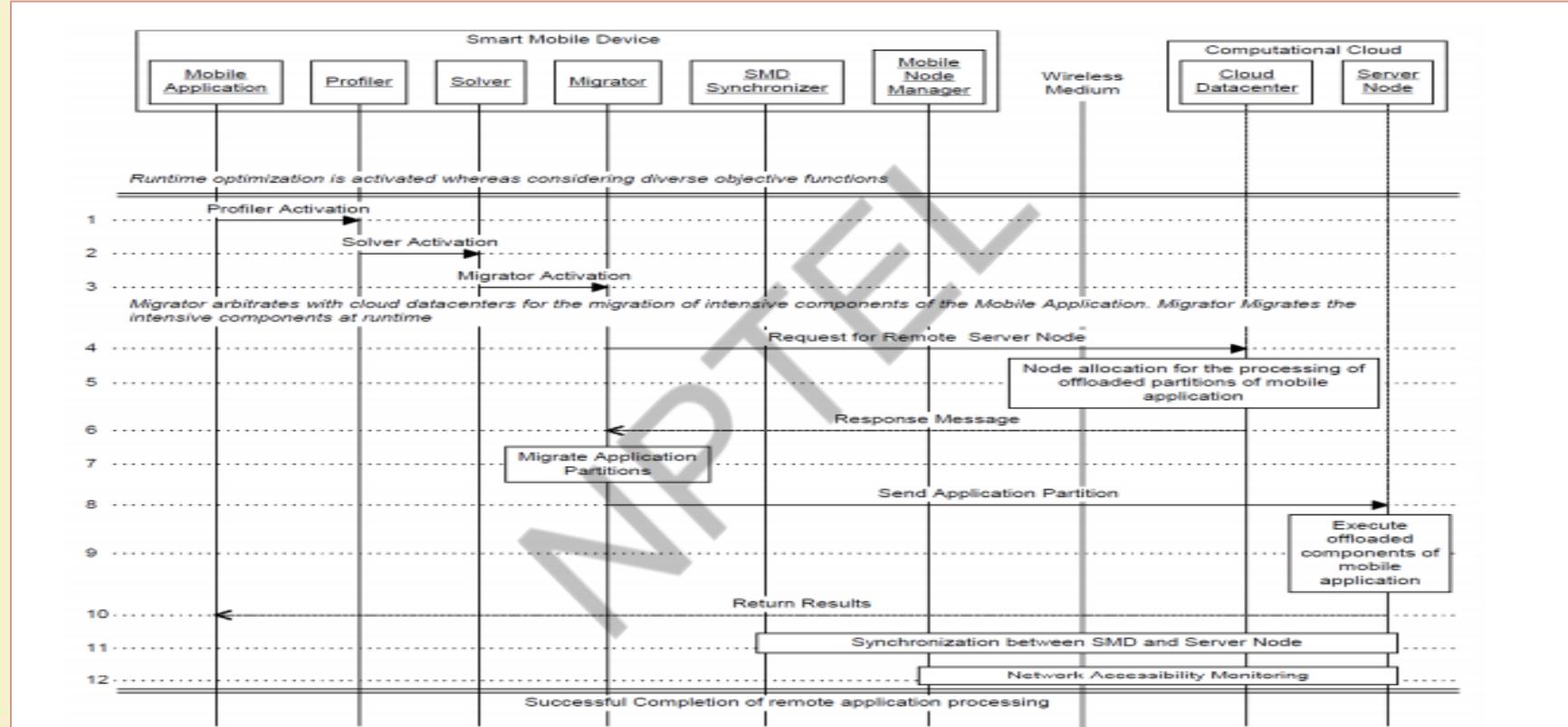
Mobile Cloud Computing

Wireless Network Technology

<u>Pros</u>	<u>Cons</u>
Saves battery power	Must send the program states (data) to the cloud server, hence consumes battery
Makes execution faster	Network latency can lead to execution delay

Mobile Cloud Computing is a framework to augment a resource constrained mobile device to execute parts of the program on cloud based servers

Typical MCC Workflow



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.

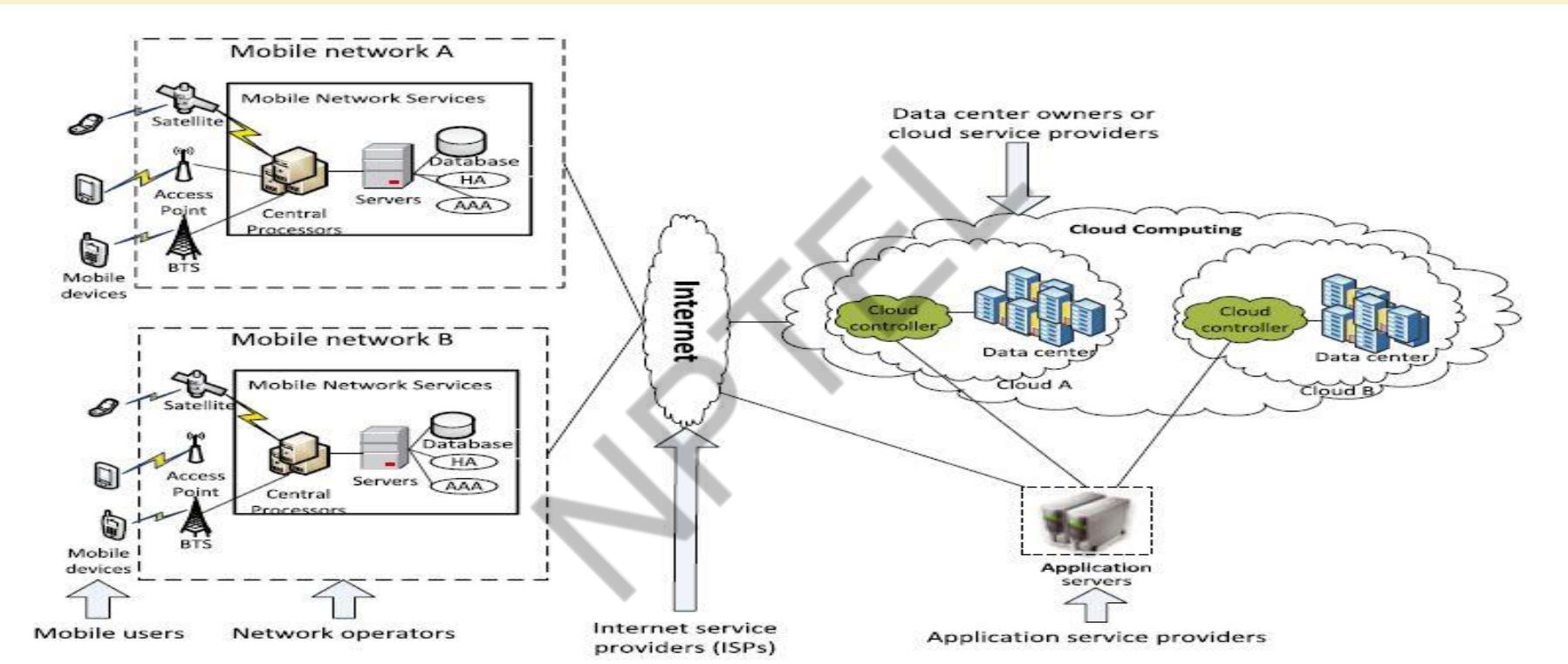
MCC key components

- 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

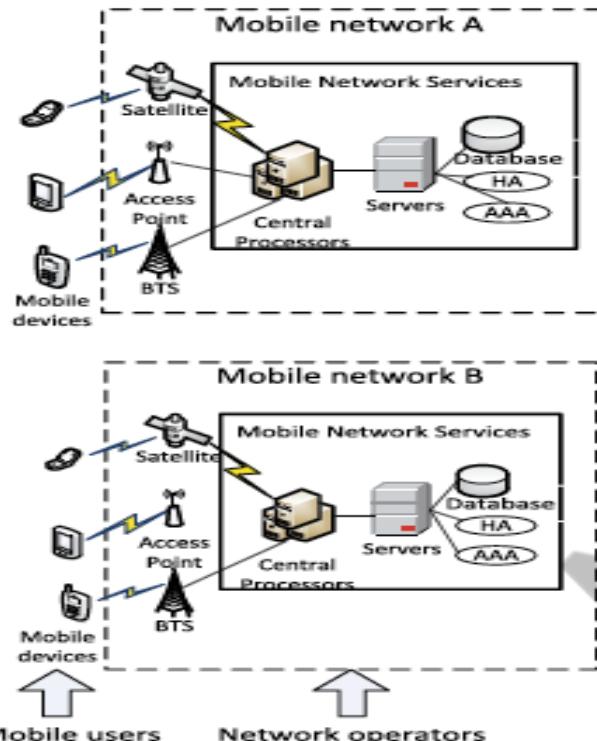
Key Requirements for MCC

- *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

Mobile Cloud Computing – Typical Architecture



Mobile Cloud Computing - Architecture



Mobile devices are connected to the mobile networks via base stations that establish and control the connections and functional interfaces between the networks and mobile devices

Mobile users' requests and information are transmitted to the central processors that are connected to servers providing mobile network services

Data center owners or cloud service providers



Cloud A

Cloud B

Internet service providers (ISPs)



Application service providers

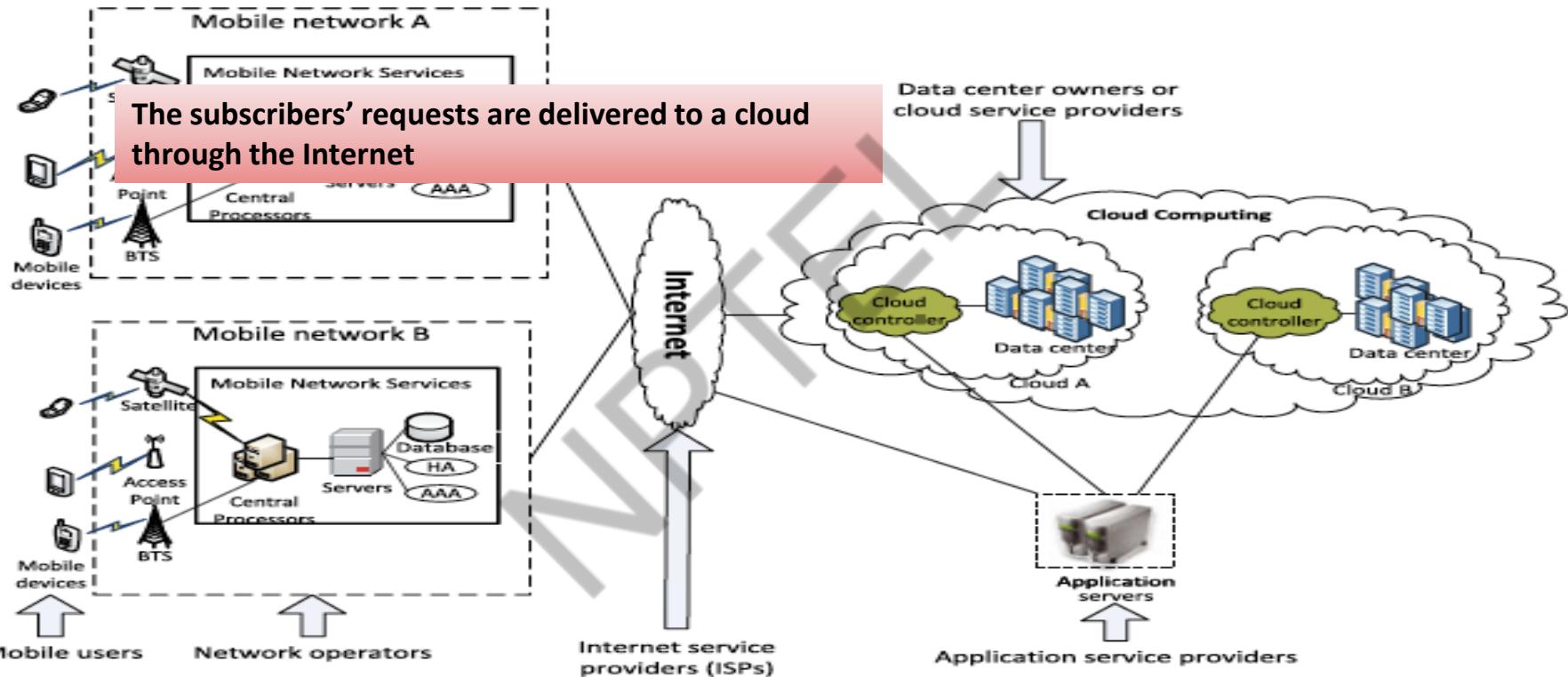


IIT KHARAGPUR

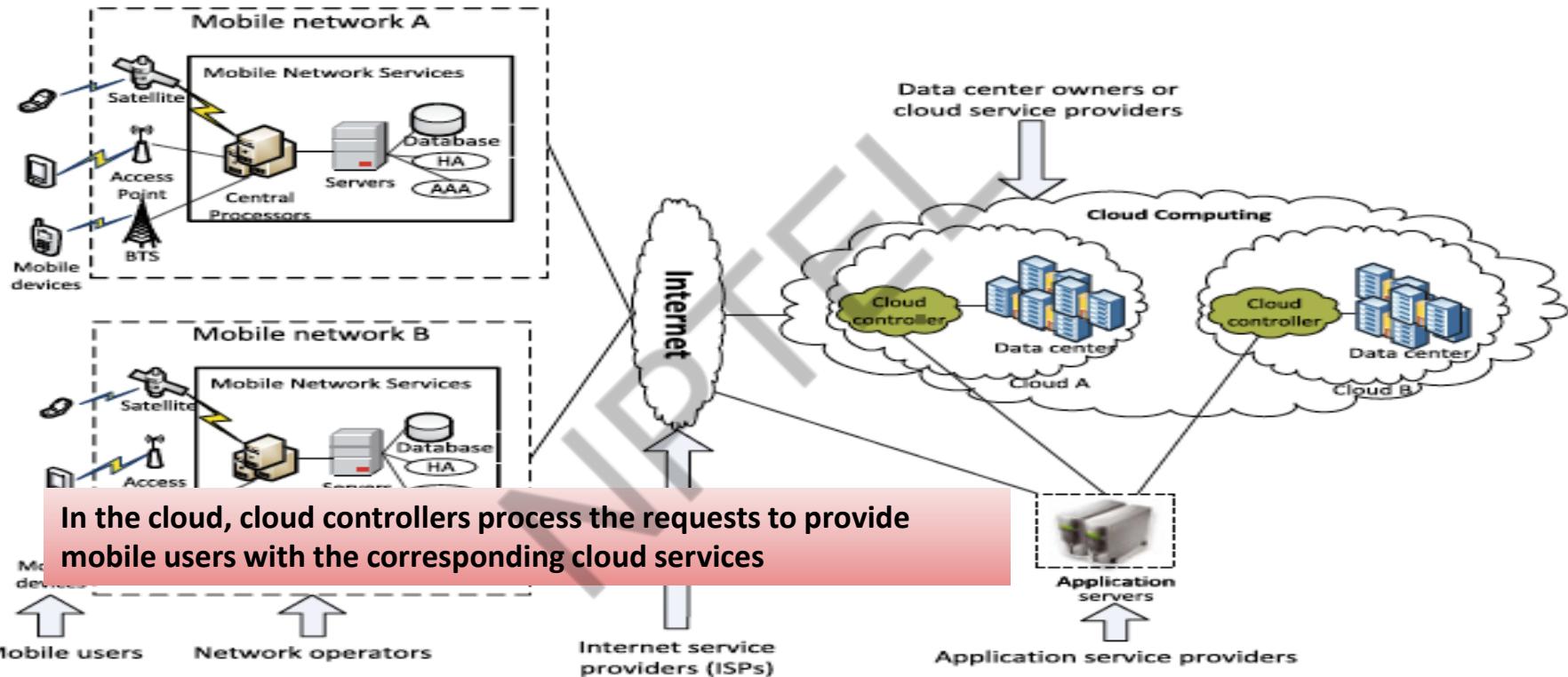


NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Mobile Cloud Computing - Architecture



Mobile Cloud Computing - Architecture



Advantages of MCC

Extending battery lifetime

- Computation offloading migrates large computations and complex processing from resource-limited devices (i.e., mobile devices) to resourceful machines (i.e., servers in clouds).
- Remote application execution can save energy significantly.
- Many mobile applications take advantages from task migration and remote processing

Improving data storage capacity and processing power

- MCC enables mobile users to store/access large data on the cloud.
- MCC helps reduce the running cost for computation intensive applications.
- Mobile applications are not constrained by storage capacity on the devices because their data now is stored on the cloud

Advantages of MCC (contd...)

Improving Reliability and Availability

- Keeping data and application in the clouds reduces the chance of lost 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.

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.

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

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.

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.

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. La 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.

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.

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.

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

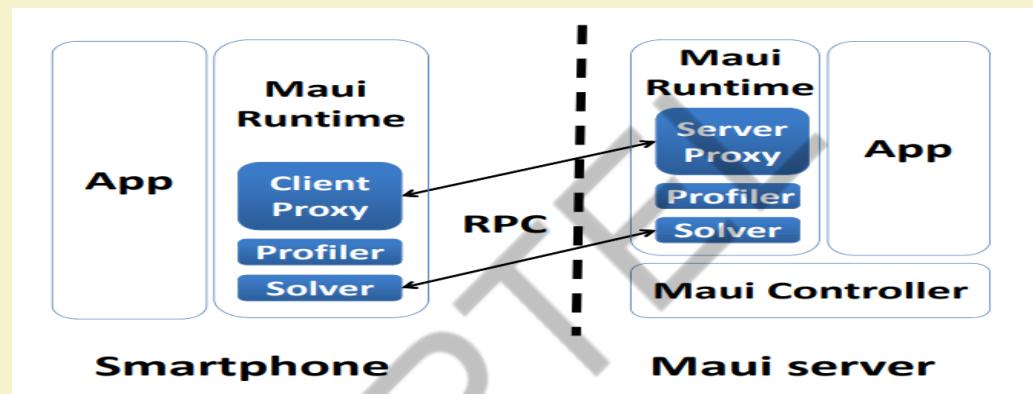


IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

MCC Systems: MAUI (Mobile Assistance Using Infrastructure)

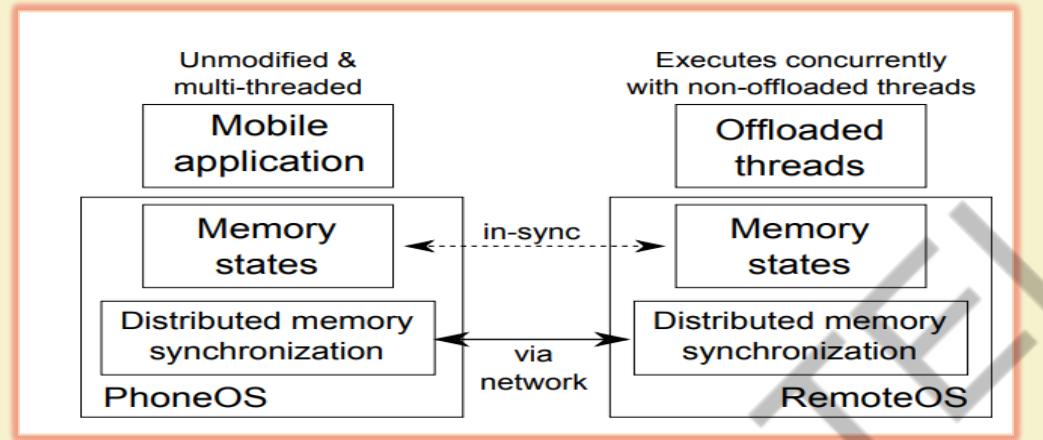


- **MAUI enables the programmer to produce an initial partition of the program**
 - Programmer marks each method as “remoteable” • or not
 - Native methods cannot be remoteable
- MAUI framework uses the annotation to decide

whether a method should be executed on cloud server to save energy and time to execute

MAUI server is the cloud component. The framework has the necessary software modules required in the workflow.

MCC Systems: COMET



- Requires only program binaries Execute multi-threaded programs correctly Improve speed of computation
- Further improvements to data traffic during migration is also possible by sending only the parts of the heap that has been modified

COMET: Code Offload by Migrating Execution Transparently

- Works on unmodified applications (no source code required)
- Allows threads to migrate between machines depending on workload
- It implements a Distributed Shared Memory (DSM) model for the runtime engine
 - ✓ *DSM allows transparent movement of threads across machines*
 - ✓ *In computer architecture, DSM is a form of memory architecture where the (physically separate) memories can be addressed as one (logically shared) address space*

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
NPTEL ONLINE
CERTIFICATION COURSES

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.

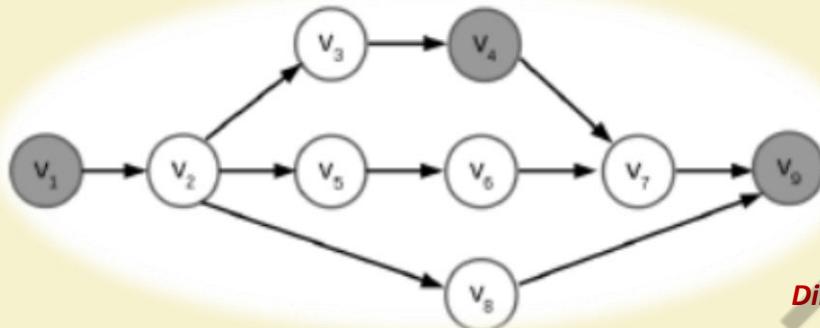


IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Mathematical Formulation



Directed Acyclic Graph represents an application Call Graph

$$\text{maximize } \sum_{v \in V} I_v \times E_v^l - \sum_{(u,v) \in E} |I_u - I_v| \times C_{u,v}$$

$$\text{such that: } \sum_{v \in V} ((1 - I_v) \times T_v^l) + (I_v \times T_v^r)$$

$$+ \sum_{(u,v) \in E} (|I_u - I_v| \times B_{u,v}) \leq L$$

$$\text{and } I_v \leq r_v, \forall v \in V$$

- Highlighted nodes must be executed on the mobile device -> called native tasks (v_1, v_4, v_9)
- Edges represent the sequence of execution - Any non-highlighted node can be executed either locally on the mobile device or on cloud

- 0-1 integer linear program,
where $I_v = 0$ if method executed locally,
 $= 1$ if method executed remotely
- E : Energy cost to execute method v locally
- $C_{u,v}$: Cost of data transfer
- L : Total execution latency
- T : Time to execute the method
- B : Time to transfer program state

Integer Linear Program to solve the Task Partitioning Problem

Mathematical Formulation (Contd..)

- 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

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
NPTEL ONLINE
CERTIFICATION COURSES

CODE OFFLOADING USING CLOUDLET

- **CLOUDLET:**

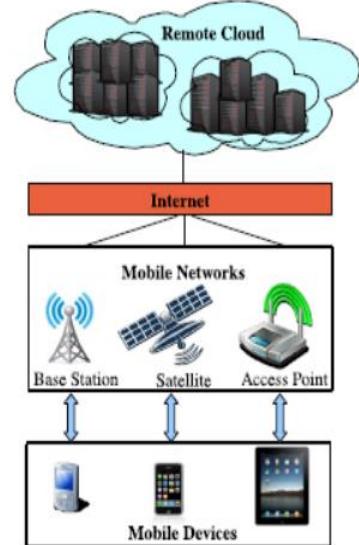
- ✓ “*a trusted, resource-rich computer or cluster of computers that is well-connected to the Internet and is available for use by nearby mobile devices.*”

- **Code Offloading :**

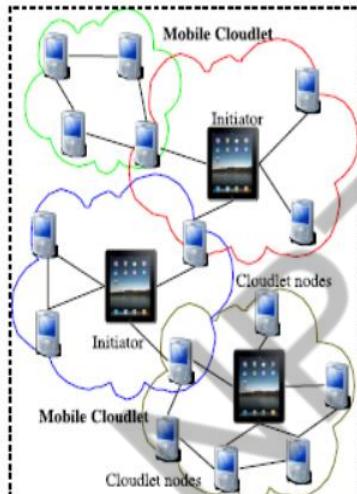
- ✓ Offloading the code to the remote server and executing it.
 - ✓ This architecture decreases latency by using a single-hop network and potentially lowers battery consumption by using Wi-Fi or short-range radio instead of broadband wireless which typically consumes more energy.

CODE OFFLOADING USING CLOUDLET

Cloudlet

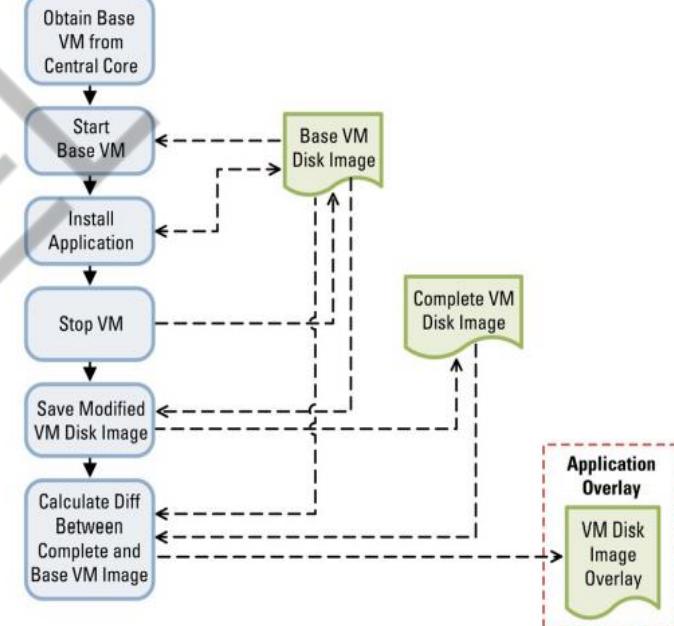


Use remote cloud



Use cloudlet

Application Overlay Creation Process



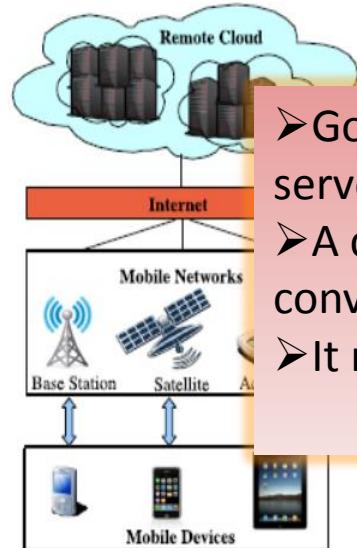
IIT KHARAGPUR



NPTEL
ONLINE
CERTIFICATION COURSES

CODE OFFLOADING USING CLOUDLET

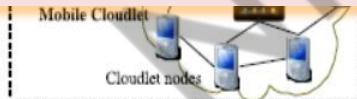
Cloudlet



Use remote cloud

- 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

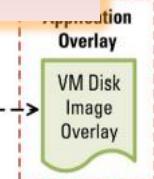


Use cloudlet

Application Overlay Creation Process

Obtain Base VM from Central Core

Calculate Diff Between Complete and Base VM Image



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

When to Offload??

The amount of energy saved is :

$$P_c \times \frac{C}{M} - P_i \times \frac{C}{S} - P_{tr} \times \frac{D}{B}$$

S: the speed of cloud to compute C instructions

M: the speed of mobile to compute C instructions

D: the data need to transmit

B: the bandwidth of the wireless Internet

P_c: the energy cost per second when the mobile phone is doing computing

P_i: the energy cost per second when the mobile phone is idle.

P_{tr}: the energy cost per second when the mobile is transmission the data.

Suppose the server is F times faster—that is, S= F × M.

We can rewrite the formula as

$$\frac{C}{M} \times (P_c - \frac{P_i}{F}) - P_{tr} \times \frac{D}{B}$$

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.

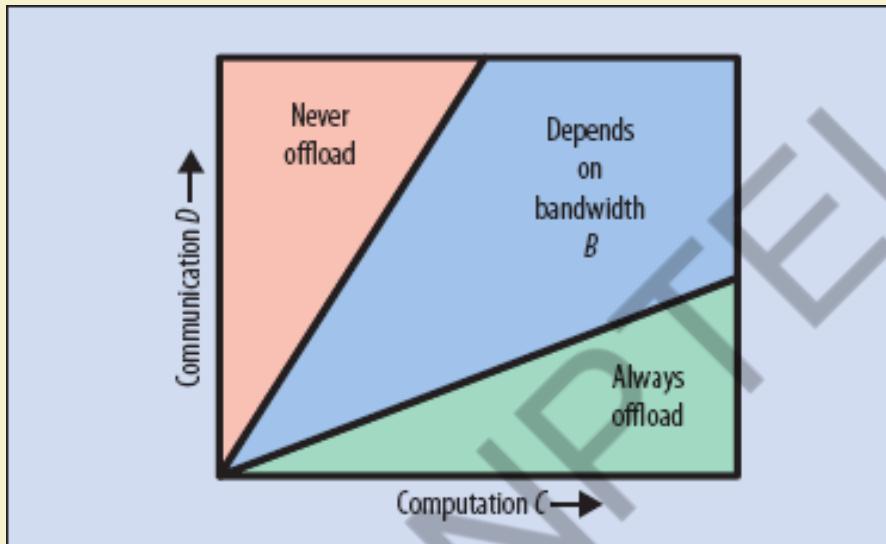
The amount of energy saved is :

$$P_c \times \frac{C}{M} - P_i \times \frac{C}{S} - P_{tr} \times \frac{D}{B}$$

We can rewrite the formula as

$$\frac{C}{M} \times (P_c - \frac{P_i}{F}) - P_{tr} \times \frac{D}{B}$$

When to Offload?? (contd..)



Offloading is beneficial when large amounts of computation C are needed with relatively small amounts of communication D

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

$$\frac{C}{M} \times (P_c - \frac{P_i}{F}) - P_{tr} \times \frac{D}{B}$$

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 Intl 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

How to evaluate MCC performance

- Energy Consumption
 - Must reduce energy usage and extend battery life
- Time to Completion
 - Should not take longer to finish the application compared to local execution
- Monetary Cost
 - Cost of network usage and server usage must be optimized
- Security
 - As offloading transfers data to the servers, ensure confidentiality and privacy of data, how to identify methods which process confidential data

Open Questions?

- How can one design a practical and usable MCC framework
 - System as well as partitioning algorithm
- Is there a scalable algorithm for partitioning
 - Optimization formulations are NP-hard
 - Heuristics fail to give any performance guarantee
- Which are the most relevant parameters to consider in the design of MCC systems?



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

Mobile Cloud Computing – Applications?

Mobile Health-care



Health-Monitoring services, Intelligent emergency management system, Health-aware mobile devices (detect pulse rate, blood pressure, alcohol-level etc.)

Mobile Gaming



It can completely offload game engine requiring large computing resource (e.g., graphic rendering) to the server in the cloud



Mobile Commerce

M-commerce allows business models for commerce using mobile (Mobile financial, mobile advertising, mobile shopping)



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Mobile Cloud Computing – Applications?



Pedestrian crossing guide for blind and visually-impaired

Mobile currency reader for blind and visually impaired

Lecture transcription for hearing impaired students

Assistive Technologies



Mobile Learning

- *M-learning combines e-learning and mobility*
- *Traditional m-learning has limitations on high cost of devices/network, low transmission rate, limited educational resources*
- *Cloud-based m-learning can solve these limitations*
- *Enhanced communication quality between students and teachers*
- *Help learners access remote learning resources*
- *A natural environment for collaborative learning*



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

MuSIC: Mobility-Aware Optimal Service Allocation in Mobile Cloud Computing

- User Mobility introduces new complexities in enabling an optimal decomposition of tasks that can execute cooperatively on mobile clients and the tiered cloud architecture while considering multiple QoS goals such application delay, device power consumption and user cost/price.
- Apart from scalability and access issues with the increased number of users, mobile applications are faced with increased *latencies* and reduced *reliability*
- As a user moves, the physical distance between the user and the cloud resources originally provisioned changes causing additional delays
- Further, the lack of effective handoff mechanisms in WiFi networks as user move rapidly causes an increase in the number of *packet losses*

In other words, user mobility, if not addressed properly, can result in suboptimal resource mapping choices and ultimately in diminished application QoS

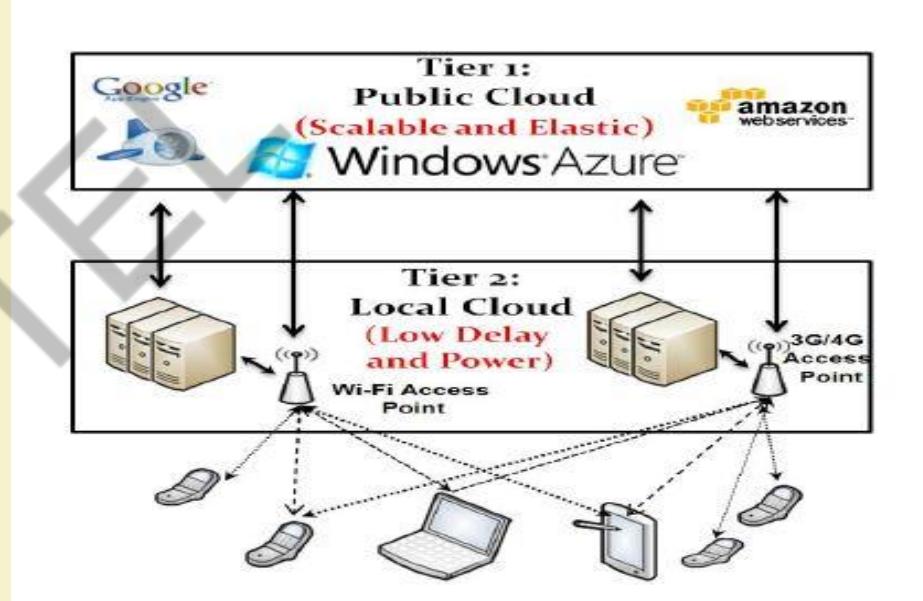
MuSIC: Mobility-Aware Optimal Service Allocation in Mobile Cloud Computing

Efficient techniques for *dynamic mapping of resources* in the presence of **mobility**; using a *tiered cloud architecture*, to meet the *multidimensional QoS* needs of mobile users

- Location-time workflow (LTW) as the modeling framework to model mobile applications and capture user mobility. Within this framework, mobile service usage patterns as a function of location and time has been formally modelled
- Given a mobile application execution expressed as a LTW, the framework optimally partitions the execution of the location-time workflow in the 2-tier architecture based on a *utility metric* that combines *service price, power consumption and delay* of the mobile applications

MuSIC: Mobility-Aware Optimal Service Allocation in Mobile Cloud Computing

- ✓ Tier 1 nodes in the system architecture represents *public cloud services* such as Amazon EC2, Microsoft Azure and Google AppEngine. Services provided by these vendors are highly *scalable* and *available*; what they lack is the ability to provide the *fine grain location granularity* required for high performance mobile applications
- ✓ This feature is provided by the second tier local cloud, that consists of nodes that are connected to access points.
- ✓ Location information of these services are available at finer levels of granularity (campus and street level).
- ✓ Mobile users are typically connected to these local clouds through WiFi (via access points) or cellular (via 3G cell towers) connectivity - the aim to intelligently select which local and which public cloud resources to utilize for task offloading.



2-Tier Mobile Cloud Architecture

Mobile Application Modelling

Cloud Service Set:

The set of all services (e.g. compute, storage and software capabilities like multimedia streaming services, content transcoding services, etc) provided by local and public cloud providers

Local Cloud Capacity:

Local cloud services can only accept a limited number of mobile client requests

Location Map:

It is a partition of the 2-D space/region in which mobile hosts and cloud resources are located

User Service Set:

The set of all services that a user has on his own device (e.g. decoders, image editors etc.)

Criteria	Definition
$q_{price}(s_i, u_k^{l_i, t_j})$	The price of using service s_i when user u_k is in location $l_i \in L$ and time t_j .
$q_{power}(s_i, u_k^{l_i, t_j})$	The power consumed on user mobile device using s_i when user u_k is in location $l_i \in L$ and time t_j .
$q_{delay}(s_i, u_k^{l_i, t_j})$	The delay of executing service s_i when user u_k is in location $l_i \in L$ and time t_j .

Mobile User Trajectory:

The trajectory of a mobile user, u_k , is represented as a list of tuples of the form $\{(1; l_1); \dots; (n; l_n)\}$ where $(i; l_i)$ implies that the mobile user is in location l_i for time duration i

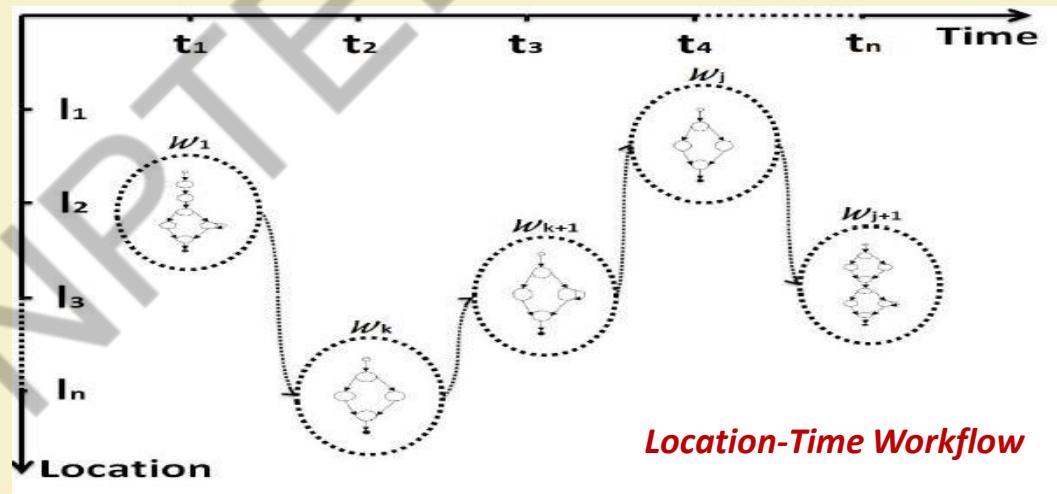
Center of Mobility:

It is the location where (or near where) a mobile user u_k spends most of its time

Mobile Application Modelling

Location-Time Workflow

Combination of the mobile application workflow concept with a user trajectory to model the mobile users and the requested services in their trajectory.



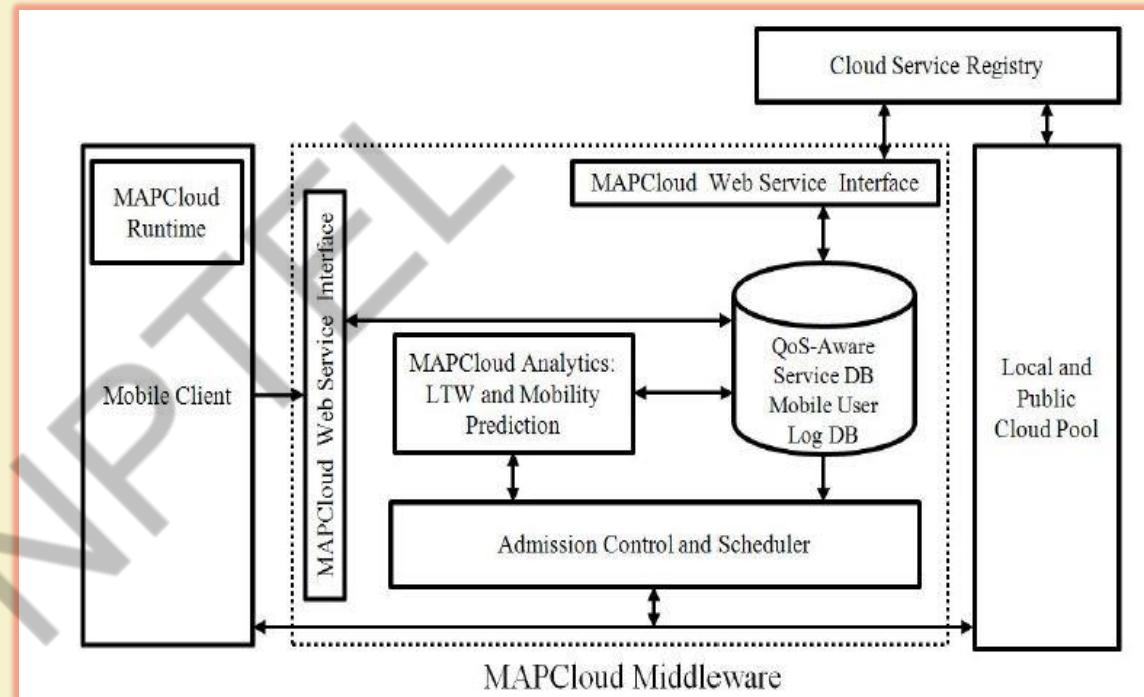
Mobile Application Modelling

Mobile User Log DB and QoS-Aware Service DB:

Unprocessed user data log such as mobile service usage, location of the user, user delay experience of getting the service, energy consumed on user mobile device, etc and service lists on local and public cloud and their QoSes in different locations respectively

MAPCloud Analytic: This module processes mobile user Log DB and updates QoS-aware cloud service DB based on user experience and LTW

Admission Control and Scheduling: This module is responsible for optimally allocate services to admitted mobile users based on MuSIC



A Case Study: Context Aware Dynamic Parking Service

- MCC can provide a flexible method of handling massive computing, storage, and software services in a scalable and virtualized manner.
- The integration of MCC and vehicular networks is expected to promote the development of cost effective, scalable, and data-driven CVC (Context-aware vehicular cyber physical systems)

An application scenario regarding the context-aware dynamic parking services by illuminating the cloud-assisted architecture and logic flow.

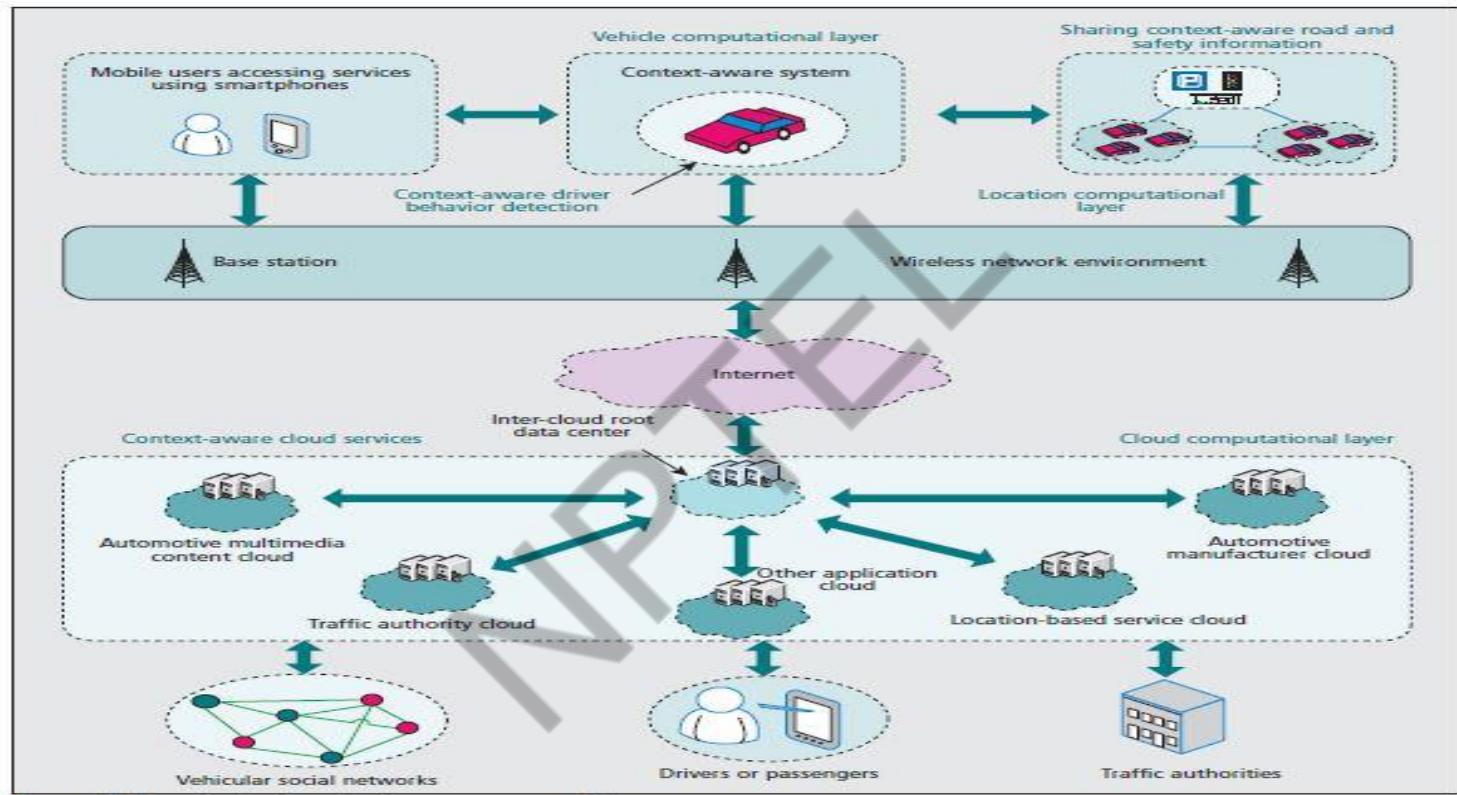
- As the number of vehicles increases, there is an increasing trend of insufficient parking spaces in many large cities, and this problem is gradually getting worse
- With the proliferation of wireless sensor networks (WSNs) and cloud computing, there exists strong potential to alleviate this problem using context information (e.g., road conditions and status of parking garages) to provide context-aware dynamic parking services
- Cloud Assisted parking services (traditional parking garages and dynamic parking services along the road) and parking reservation service using smart terminals such as smartphones.

A Case Study: Context Aware Dynamic Parking Service

- MCC can provide a flexible method of handling massive computing, storage, and software services in a scalable and virtualized manner.
- The integration of MCC and vehicular networks is expected to promote the development of cost effective, scalable, and data-driven CVC (Context-aware vehicular cyber physical systems)

An application scenario regarding the context-aware dynamic parking services by illuminating the cloud-assisted architecture and logic flow.

- As the number of vehicles increases, there is an increasing trend of insufficient parking spaces in many large cities, and this problem is gradually getting worse
- With the proliferation of wireless sensor networks (WSNs) and cloud computing, there exists strong potential to alleviate this problem using context information (e.g., road conditions and status of parking garages) to provide context-aware dynamic parking services
- Cloud Assisted parking services (traditional parking garages and dynamic parking services along the road) and parking reservation service using smart terminals such as smartphones.



IIT KHARAGPUR

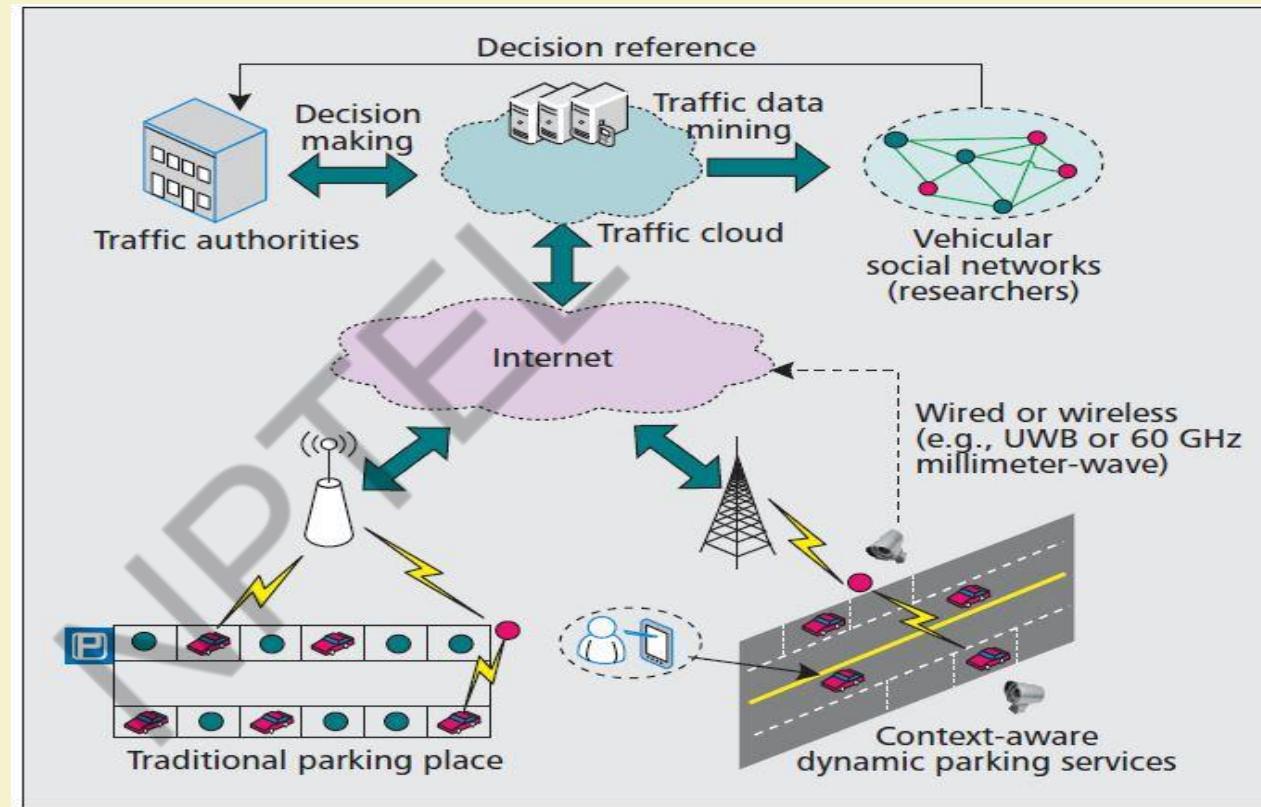


NPTEL
ONLINE
CERTIFICATION COURSES

Example cloud-assisted context-aware architecture

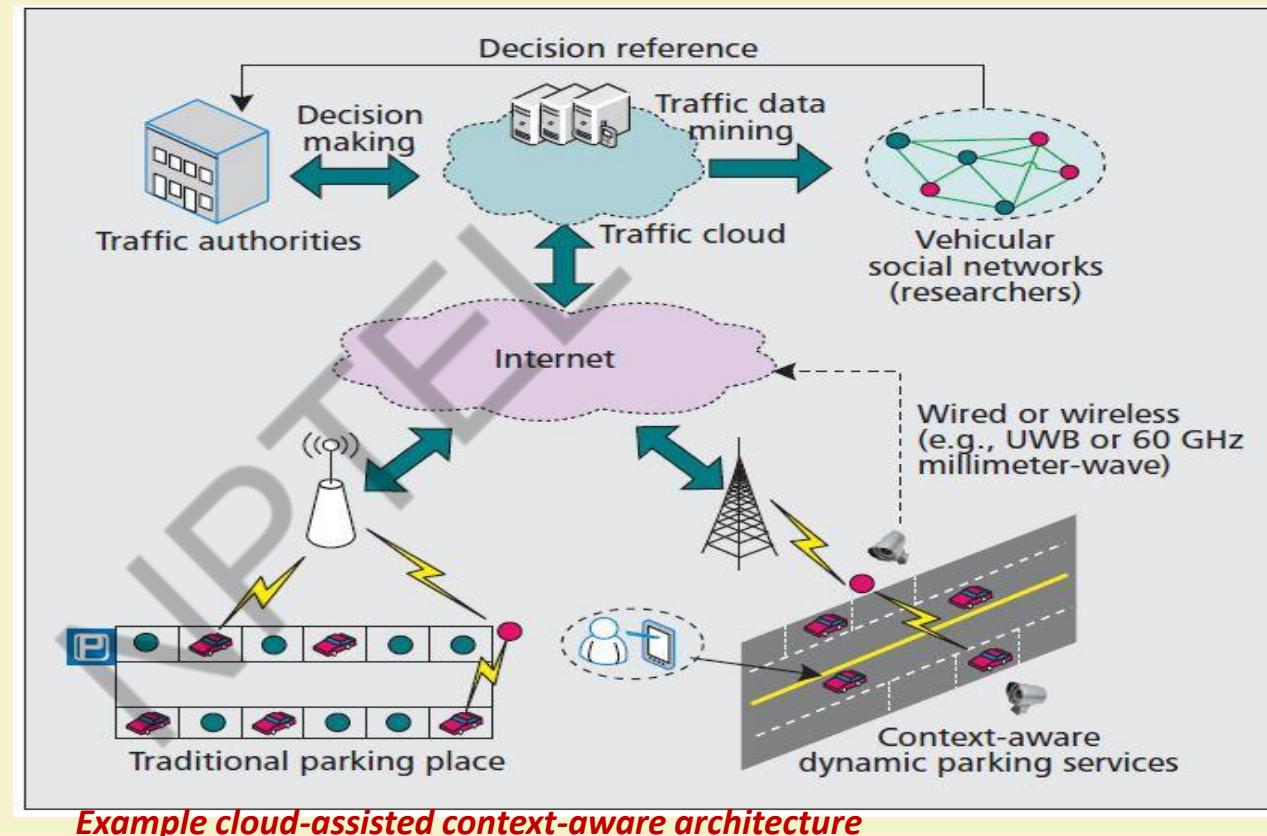
Traditional parking garages:

- The context information of each parking space detected by a WSN is forwarded to the traffic cloud by WSNs, third-generation (3G) communications, and the Internet.
- The collected data are processed in the cloud and then selectively transmitted to the users.
- This is helpful for providing more convenience services and evaluating the utilization levels of the parking garage.
- Also, the status of the parking garage may be dynamically published on a nearby billboard to users who have no ability to get the status by smart terminals.



Dynamic parking services:

- In this scenario, we consider a situation in which we may temporarily park a vehicle along the road if it does not impede the passage of other vehicles or pedestrians.
- We envision this application scenario based on the common observation that the traffic flow capacity is usually regular for each road. For example, there is usually heavy traffic during morning and evening rush hours.
- Therefore, considering the context information such as rush hours and road conditions, we may dynamically arrange the parking services for a very wide road.
- With the support of many new technologies (e.g., MCC and WSNs), the traffic authorities can carry out the dynamic management of this kind of service.



A Case Study: Context Aware Dynamic Parking Service

Three aspects, including service planning of traffic authorities, reservation service process, and context-aware optimization have been studied.

Decision making of traffic authorities

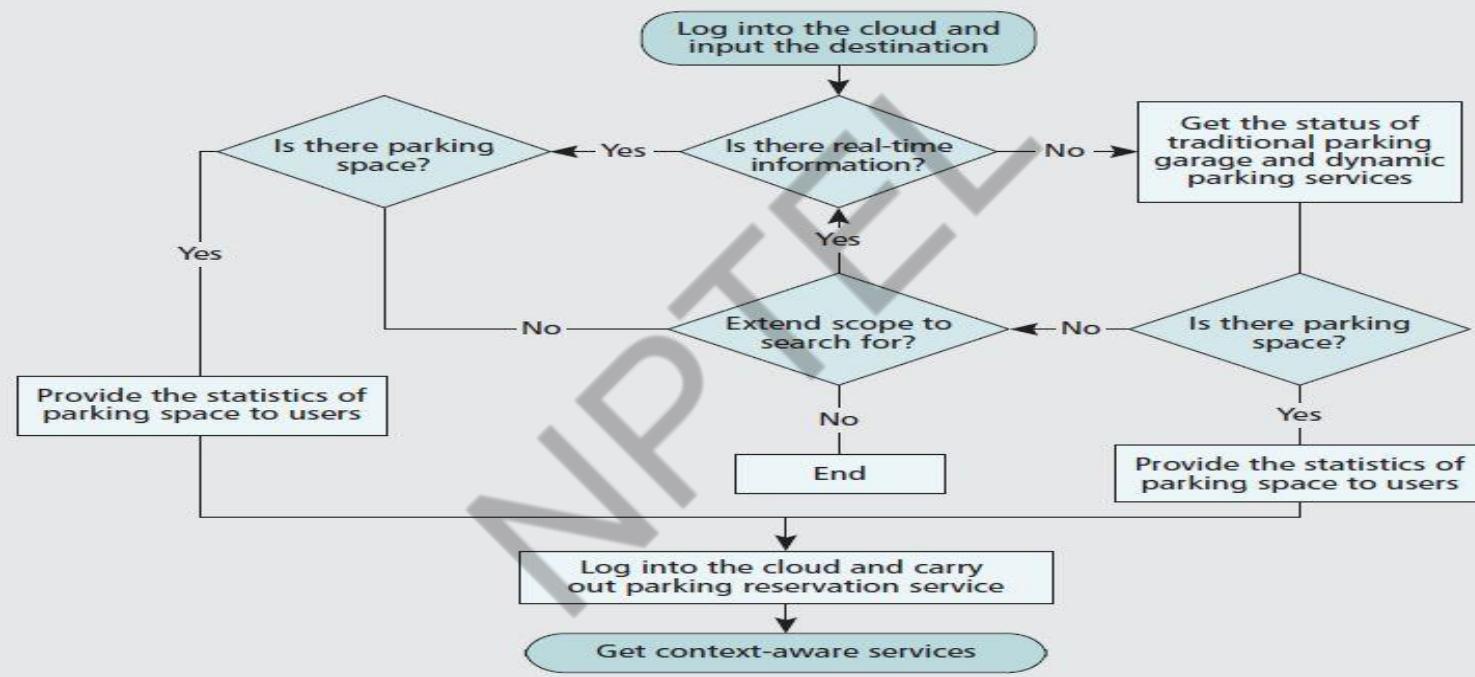
- The decision-making process of the proposed scheme heavily depends on many factors, such as historical traffic flow capacity, road conditions, weather conditions, and traffic flow forecasting
- In order to make an effective prediction, researchers on vehicular social networks carry out traffic data mining to discover useful information and knowledge from collected big data. The prediction process depends on classifying the influence factors and designing a decision tree
- By the method of probability analysis, the traffic authorities dynamically arrange whether the road can be authorized to provide context-aware parking services. In some particular cases, a fatal factor may directly affect the decision making. For example, when a typhoon is approaching, traffic authorities may immediately terminate services

A Case Study: Context Aware Dynamic Parking Service

Parking reservation services:

- The status of a parking space can be monitored as determined by the corresponding system, and subsequently updated in the traffic cloud.
- The drivers or passengers can quickly obtain the parking space's information by various smart terminals such as smartphones. If a proper parking space cannot be found, further search scope is extended.
- Within a given time, we may log into the traffic cloud and subscribe to a parking space.

A Case Study: Context Aware Dynamic Parking Service



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

A Case Study: Context Aware Dynamic Parking Service

Context-aware optimization:

- The context information includes not only road conditions and the status of the parking garage, but also the expected duration of parking as well.
- Since the purpose of a visit to the place in question can determine the expected duration of parking, this context information can be used to optimize the best parking locations for drivers.
- For the parked vehicles, the expected duration of parking can be uploaded to the traffic cloud and shared with potential drivers after analysis.
- In this way, even when the parking garage has no empty parking spaces available, drivers still can inquire as to the status of the parking garage and get the desired service by context-aware optimization.
- The proposed context-aware dynamic parking service is a promising solution for alleviating parking difficulties and improving the QoS of CVC. Many technologies such as WSNs, traffic clouds, and traffic data mining are enabling this application scenario to become a reality

Summary

- Mobile cloud computing is one of the mobile technology trends in the future because it combines the advantages of both MC and CC, thereby providing optimal services for mobile users
- MCC focuses more on user experience : Lower battery consumption , Faster application execution
- MCC architectures design the middleware to partition an application execution transparently between mobile device and cloud servers
- The applications supported by MCC including m-commerce, mlearning, and mobile healthcare show the applicability of the MCC to a wide range.
- The issues and challenges for MCC (i.e., from communication and computing sides) demonstrates future research avenues and directions.

References

- Dinh, Hoang T., et al. "A survey of mobile cloud computing: architecture, applications, and approaches." *Wireless communications and mobile computing* 13.18 (2013): 1587-1611
- Z. Li, C. Wang, and R. Xu, "Computation offloading to save energy on handheld devices: a partition scheme," in *Proc 2001 Intl 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
- H. H. La 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
- Gordon, Mark S., et al. "COMET: Code Offload by Migrating Execution Transparently." *OSDI*. 2012.
- Yang, Seungjun, et al. "Fast dynamic execution offloading for efficient mobile cloud computing." *Pervasive Computing and Communications (PerCom), 2013 IEEE International Conference on*. IEEE, 2013
- Shiraz, Muhammad, et al. "A review on distributed application processing frameworks in smart mobile devices for mobile cloud computing." *Communications Surveys & Tutorials, IEEE* 15.3 (2013): 1294-1313
- <https://www.ibm.com/cloud-computing/learn-more/what-is-mobile-cloud-computing/>

Thank You!!

NPTEL



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

CLOUD COMPUTING

Mobile Cloud Computing - II

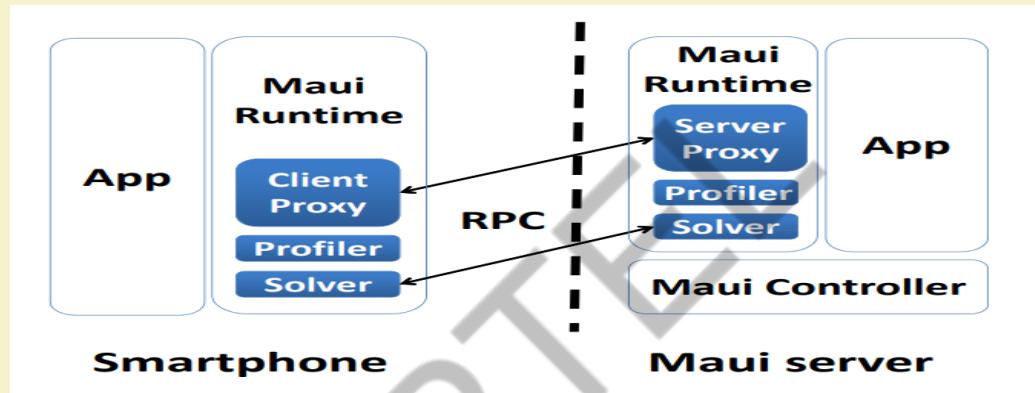
Prof. Soumya K Ghosh

Department of Computer Science and Engineering
IIT KHARAGPUR

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

MCC Systems: MAUI (Mobile Assistance Using Infrastructure)

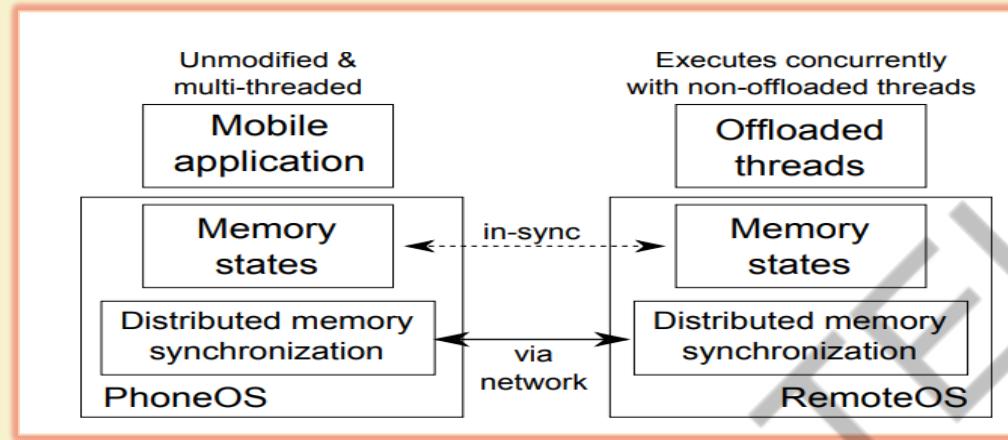


- **MAUI enables the programmer to produce an initial partition of the program**
 - Programmer marks each method as “remoteable” • or not
 - Native methods cannot be remoteable
- MAUI framework uses the annotation to decide

whether a method should be executed on cloud server to save energy and time to execute

MAUI server is the cloud component. The framework has the necessary software modules required in the workflow.

MCC Systems: COMET



- Requires only program binaries Execute multi-threaded programs correctly Improve speed of computation
- Further improvements to data traffic during migration is also possible by sending only the parts of the heap that has been modified

COMET: Code Offload by Migrating Execution Transparently

- Works on unmodified applications (no source code required)
- Allows threads to migrate between machines depending on workload
- It implements a Distributed Shared Memory (DSM) model for the runtime engine
 - ✓ *DSM allows transparent movement of threads across machines*
 - ✓ *In computer architecture, DSM is a form of memory architecture where the (physically separate) memories can be addressed as one (logically shared) address space*

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
NPTEL ONLINE
CERTIFICATION COURSES

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.

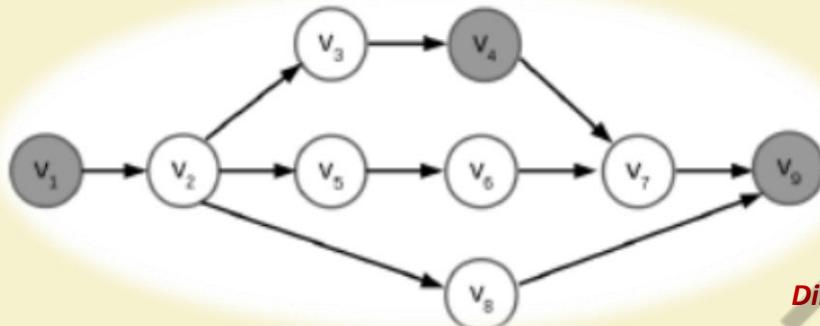


IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

Mathematical Formulation



Directed Acyclic Graph represents an application Call Graph

$$\text{maximize } \sum_{v \in V} I_v \times E_v^l - \sum_{(u,v) \in E} |I_u - I_v| \times C_{u,v}$$

$$\text{such that: } \sum_{v \in V} ((1 - I_v) \times T_v^l) + (I_v \times T_v^r)$$

$$+ \sum_{(u,v) \in E} (|I_u - I_v| \times B_{u,v}) \leq L$$

$$\text{and } I_v \leq r_v, \forall v \in V$$

- Highlighted nodes must be executed on the mobile device -> called native tasks (v_1, v_4, v_9)
- Edges represent the sequence of execution - Any non-highlighted node can be executed either locally on the mobile device or on cloud

- 0-1 integer linear program,
where $I_v = 0$ if method executed locally,
 $= 1$ if method executed remotely
- E : Energy cost to execute method v locally
- $C_{u,v}$: Cost of data transfer
- L : Total execution latency
- T : Time to execute the method
- B : Time to transfer program state

Integer Linear Program to solve the Task Partitioning Problem

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

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
NPTEL ONLINE
CERTIFICATION COURSES

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.

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
NPTEL ONLINE
CERTIFICATION COURSES

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

$$\frac{C}{M} \times (P_c - \frac{P_i}{F}) - P_{tr} \times \frac{D}{B}$$

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.

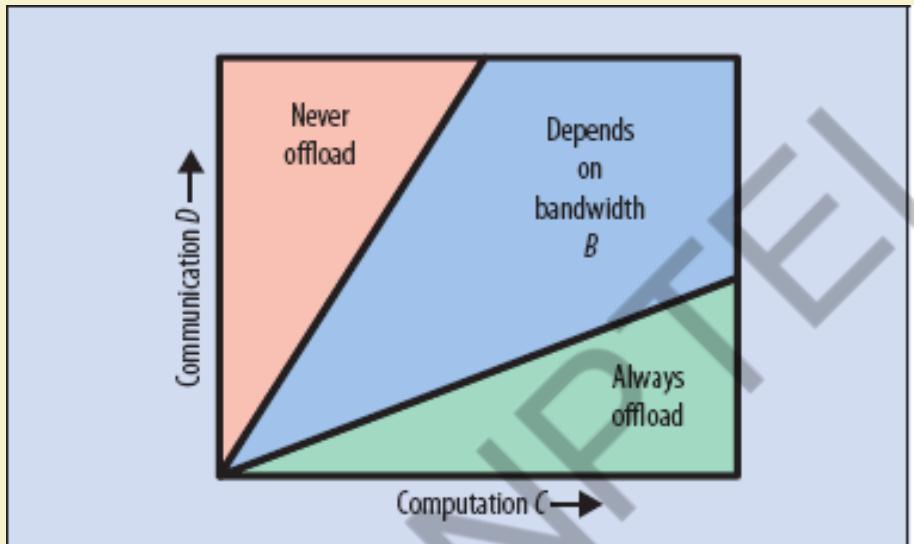
The amount of energy saved is :

$$P_c \times \frac{C}{M} - P_i \times \frac{C}{S} - P_{tr} \times \frac{D}{B}$$

We can rewrite the formula as

$$\frac{C}{M} \times (P_c - \frac{P_i}{F}) - P_{tr} \times \frac{D}{B}$$

When to Offload?? (contd..)



Offloading is beneficial when large amounts of computation C are needed with relatively small amounts of communication D

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

$$\frac{C}{M} \times (P_c - \frac{P_i}{F}) - P_{tr} \times \frac{D}{B}$$

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 Intl 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

How to evaluate MCC performance

- Energy Consumption
 - Must reduce energy usage and extend battery life
- Time to Completion
 - Should not take longer to finish the application compared to local execution
- Monetary Cost
 - Cost of network usage and server usage must be optimized
- Security
 - As offloading transfers data to the servers, ensure confidentiality and privacy of data, how to identify methods which process confidential data

Challenges

- How can one design a practical and usable MCC framework
 - System as well as partitioning algorithm
- Is there a scalable algorithm for partitioning
 - Optimization formulations are NP-hard
 - Heuristics fail to give any performance guarantee
- Which are the most relevant parameters to consider in the design of MCC systems?

Mobile Cloud Computing – Applications?

Mobile Health-care



Health-Monitoring services, Intelligent emergency management system, Health-aware mobile devices (detect pulse rate, blood pressure, alcohol-level etc.)

Mobile Gaming



It can completely offload game engine requiring large computing resource (e.g., graphic rendering) to the server in the cloud



Mobile Commerce

M-commerce allows business models for commerce using mobile (Mobile financial, mobile advertising, mobile shopping)



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Mobile Cloud Computing – Applications?



Pedestrian crossing guide for blind and visually-impaired

Mobile currency reader for blind and visually impaired

Lecture transcription for hearing impaired students

Assistive Technologies



Mobile Learning

- *M-learning combines e-learning and mobility*
- *Traditional m-learning has limitations on high cost of devices/network, low transmission rate, limited educational resources*
- *Cloud-based m-learning can solve these limitations*
- *Enhanced communication quality between students and teachers*
- *Help learners access remote learning resources*
- *A natural environment for collaborative learning*



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

MuSIC: Mobility-Aware Optimal Service Allocation in Mobile Cloud Computing

- User Mobility introduces new complexities in enabling an optimal decomposition of tasks that can execute cooperatively on mobile clients and the tiered cloud architecture while considering multiple QoS goals such application delay, device power consumption and user cost/price.
- Apart from scalability and access issues with the increased number of users, mobile applications are faced with increased *latencies* and reduced *reliability*
- As a user moves, the physical distance between the user and the cloud resources originally provisioned changes causing additional delays
- Further, the lack of effective handoff mechanisms in WiFi networks as user move rapidly causes an increase in the number of *packet losses*

In other words, user mobility, if not addressed properly, can result in suboptimal resource mapping choices and ultimately in diminished application QoS

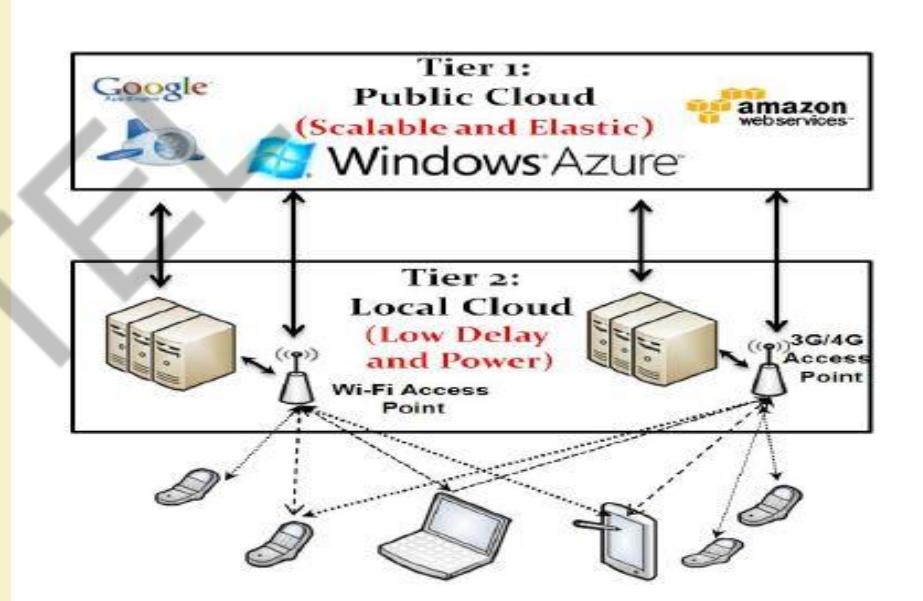
MuSIC: Mobility-Aware Optimal Service Allocation in Mobile Cloud Computing

Efficient techniques for *dynamic mapping of resources* in the presence of **mobility**; using a *tiered cloud architecture*, to meet the *multidimensional QoS* needs of mobile users

- Location-time workflow (LTW) as the modeling framework to model mobile applications and capture user mobility. Within this framework, mobile service usage patterns as a function of location and time has been formally modelled
- Given a mobile application execution expressed as a LTW, the framework optimally partitions the execution of the location-time workflow in the 2-tier architecture based on a *utility metric* that combines *service price, power consumption and delay* of the mobile applications

MuSIC: Mobility-Aware Optimal Service Allocation in Mobile Cloud Computing

- ✓ Tier 1 nodes in the system architecture represents *public cloud services* such as Amazon EC2, Microsoft Azure and Google AppEngine. Services provided by these vendors are highly *scalable* and *available*; what they lack is the ability to provide the *fine grain location granularity* required for high performance mobile applications
- ✓ This feature is provided by the second tier local cloud, that consists of nodes that are connected to access points.
- ✓ Location information of these services are available at finer levels of granularity (campus and street level).
- ✓ Mobile users are typically connected to these local clouds through WiFi (via access points) or cellular (via 3G cell towers) connectivity - the aim to intelligently select which local and which public cloud resources to utilize for task offloading.



2-Tier Mobile Cloud Architecture

Mobile Application Modelling

Cloud Service Set:

The set of all services (e.g. compute, storage and software capabilities like multimedia streaming services, content transcoding services, etc) provided by local and public cloud providers

Local Cloud Capacity:

Local cloud services can only accept a limited number of mobile client requests

Location Map:

It is a partition of the 2-D space/region in which mobile hosts and cloud resources are located

User Service Set:

The set of all services that a user has on his own device (e.g. decoders, image editors etc.)

Criteria	Definition
$q_{price}(s_i, u_k^{l_i, t_j})$	The price of using service s_i when user u_k is in location $l_i \in L$ and time t_j .
$q_{power}(s_i, u_k^{l_i, t_j})$	The power consumed on user mobile device using s_i when user u_k is in location $l_i \in L$ and time t_j .
$q_{delay}(s_i, u_k^{l_i, t_j})$	The delay of executing service s_i when user u_k is in location $l_i \in L$ and time t_j .

Mobile User Trajectory:

The trajectory of a mobile user, u_k , is represented as a list of tuples of the form $\{(1; l_1); \dots; (n; l_n)\}$ where $(i; l_i)$ implies that the mobile user is in location l_i for time duration i

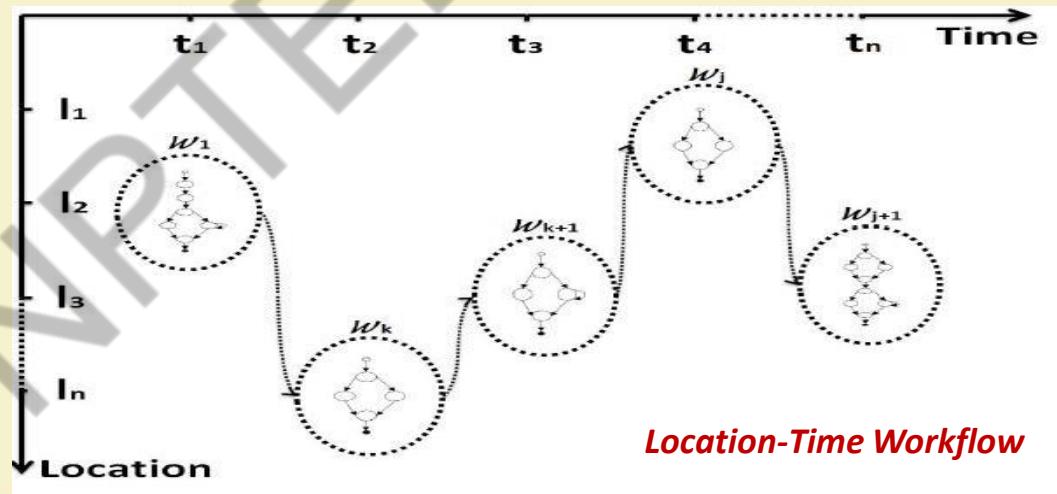
Center of Mobility:

It is the location where (or near where) a mobile user u_k spends most of its time

Mobile Application Modelling

Location-Time Workflow

Combination of the mobile application workflow concept with a user trajectory to model the mobile users and the requested services in their trajectory.



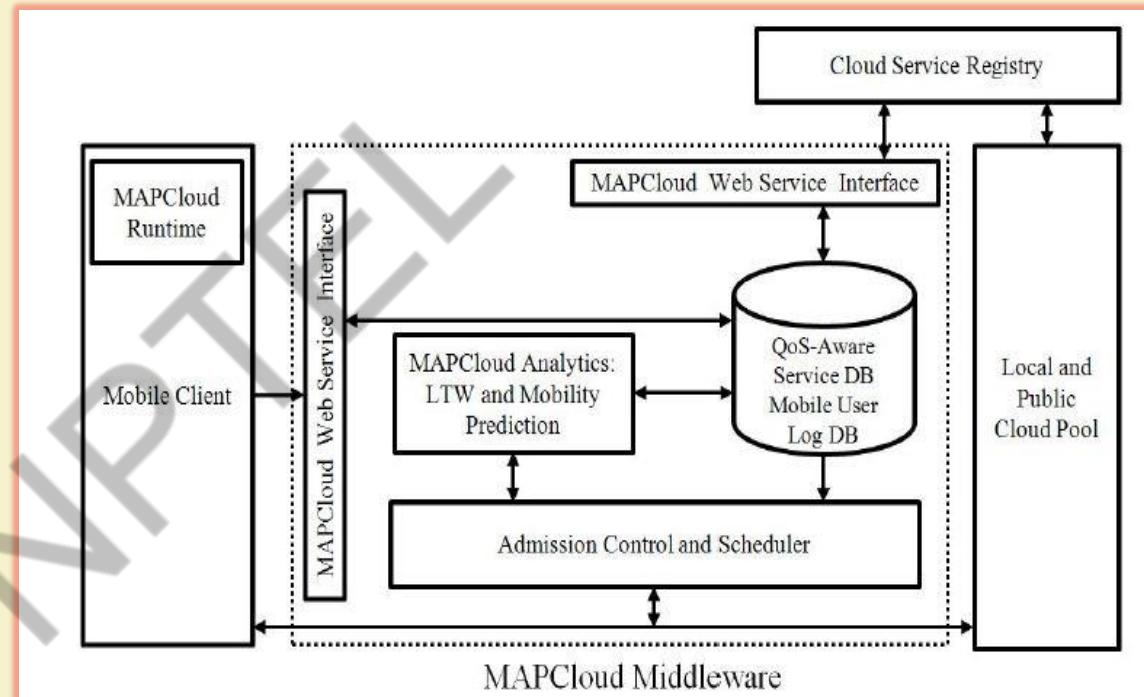
Mobile Application Modelling

Mobile User Log DB and QoS-Aware Service DB:

Unprocessed user data log such as mobile service usage, location of the user, user delay experience of getting the service, energy consumed on user mobile device, etc and service lists on local and public cloud and their QoSes in different locations respectively

MAPCloud Analytic: This module processes mobile user Log DB and updates QoS-aware cloud service DB based on user experience and LTW

Admission Control and Scheduling: This module is responsible for optimally allocate services to admitted mobile users based on MuSIC



A Case Study: Context Aware Dynamic Parking Service

- MCC can provide a flexible method of handling massive computing, storage, and software services in a scalable and virtualized manner.
- The integration of MCC and vehicular networks is expected to promote the development of cost effective, scalable, and data-driven CVC (Context-aware vehicular cyber physical systems)

An application scenario regarding the context-aware dynamic parking services by illuminating the cloud-assisted architecture and logic flow.

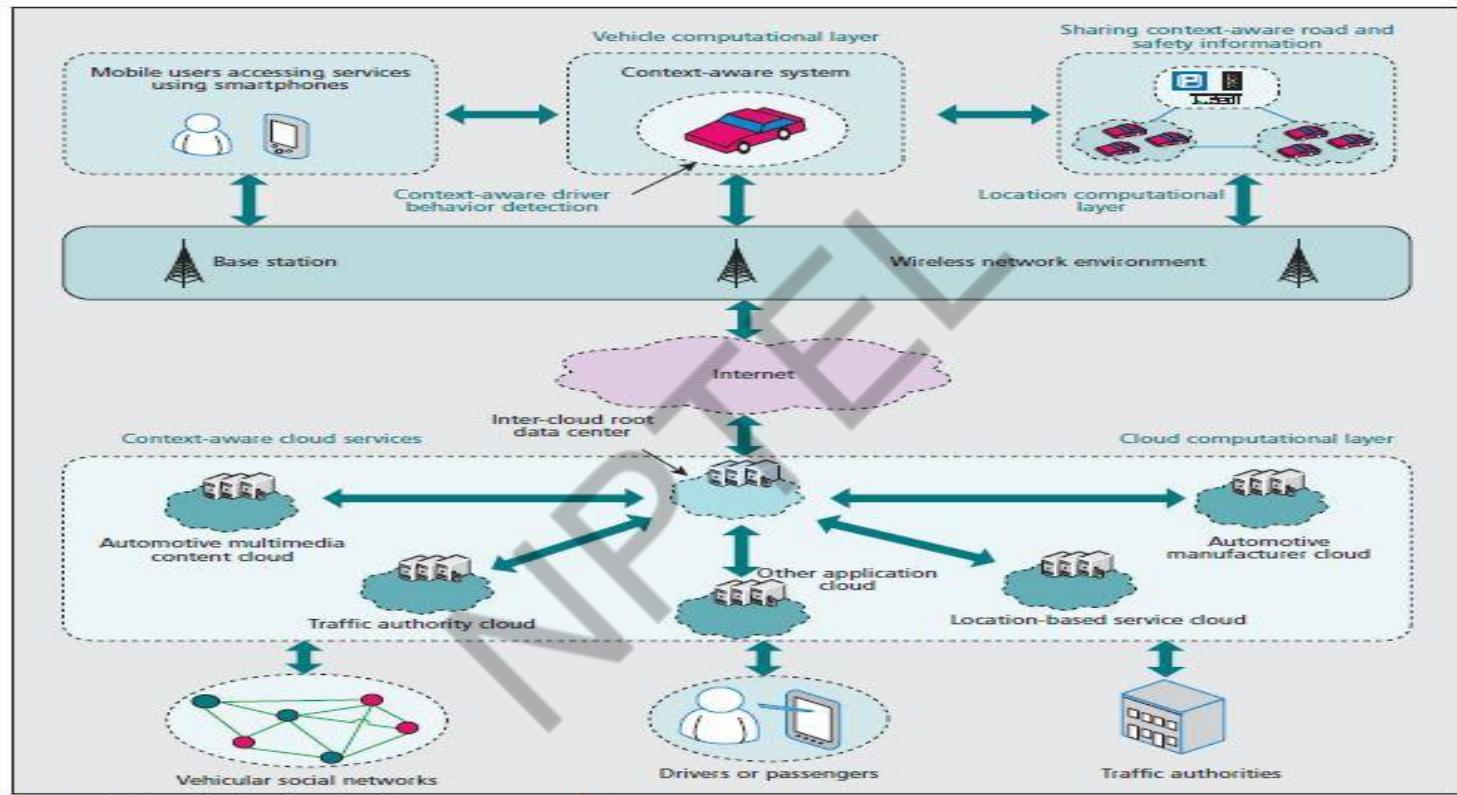
- As the number of vehicles increases, there is an increasing trend of insufficient parking spaces in many large cities, and this problem is gradually getting worse
- With the proliferation of wireless sensor networks (WSNs) and cloud computing, there exists strong potential to alleviate this problem using context information (e.g., road conditions and status of parking garages) to provide context-aware dynamic parking services
- Cloud Assisted parking services (traditional parking garages and dynamic parking services along the road) and parking reservation service using smart terminals such as smartphones.

A Case Study: Context Aware Dynamic Parking Service

- MCC can provide a flexible method of handling massive computing, storage, and software services in a scalable and virtualized manner.
- The integration of MCC and vehicular networks is expected to promote the development of cost effective, scalable, and data-driven CVC (Context-aware vehicular cyber physical systems)

An application scenario regarding the context-aware dynamic parking services by illuminating the cloud-assisted architecture and logic flow.

- As the number of vehicles increases, there is an increasing trend of insufficient parking spaces in many large cities, and this problem is gradually getting worse
- With the proliferation of wireless sensor networks (WSNs) and cloud computing, there exists strong potential to alleviate this problem using context information (e.g., road conditions and status of parking garages) to provide context-aware dynamic parking services
- Cloud Assisted parking services (traditional parking garages and dynamic parking services along the road) and parking reservation service using smart terminals such as smartphones.



IIT KHARAGPUR

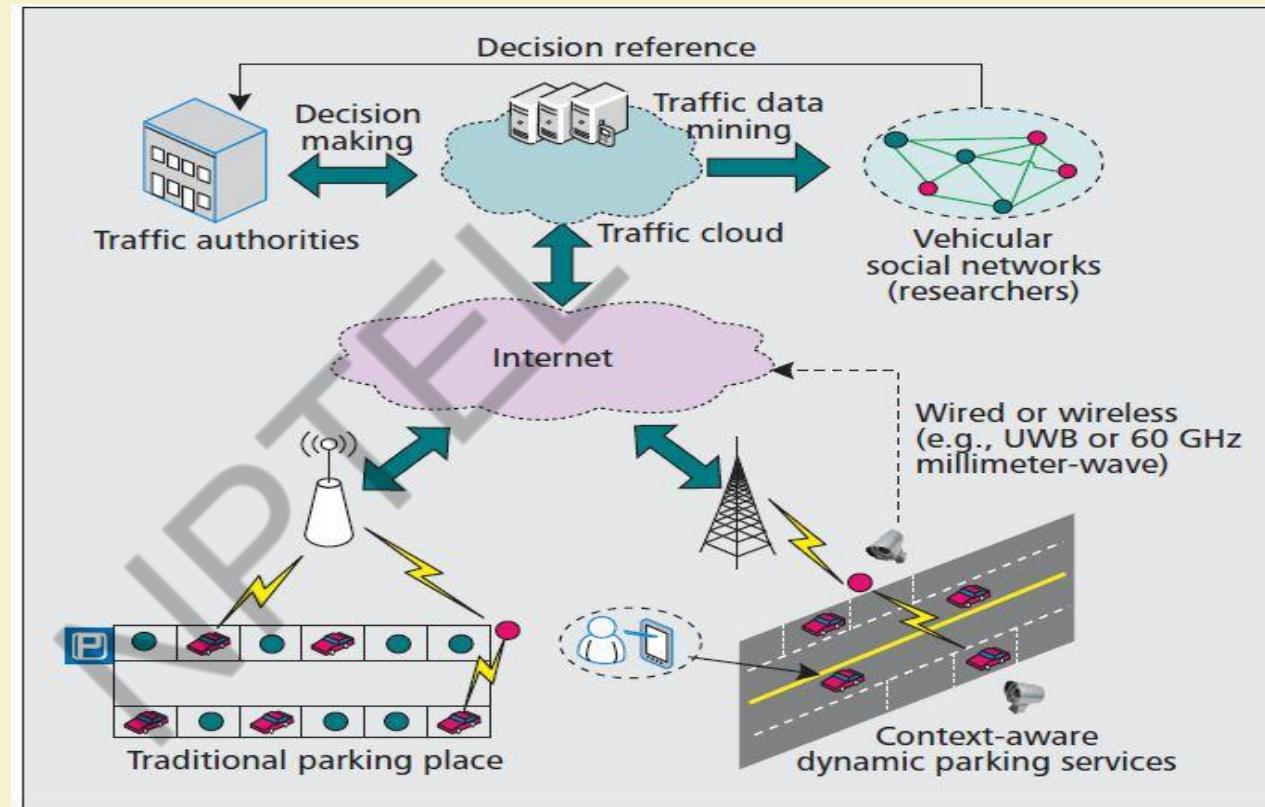


NPTEL
ONLINE
CERTIFICATION COURSES

Example cloud-assisted context-aware architecture

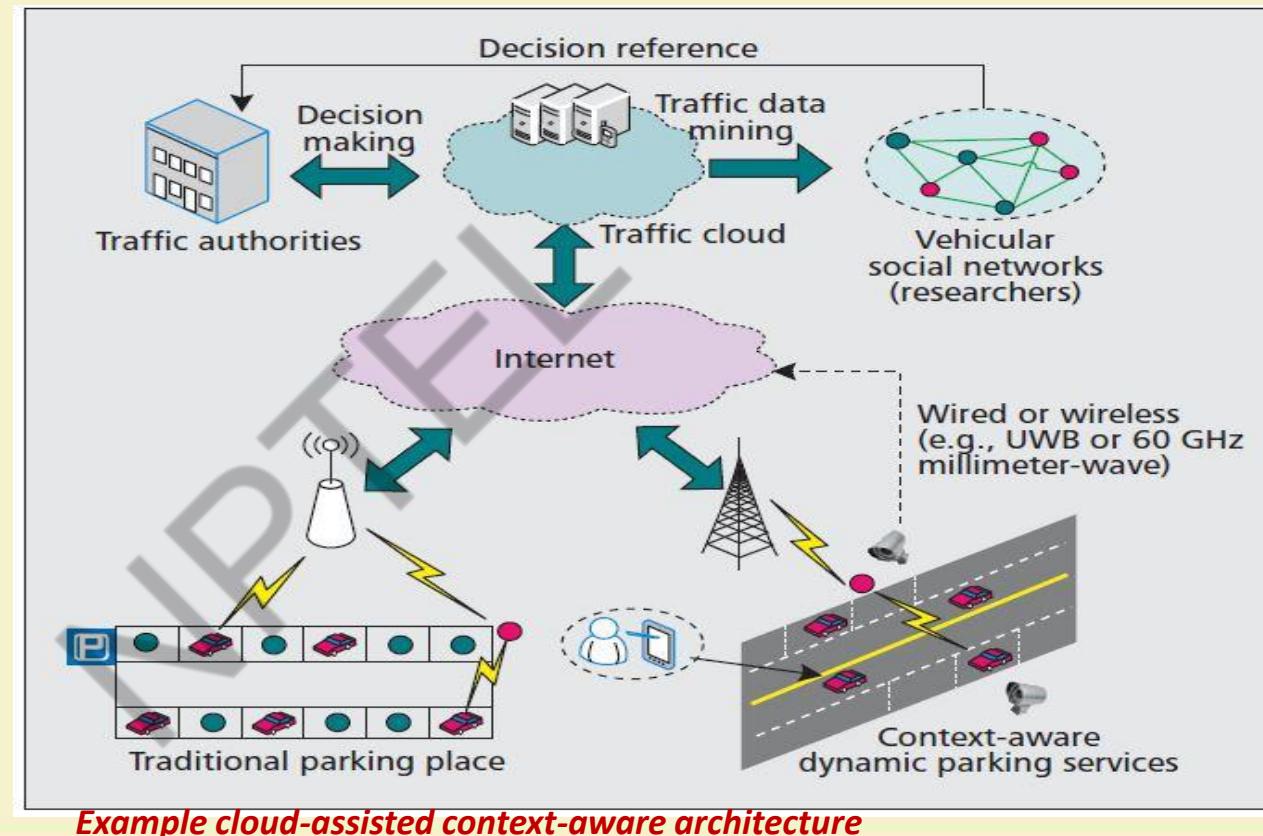
Traditional parking garages:

- The context information of each parking space detected by a WSN is forwarded to the traffic cloud by WSNs, third-generation (3G) communications, and the Internet.
- The collected data are processed in the cloud and then selectively transmitted to the users.
- This is helpful for providing more convenience services and evaluating the utilization levels of the parking garage.
- Also, the status of the parking garage may be dynamically published on a nearby billboard to users who have no ability to get the status by smart terminals.



Dynamic parking services:

- In this scenario, we consider a situation in which we may temporarily park a vehicle along the road if it does not impede the passage of other vehicles or pedestrians.
- We envision this application scenario based on the common observation that the traffic flow capacity is usually regular for each road. For example, there is usually heavy traffic during morning and evening rush hours.
- Therefore, considering the context information such as rush hours and road conditions, we may dynamically arrange the parking services for a very wide road.
- With the support of many new technologies (e.g., MCC and WSNs), the traffic authorities can carry out the dynamic management of this kind of service.



Example cloud-assisted context-aware architecture

A Case Study: Context Aware Dynamic Parking Service

Three aspects, including service planning of traffic authorities, reservation service process, and context-aware optimization have been studied.

Decision making of traffic authorities

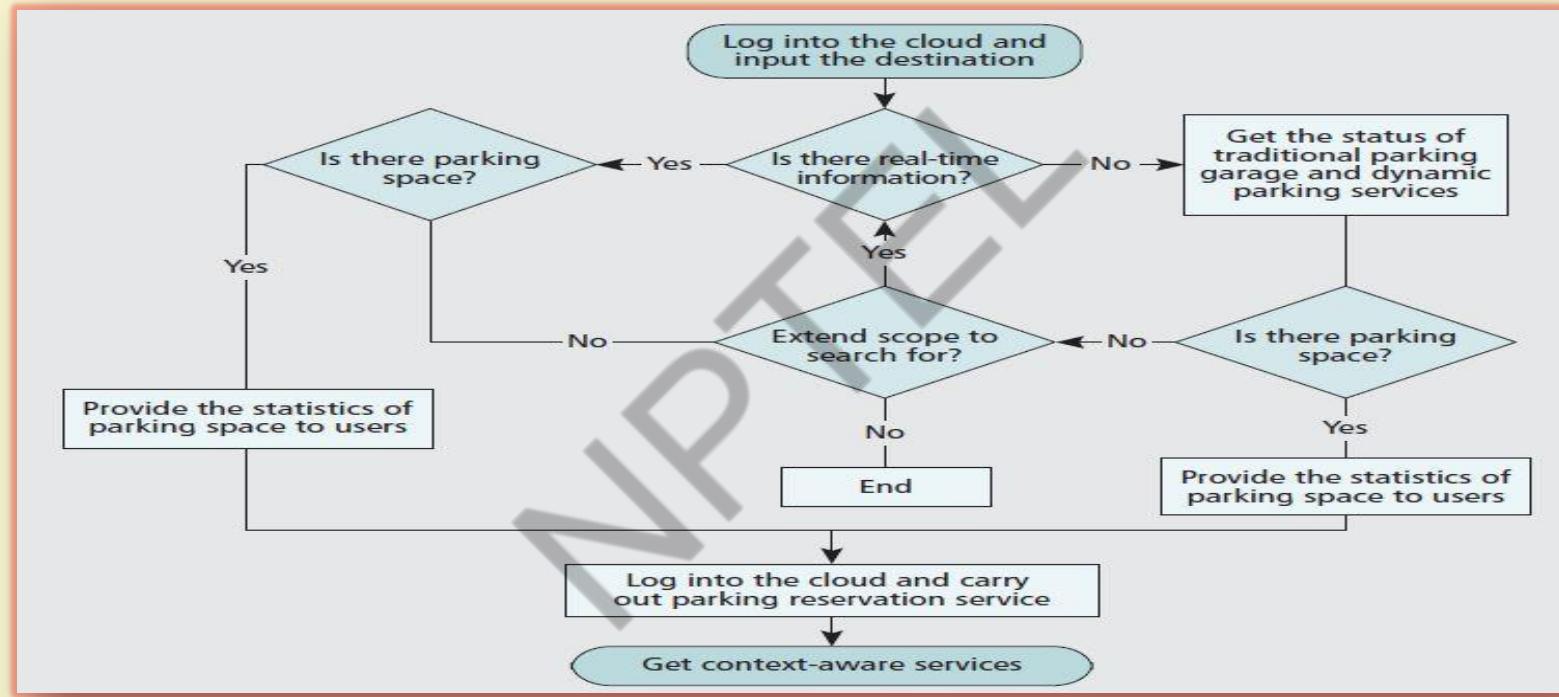
- The decision-making process of the proposed scheme heavily depends on many factors, such as historical traffic flow capacity, road conditions, weather conditions, and traffic flow forecasting
- In order to make an effective prediction, researchers on vehicular social networks carry out traffic data mining to discover useful information and knowledge from collected big data. The prediction process depends on classifying the influence factors and designing a decision tree
- By the method of probability analysis, the traffic authorities dynamically arrange whether the road can be authorized to provide context-aware parking services. In some particular cases, a fatal factor may directly affect the decision making. For example, when a typhoon is approaching, traffic authorities may immediately terminate services

A Case Study: Context Aware Dynamic Parking Service

Parking reservation services:

- The status of a parking space can be monitored as determined by the corresponding system, and subsequently updated in the traffic cloud.
- The drivers or passengers can quickly obtain the parking space's information by various smart terminals such as smartphones. If a proper parking space cannot be found, further search scope is extended.
- Within a given time, we may log into the traffic cloud and subscribe to a parking space.

A Case Study: Context Aware Dynamic Parking Service



A Case Study: Context Aware Dynamic Parking Service

Context-aware optimization:

- The context information includes not only road conditions and the status of the parking garage, but also the expected duration of parking as well.
- Since the purpose of a visit to the place in question can determine the expected duration of parking, this context information can be used to optimize the best parking locations for drivers.
- For the parked vehicles, the expected duration of parking can be uploaded to the traffic cloud and shared with potential drivers after analysis.
- In this way, even when the parking garage has no empty parking spaces available, drivers still can inquire as to the status of the parking garage and get the desired service by context-aware optimization.
- The proposed context-aware dynamic parking service is a promising solution for alleviating parking difficulties and improving the QoS of CVC. Many technologies such as WSNs, traffic clouds, and traffic data mining are enabling this application scenario to become a reality

Summary

- Mobile cloud computing is one of the mobile technology trends in the future because it combines the advantages of both MC and CC, thereby providing optimal services for mobile users
- MCC focuses more on user experience : Lower battery consumption , Faster application execution
- MCC architectures design the middleware to partition an application execution transparently between mobile device and cloud servers
- The applications supported by MCC including m-commerce, mlearning, and mobile healthcare show the applicability of the MCC to a wide range.
- The issues and challenges for MCC (i.e., from communication and computing sides) demonstrates future research avenues and directions.

References

- Dinh, Hoang T., et al. "A survey of mobile cloud computing: architecture, applications, and approaches." *Wireless communications and mobile computing* 13.18 (2013): 1587-1611
- Z. Li, C. Wang, and R. Xu, "Computation offloading to save energy on handheld devices: a partition scheme," in *Proc 2001 Intl 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
- H. H. La 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
- Gordon, Mark S., et al. "COMET: Code Offload by Migrating Execution Transparently." *OSDI*. 2012.
- Yang, Seungjun, et al. "Fast dynamic execution offloading for efficient mobile cloud computing." *Pervasive Computing and Communications (PerCom), 2013 IEEE International Conference on*. IEEE, 2013
- Shiraz, Muhammad, et al. "A review on distributed application processing frameworks in smart mobile devices for mobile cloud computing." *Communications Surveys & Tutorials, IEEE* 15.3 (2013): 1294-1313
- <https://www.ibm.com/cloud-computing/learn-more/what-is-mobile-cloud-computing/>

Thank You!!

NPTEL



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

CLOUD COMPUTING

Fog Computing - I

PROF. SOUMYA K. GHOSH

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

IIT KHARAGPUR

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
NPTEL ONLINE
CERTIFICATION COURSES

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

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
NPTEL ONLINE
CERTIFICATION COURSES

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



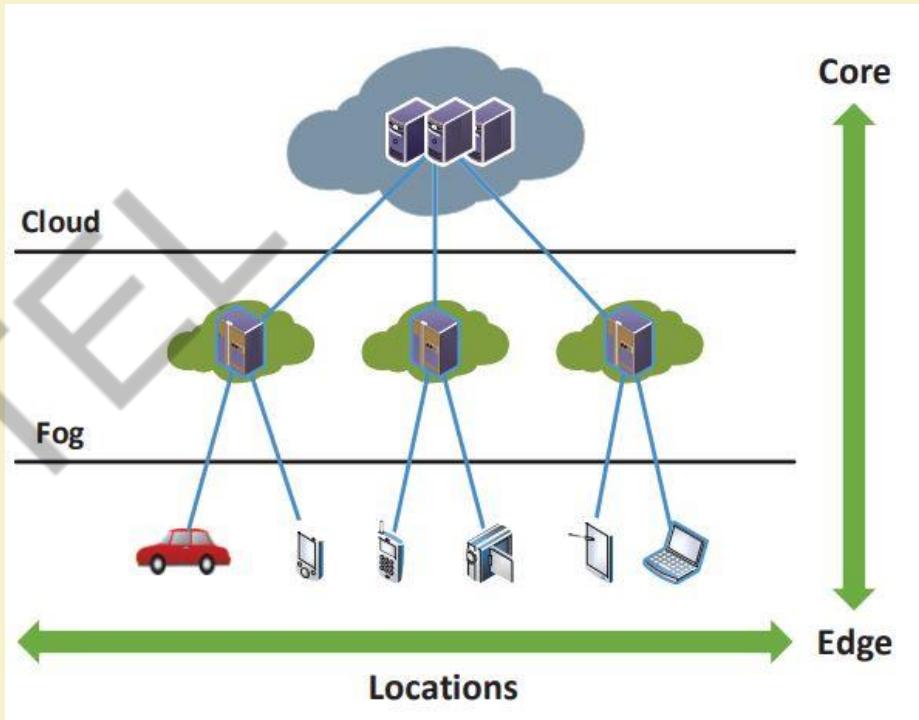
IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Fog Computing

- Bringing intelligence down from the cloud close to the ground/ end-user.
- Cellular base stations, Network routers, WiFi Gateways will be capable of running applications.
- End devices, like sensors, are able to perform basic data processing.
- Processing close to devices lowers response time, enabling real-time applications.



Source: *The Fog Computing Paradigm: Scenarios and Security Issues*,
Ivan Stojmenovic and Sheng Wen



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

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

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

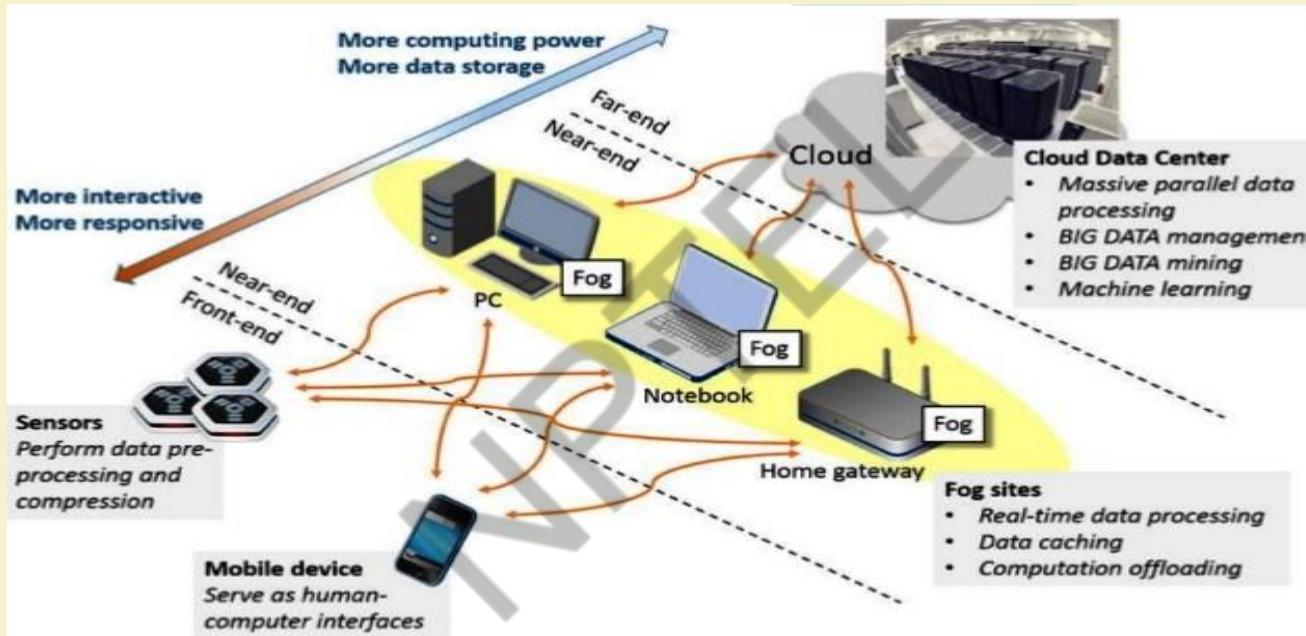


IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Fog Computing



Source: Internet



IIT KHARAGPUR



NPTEL
ONLINE
CERTIFICATION COURSES

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
NPTEL ONLINE
CERTIFICATION COURSES

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

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

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

Limitations of Cloud Computing

- High capacity(bandwidth) requirement
- Client access link
- High latency
- Security

“Fog” Solution?

- Reduction in data movement across the network resulting in reduced congestion
- Elimination of bottlenecks resulting from centralized computing systems
- Improved security of encrypted data as it stays closer to the end user



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

Fog Computing and Cloud Computing

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 enrouter	High probability	Very Less probability
Location awareness	No	Yes

Source: Internet



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

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

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

Applicability

- Smart Grids
- Smart Traffic Lights
- Wireless Sensors
- Internet of Things
- Software Defined Network

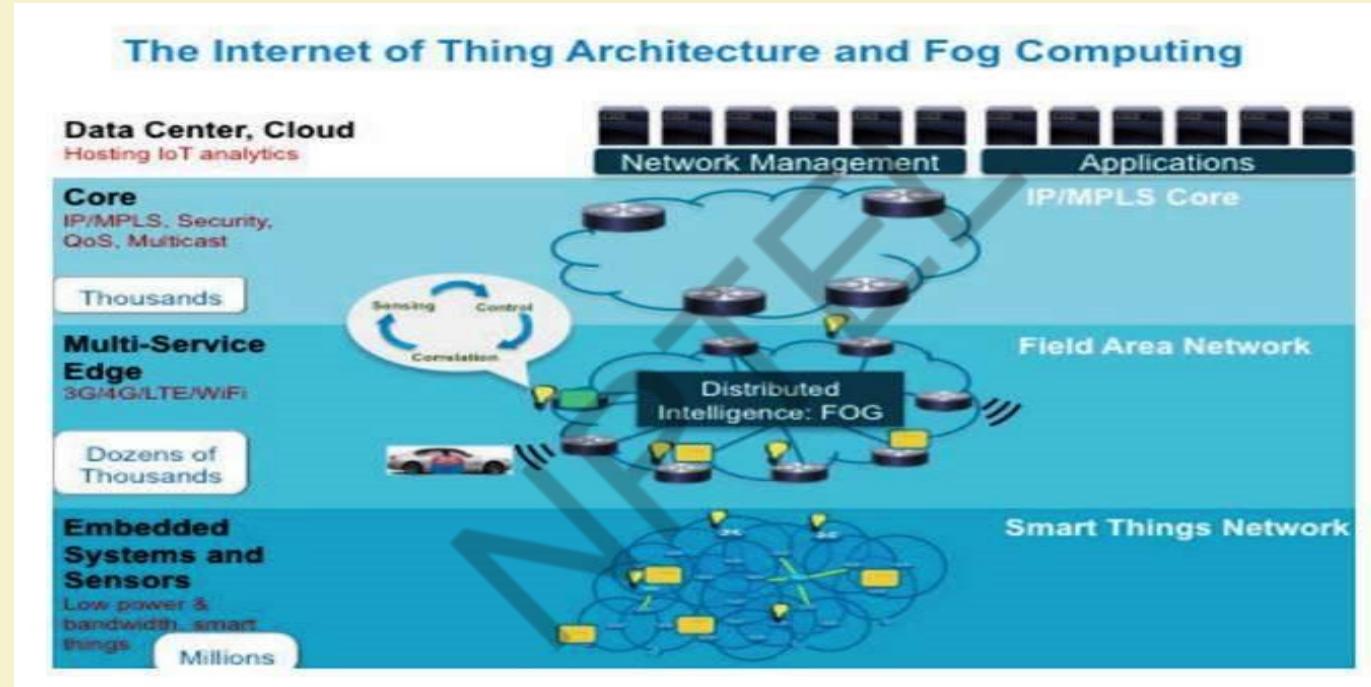


IIT KHARAGPUR



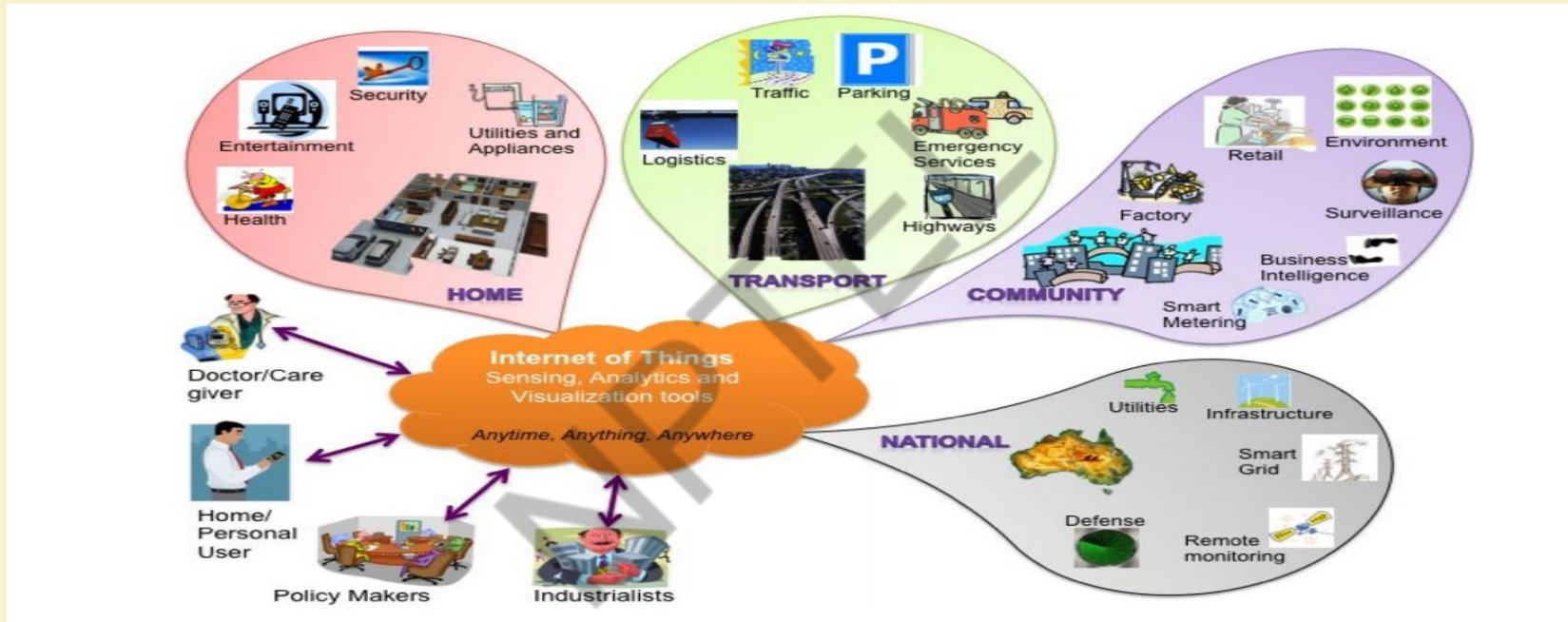
NPTEL ONLINE
CERTIFICATION COURSES

Fog Computing and IoT (Internet of Things)



Source: *Fog Computing and Its Role in the Internet of Things*, Flavio Bonomi, Rodolfo Milito, Jiang Zhu, Sateesh Addepalli

Internet of Things



Source: Internet of Things (IoT): A vision, architectural elements, and future directions, Jayavardhana Gubbi, Rajkumar Buyya, Slaven Marusic, Marimuthu Palaniswami



IIT KHARAGPUR



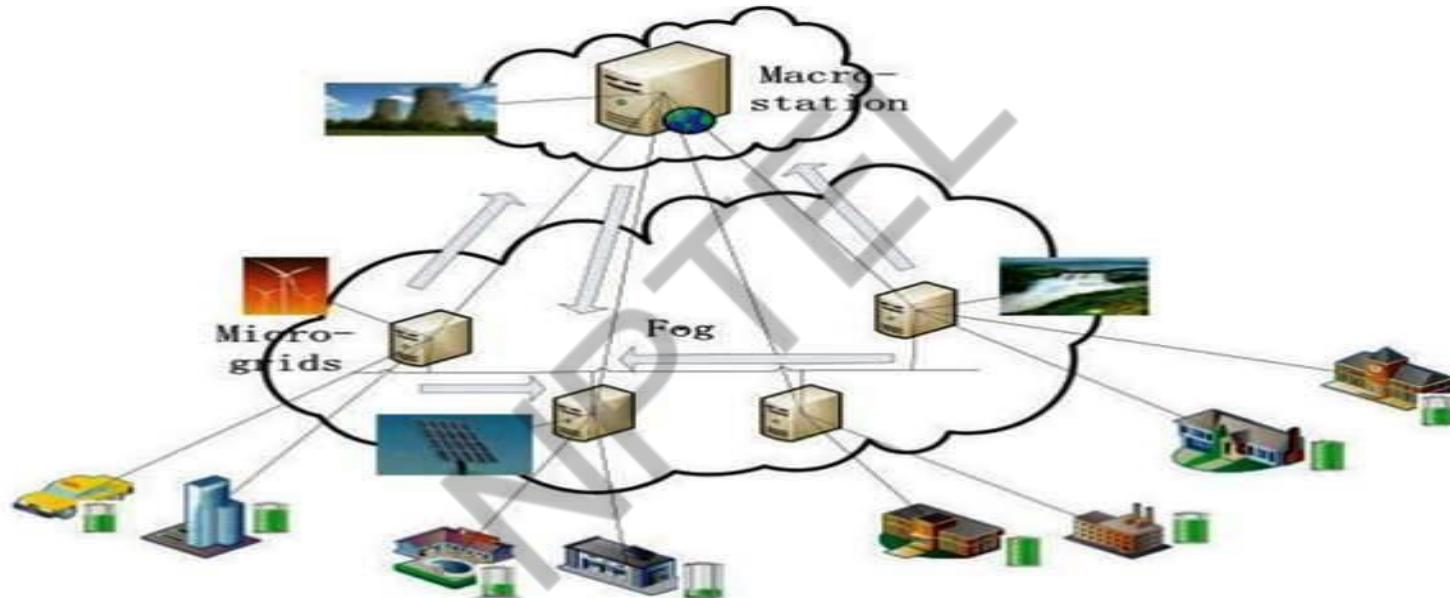
NPTEL ONLINE
CERTIFICATION COURSES

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. The Fog has a number of attributes that make it the ideal platform to deliver a rich menu of SCV services in 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 Milito, Jiang Zhu, Sateesh Addepalli

Smart Grid and Fog Computing



Source: Source: *The Fog Computing Paradigm: Scenarios and Security Issues*, Ivan Stojmenovic and Sheng Wen

Fog computing in Smart Traffic Lights and Connected Vehicles



Source: Source: The Fog Computing Paradigm: Scenarios and Security Issues, Ivan Stojmenovic and Sheng Wen



IIT KHARAGPUR



NPTEL
ONLINE
CERTIFICATION COURSES

Thank You!!

NPTEL



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

CLOUD COMPUTING

Fog Computing - II

PROF. SOUMYA K. GHOSH

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
IIT KHARAGPUR

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.

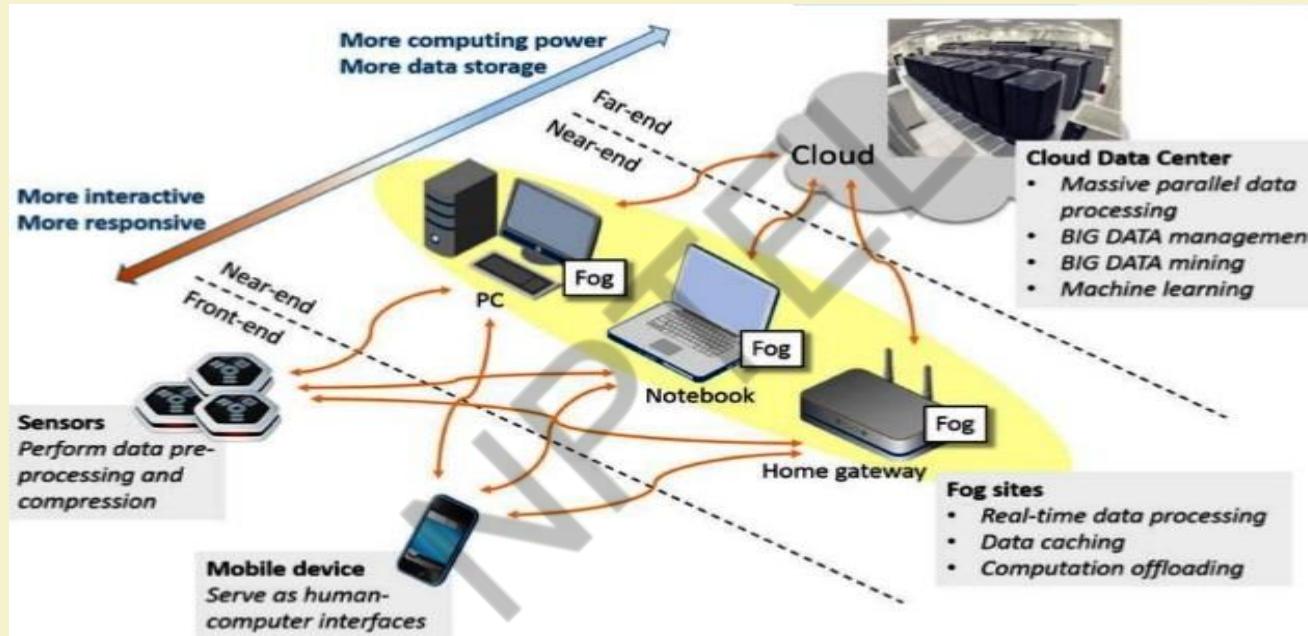


IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

Fog Computing



Source: Internet



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

Fog Computing and Cloud Computing

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 enrouter	High probability	Very Less probability
Location awareness	No	Yes

Source: Internet



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

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

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
NPTEL ONLINE
CERTIFICATION COURSES

Applicability

- Smart Traffic Lights
- Connected Vehicles
- Smart Grids
- Wireless Sensors
- Internet of Things
- Software Defined Network



IIT KHARAGPUR



NPTEL
ONLINE
CERTIFICATION COURSES

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 Milito, Jiang Zhu, Sateesh Addepalli

Fog Computing in Smart Traffic Lights and Connected Vehicles



Source: Source: The Fog Computing Paradigm: Scenarios and Security Issues, Ivan Stojmenovic and Sheng Wen

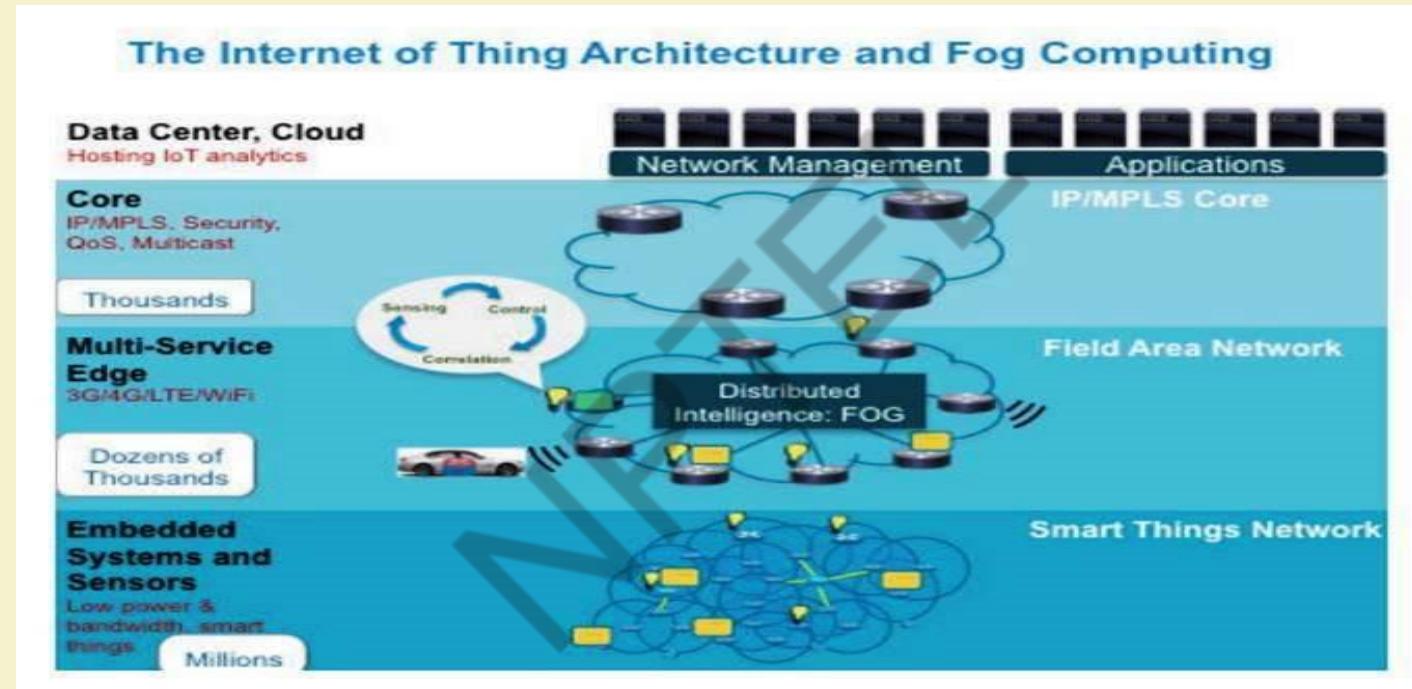


IIT KHARAGPUR



NPTEL
ONLINE
CERTIFICATION COURSES

Fog Computing and IoT (Internet of Things)



Source: *Fog Computing and Its Role in the Internet of Things*, Flavio Bonomi, Rodolfo Milito, Jiang Zhu, Sateesh Addepalli

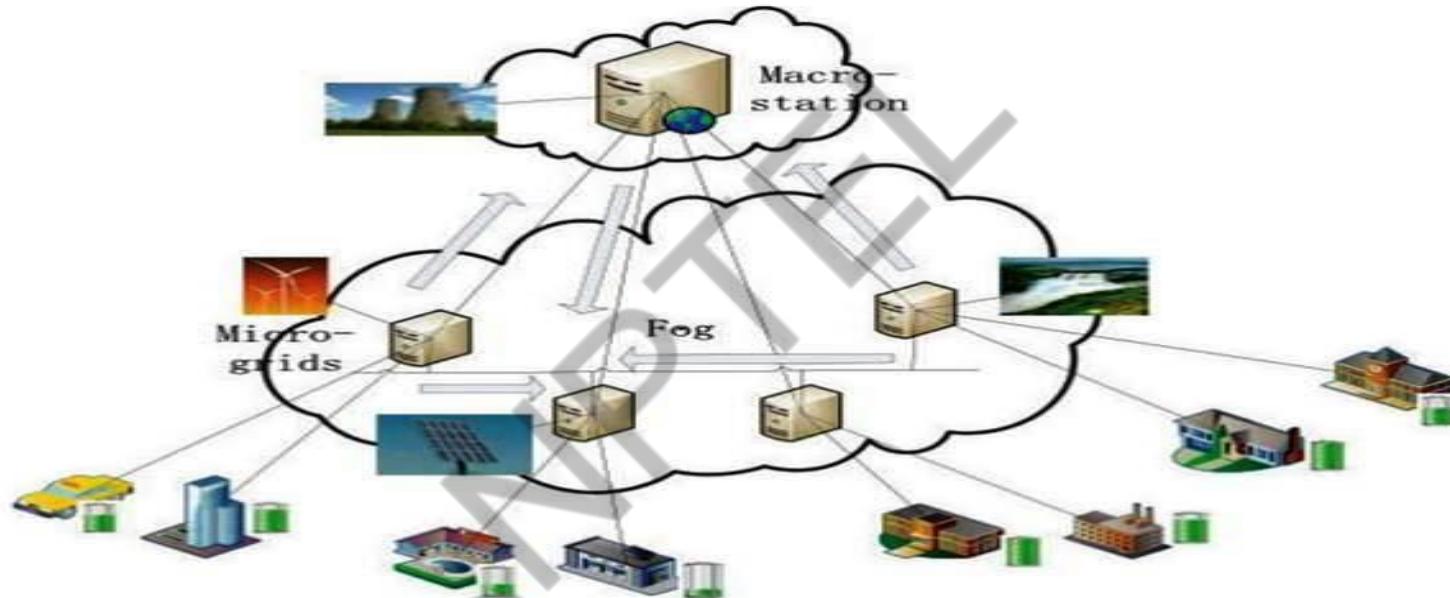


IIT KHARAGPUR



NPTEL
ONLINE
CERTIFICATION COURSES

Fog Computing and Smart Grid



Source: Source: *The Fog Computing Paradigm: Scenarios and Security Issues*, Ivan Stojmenovic and Sheng Wen

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

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

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
NPTEL ONLINE
CERTIFICATION COURSES

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 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

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
NPTEL ONLINE
CERTIFICATION COURSES

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

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

Thank You!!

NPTEL



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

Cloud Computing

Use Case: Geospatial Cloud

Soumya K Ghosh

Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur
skg@cse.iitkgp.ernet.in

Broad Agenda

- ▶ Geospatial Information
- ▶ Geospatial Cloud
- ▶ IIT Kharagpur Geo-Cloud

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



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

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)



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)

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).

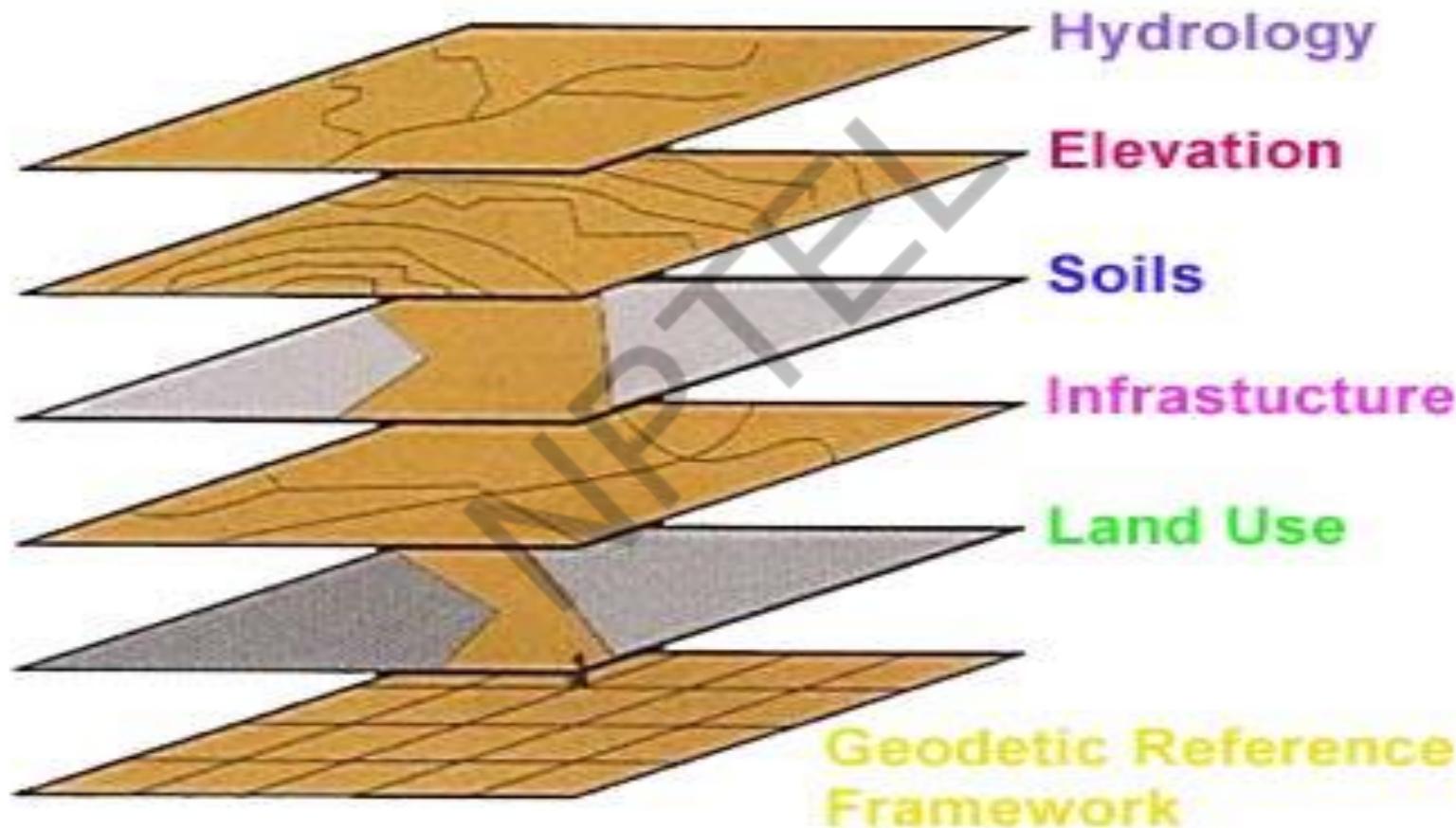
Components of a GIS

- ▶ Computer hardware
- ▶ Software
- ▶ Data management and analysis procedures (this could be considered part of the software)
- ▶ Spatial data
- ▶ People needed to operate the GIS

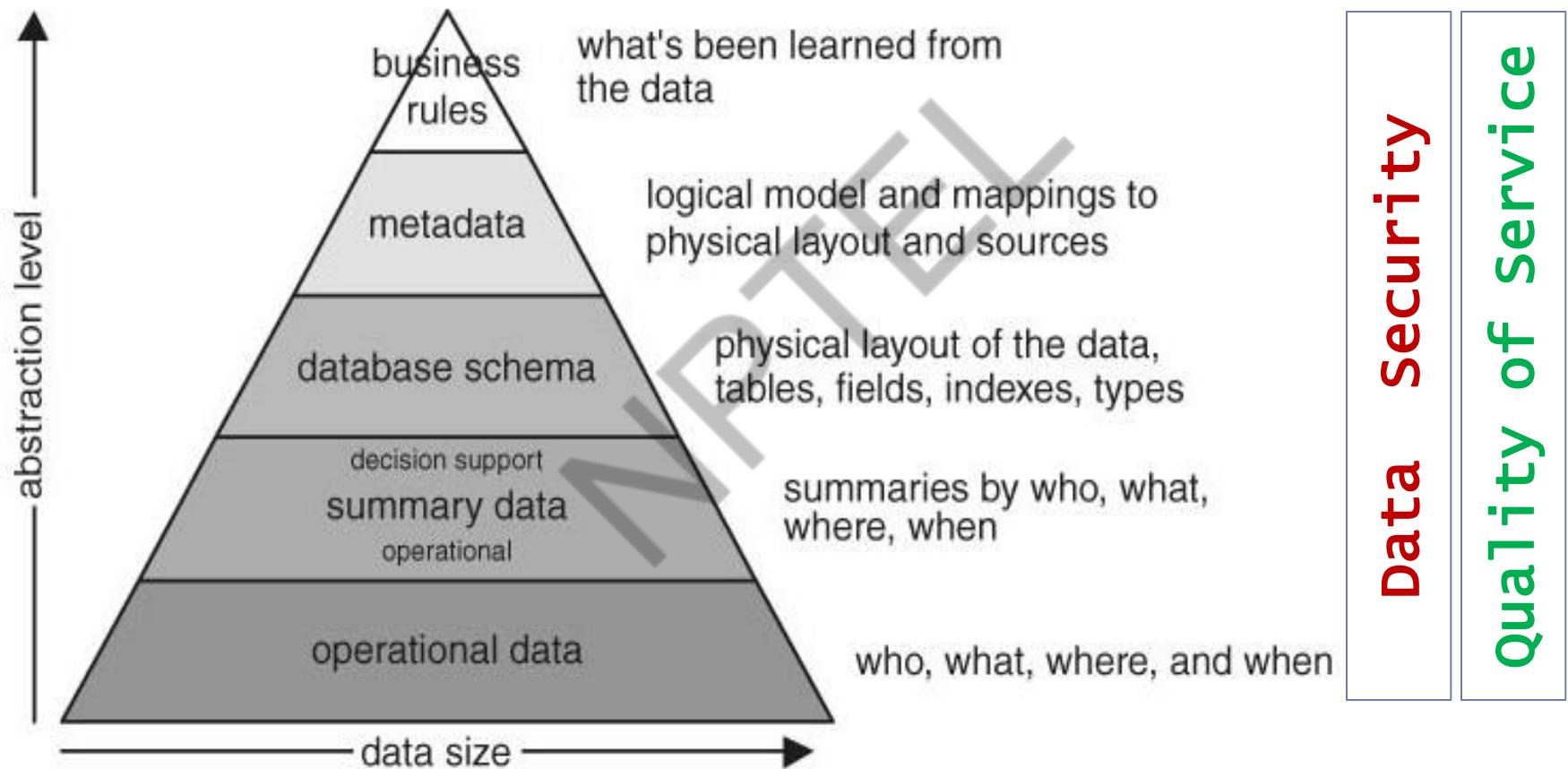
Geospatial Information System - Challenges

- ▶ 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

Geospatial Layers



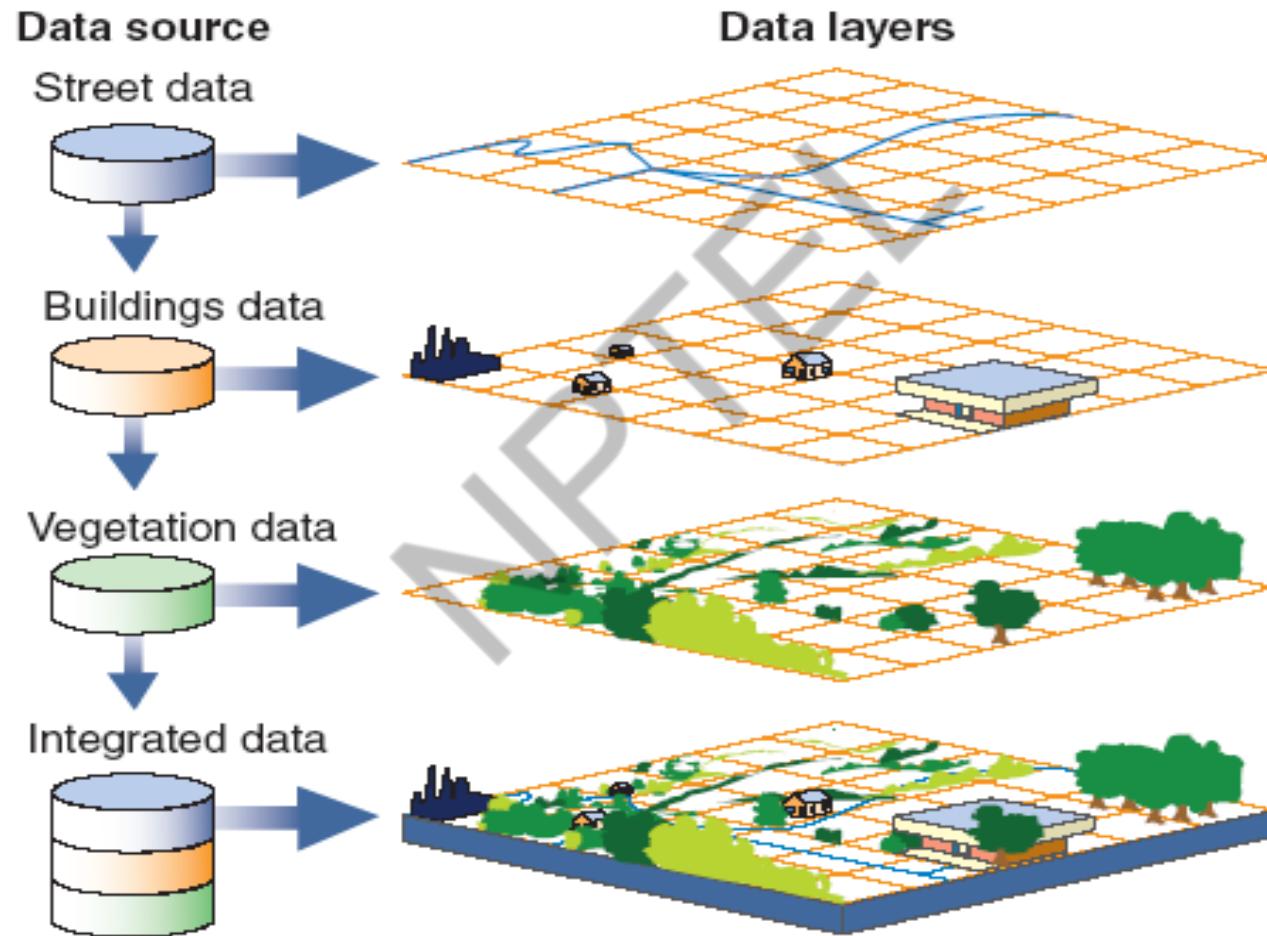
Generic Architecture of Data



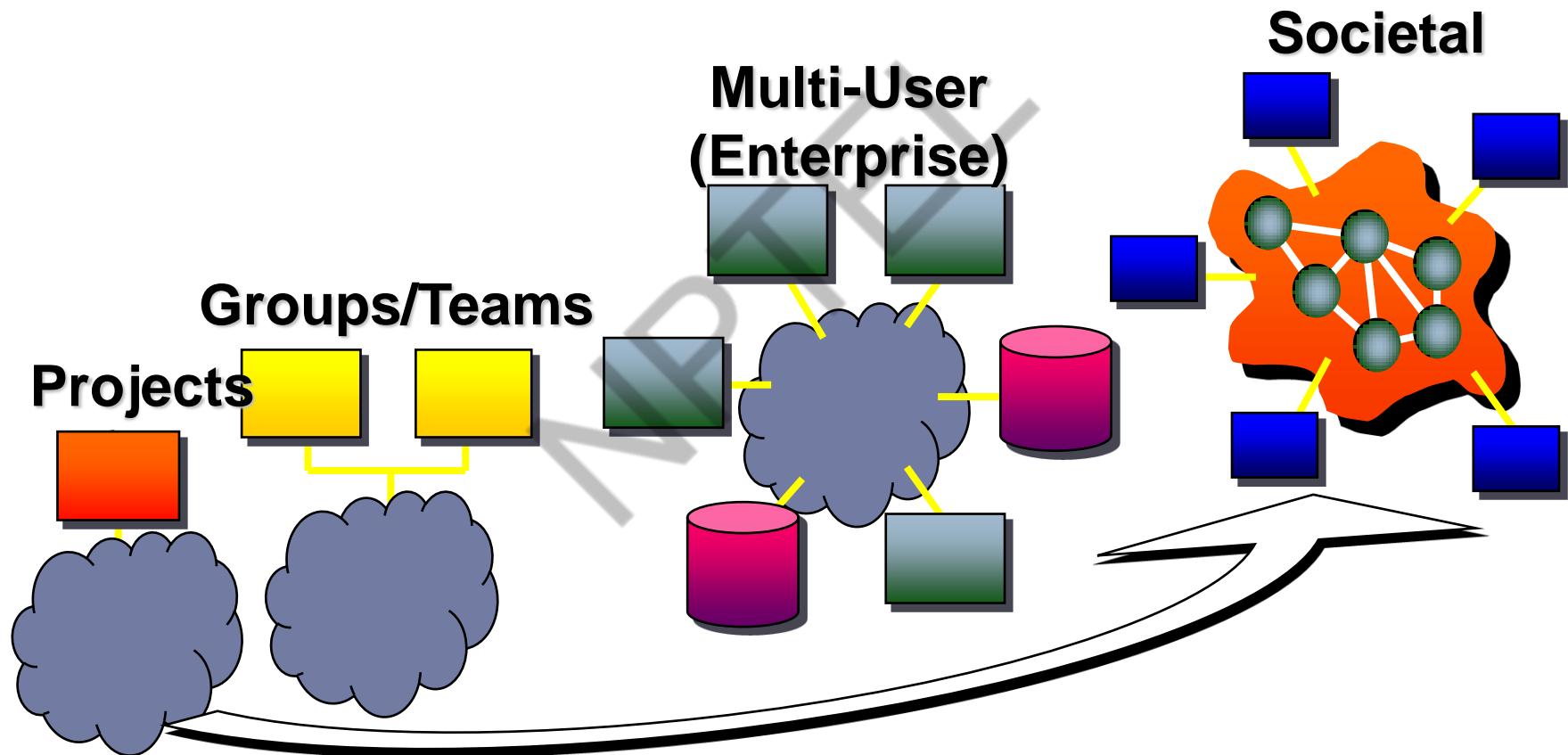
Heterogeneity Issue

- ▶ **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

Homogeneity (Needs to be achieved !)



GIS Users - Trend

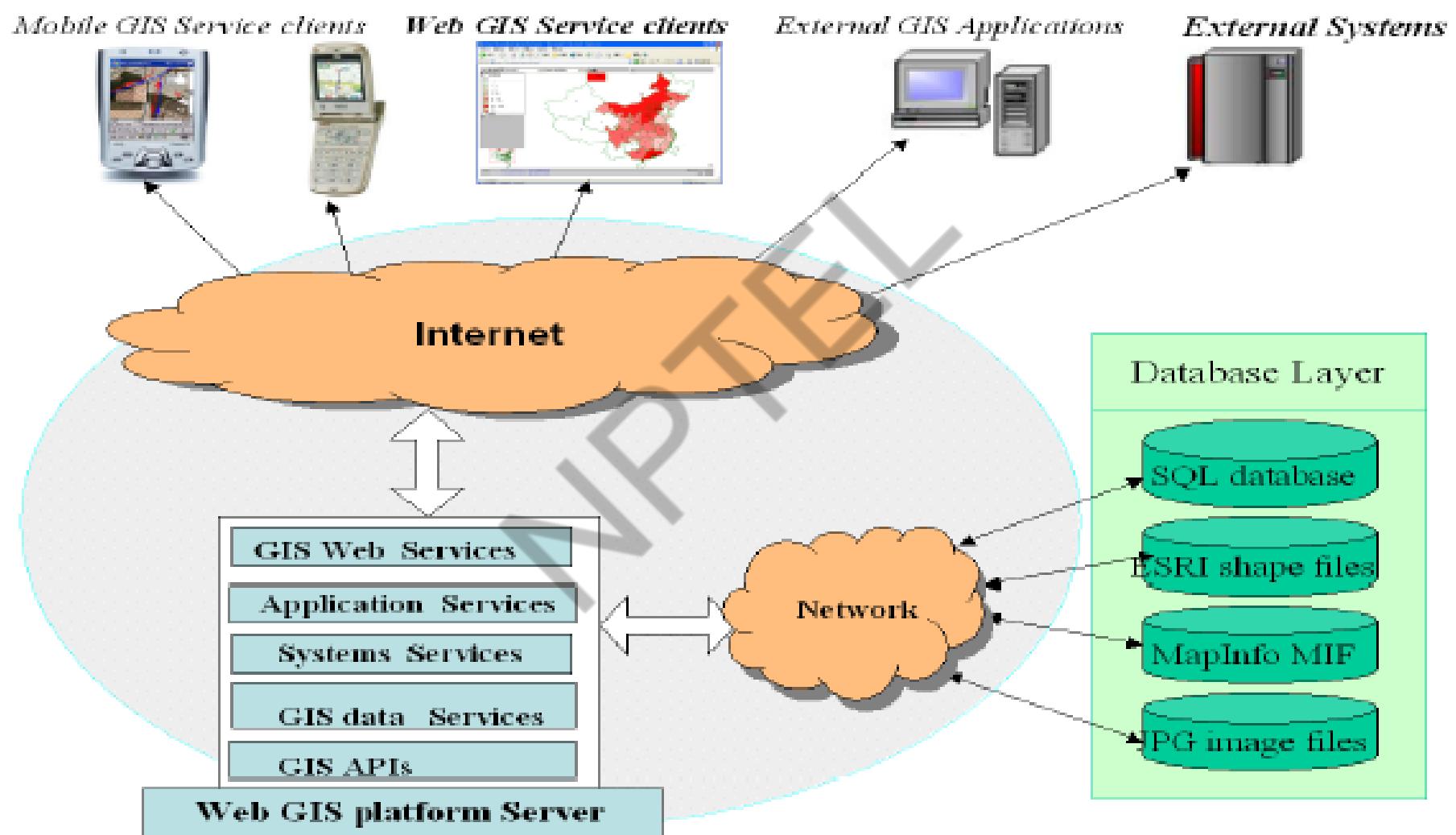


Spatial Data Infrastructure (SDI)

- ▶ “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

Interoperable GIS – Service driven



Need for Geospatial Cloud

- ▶ “Huge” volume of Data and Metadata
- ▶ Need of Services and Service Orchestration
- ▶ Evolving Standards and Policies
- ▶ Need for **Geospatial Cloud**



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



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.



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.*”



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



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

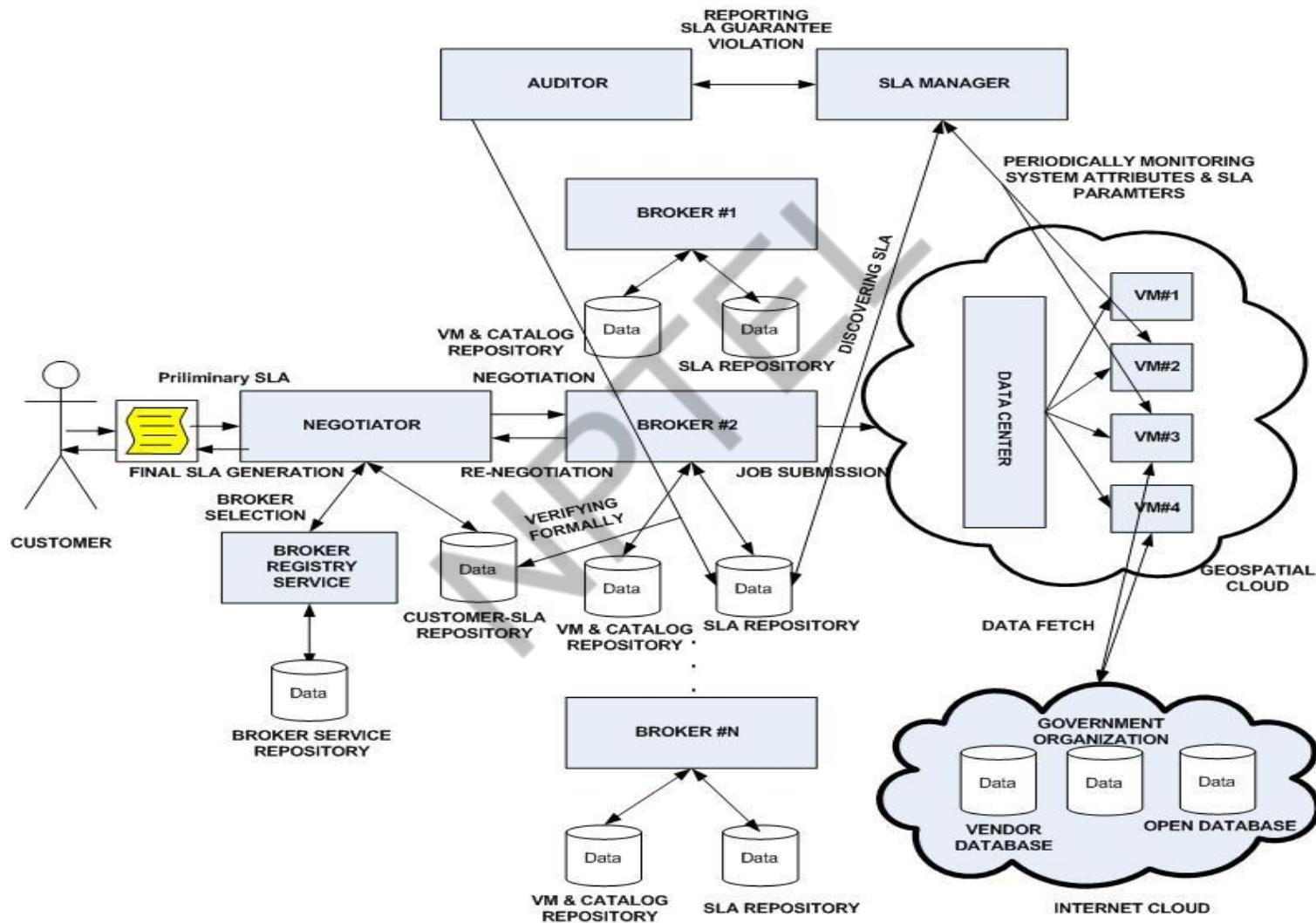


Cloud Actors

- ▶ 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)



Typical Geospatial Cloud Architecture

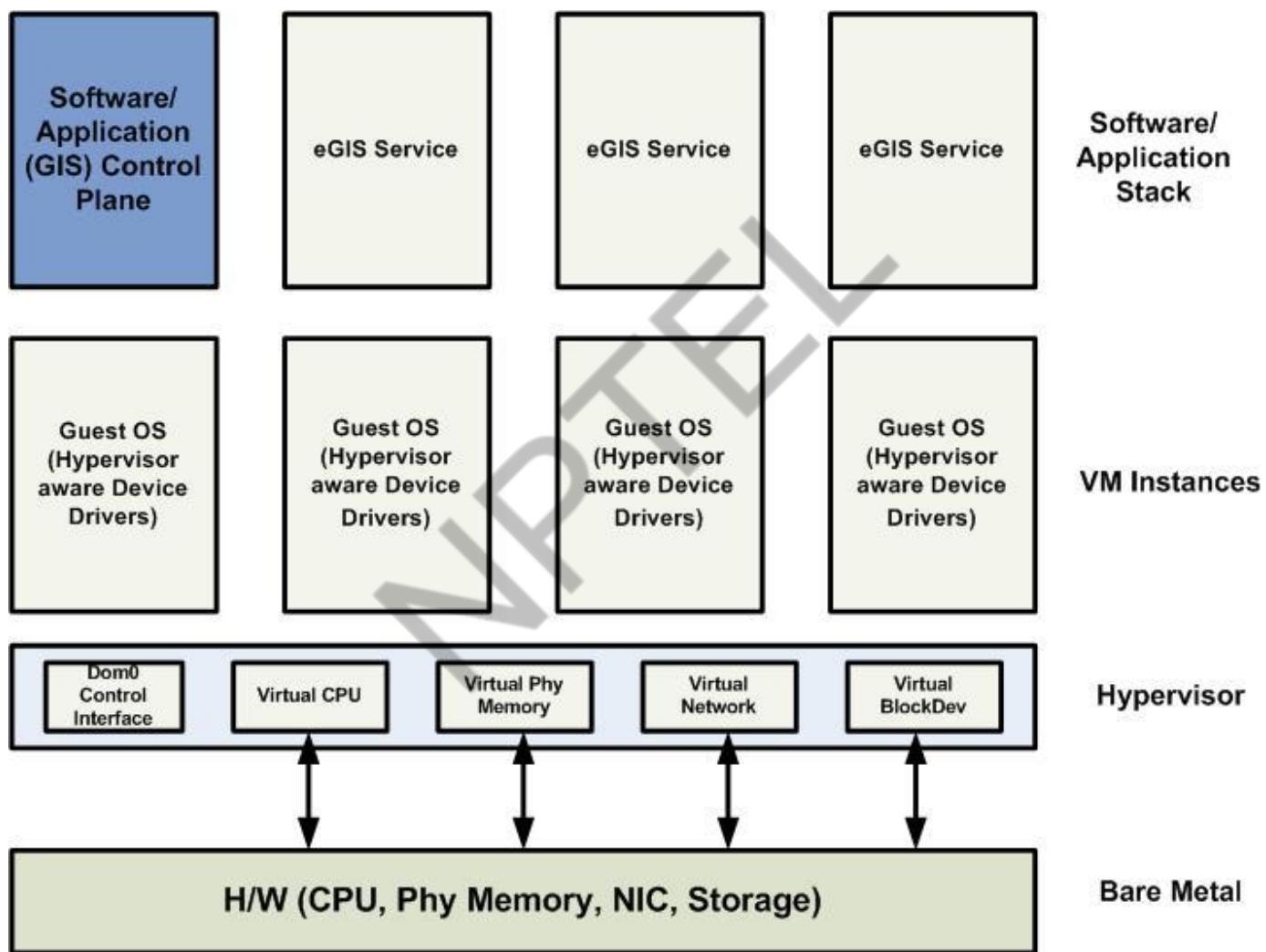


Cloud as Service Provider

- 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.



Geospatial Cloud

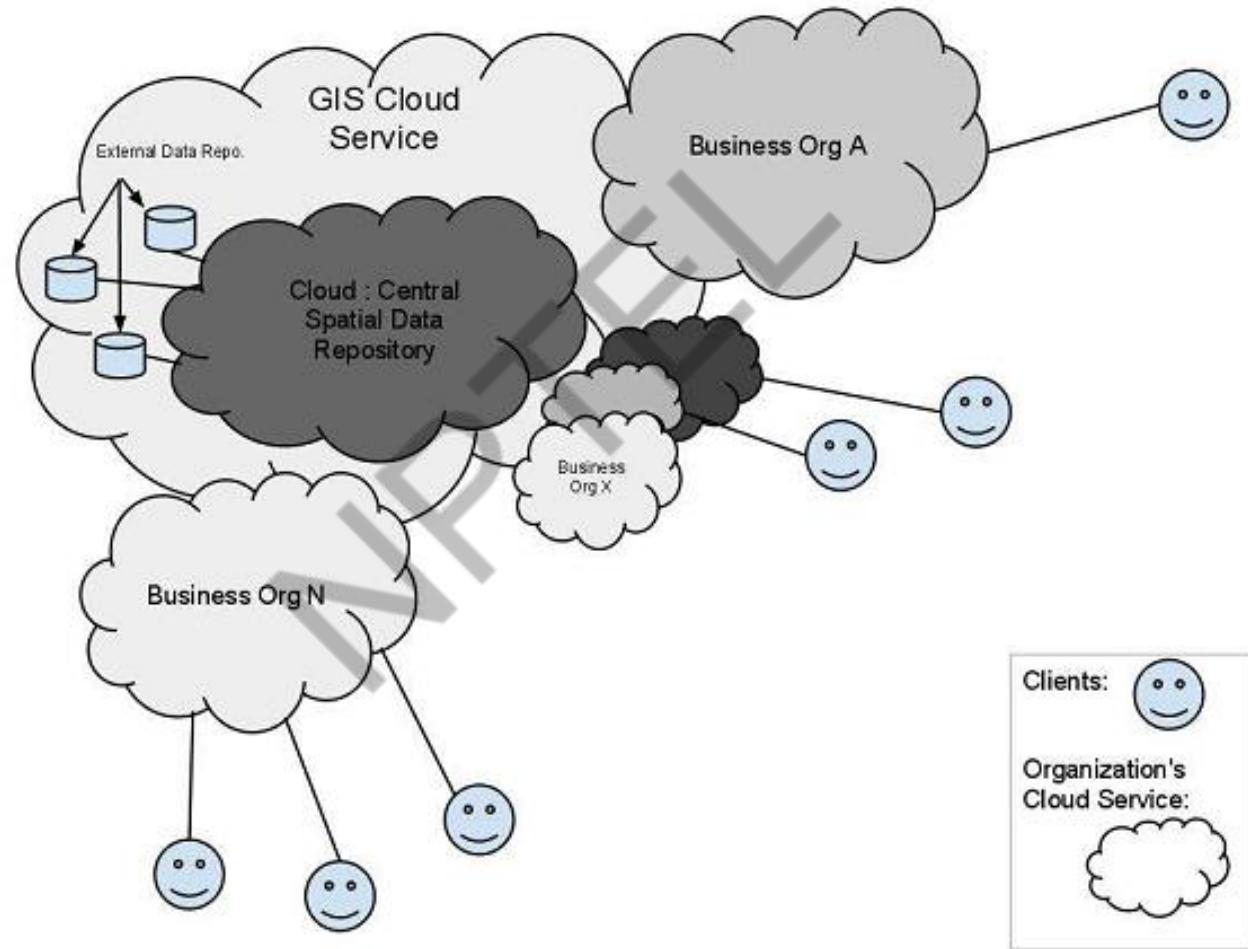


Geospatial Cloud Model

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.

Geospatial Cloud – Typical Scenario



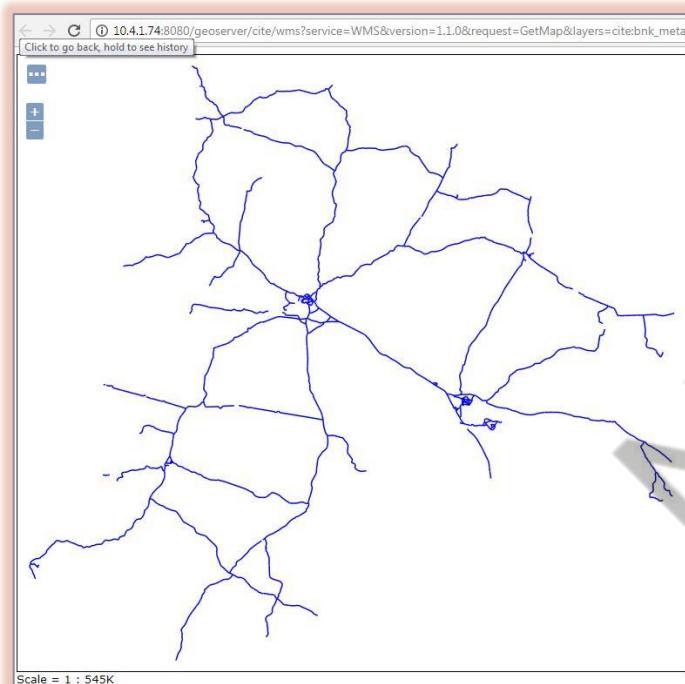
Geospatial Cloud

- ▶ 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.

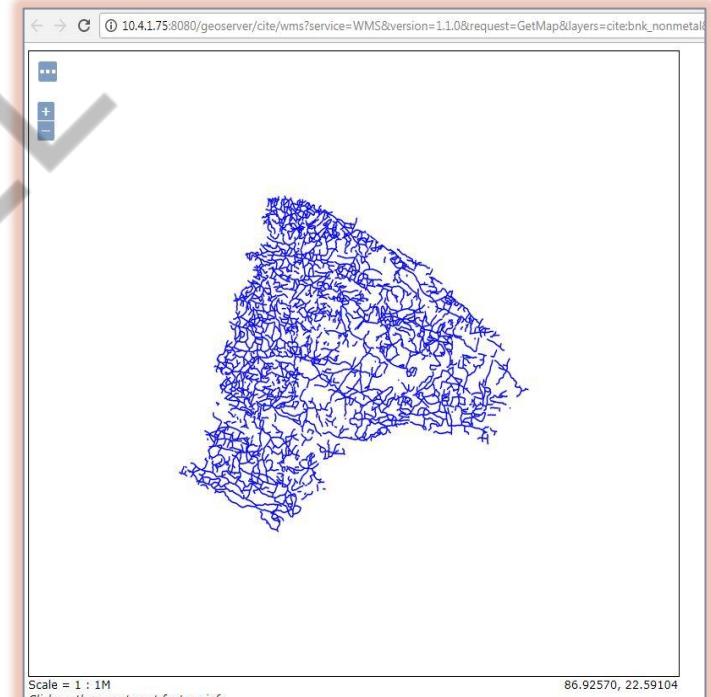
Experimental GeoSpatial-Cloud @IITKgp



Service Integration for Query in Cloud (Case Study 1)



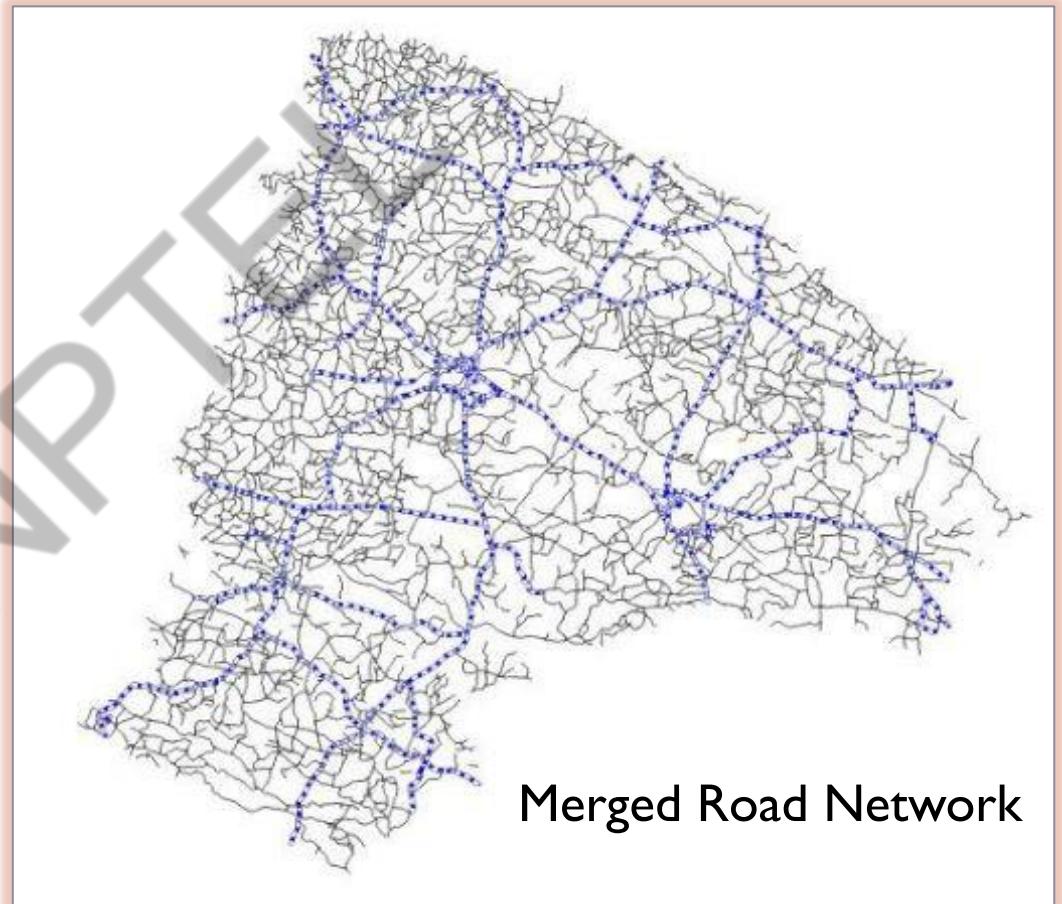
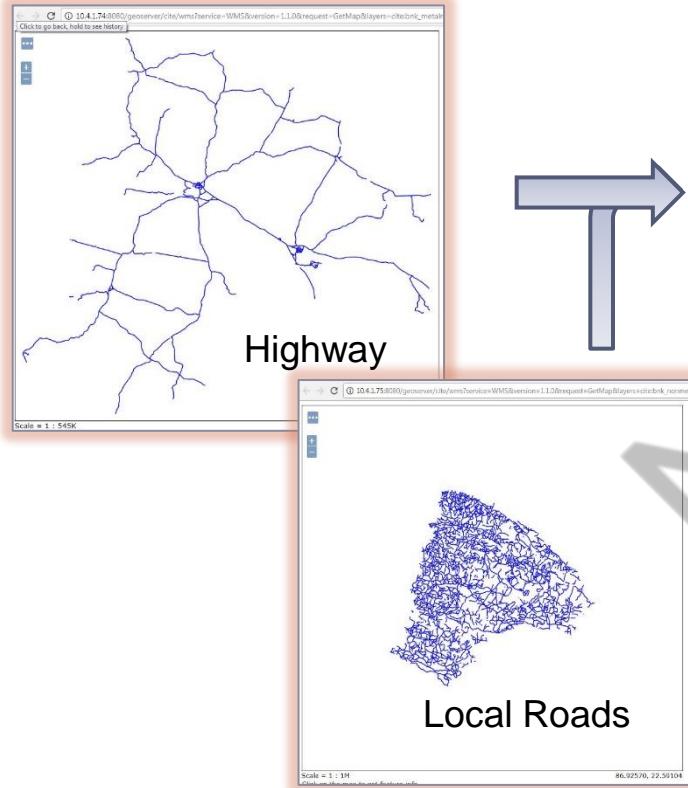
Highway



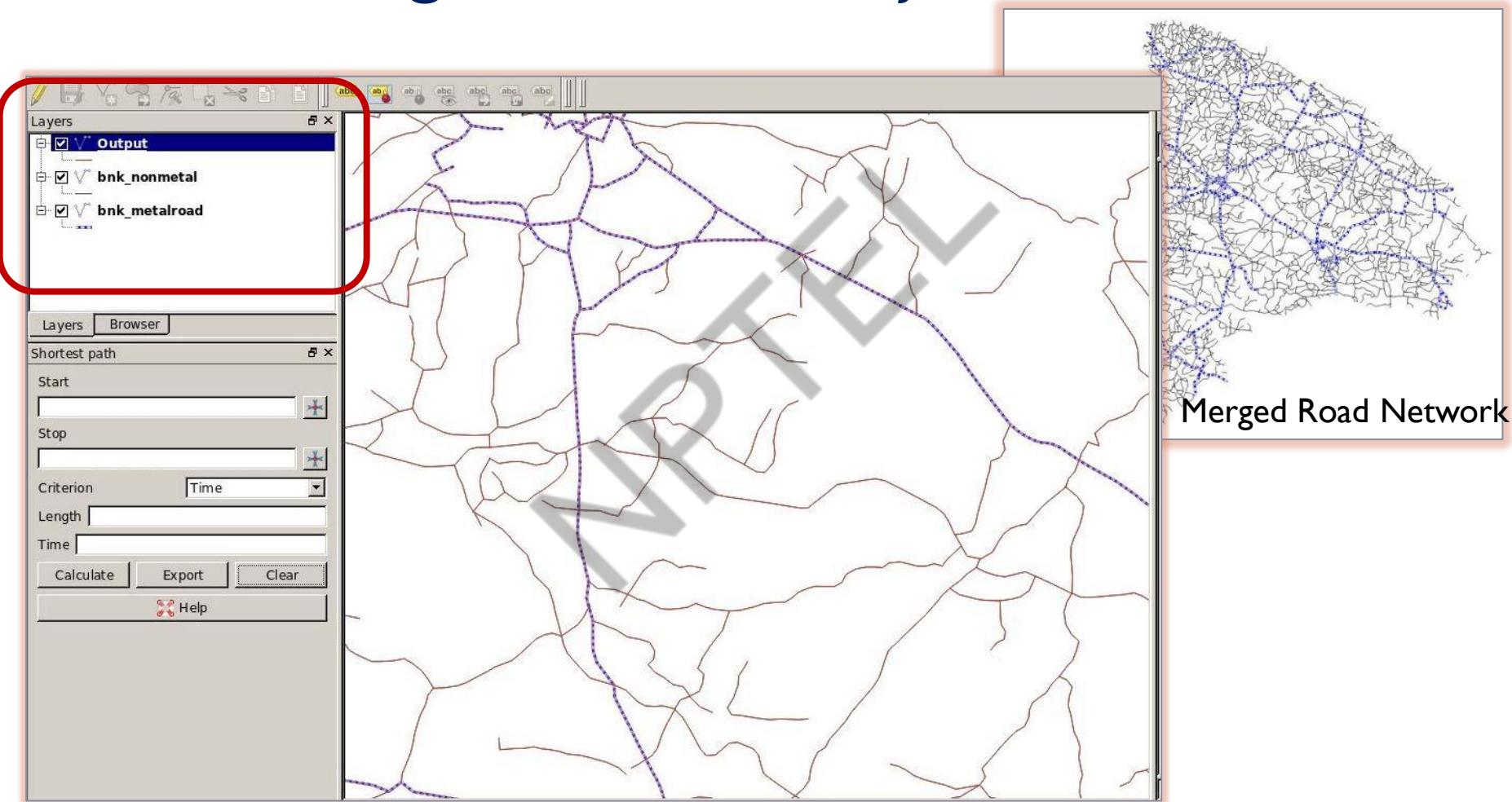
Local Roads



Service Integration for Query in Cloud



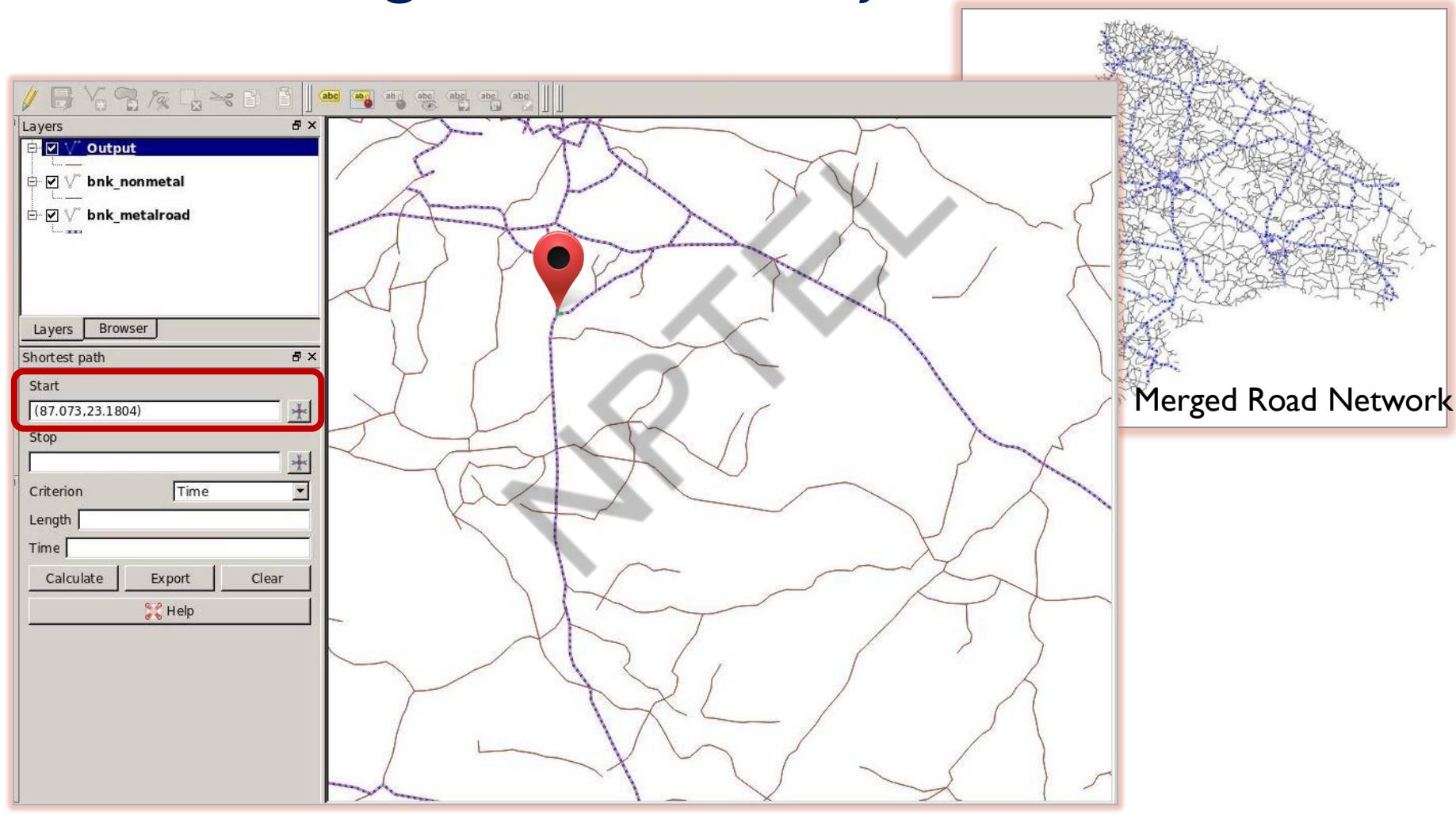
Service Integration for Query in Cloud



Shortest Path Calculation
CSE, IIT Kharagpur

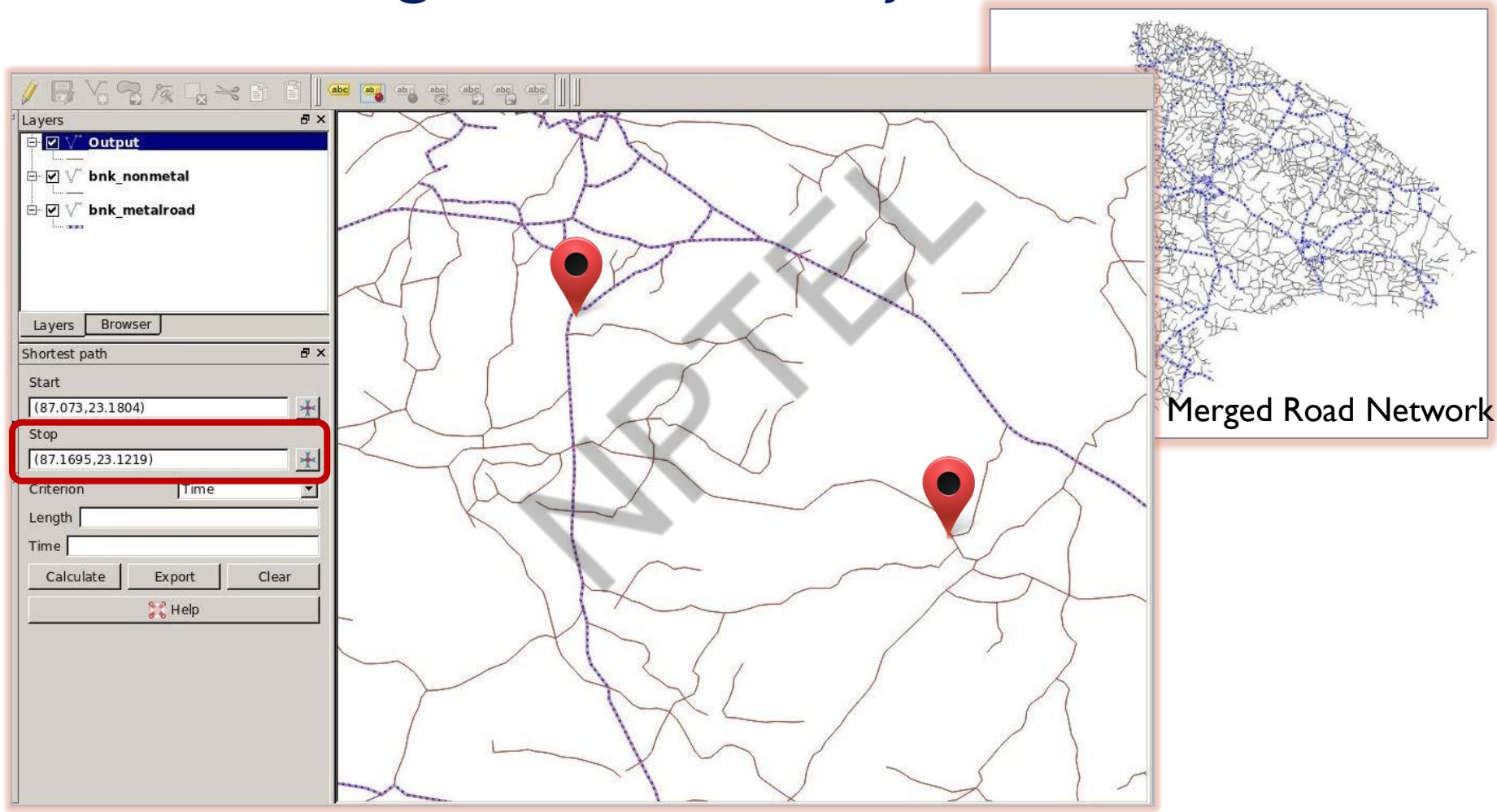


Service Integration for Query in Cloud



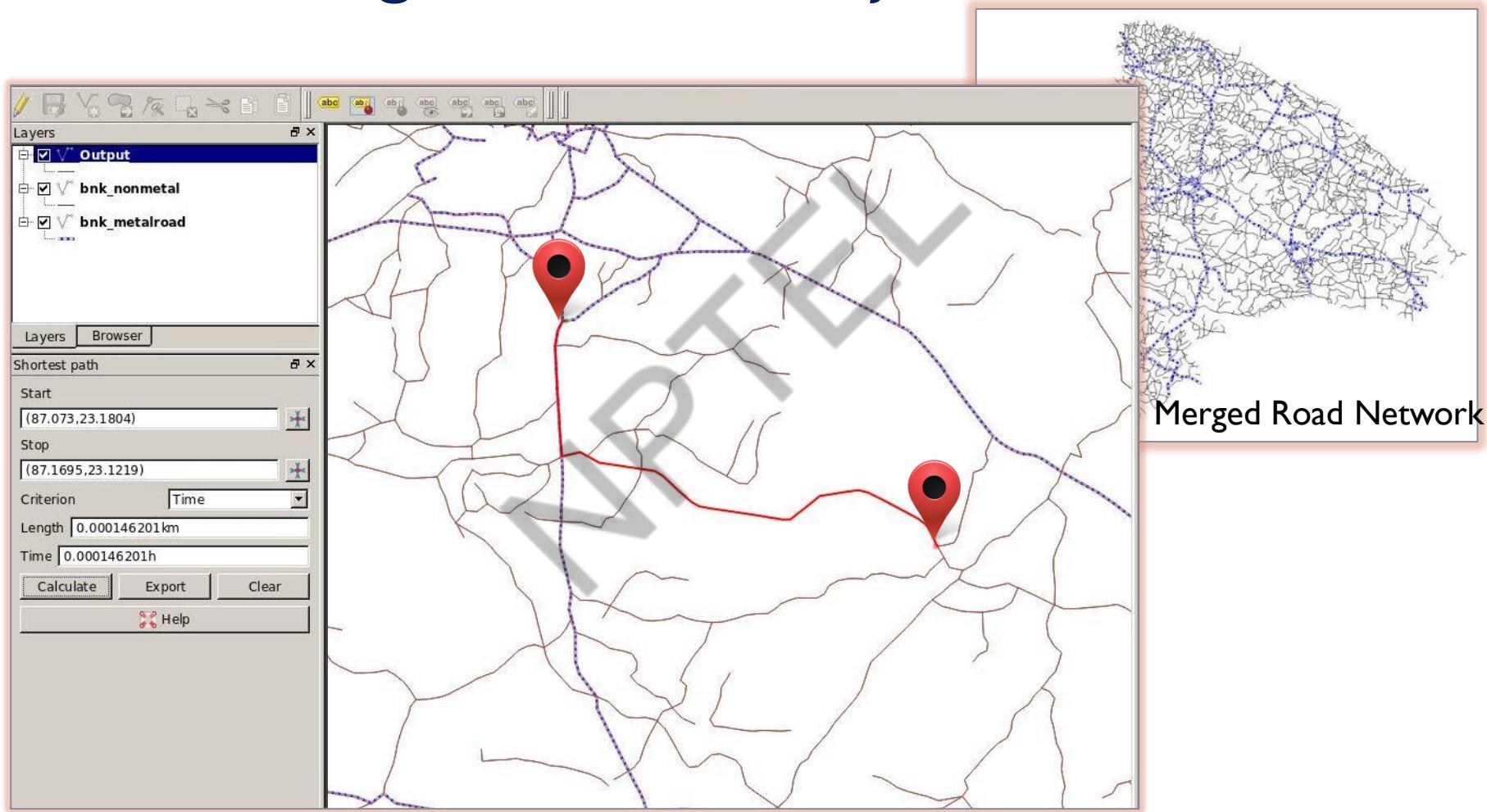
Shortest Path Calculation
CSE, IIT Kharagpur

Service Integration for Query in Cloud



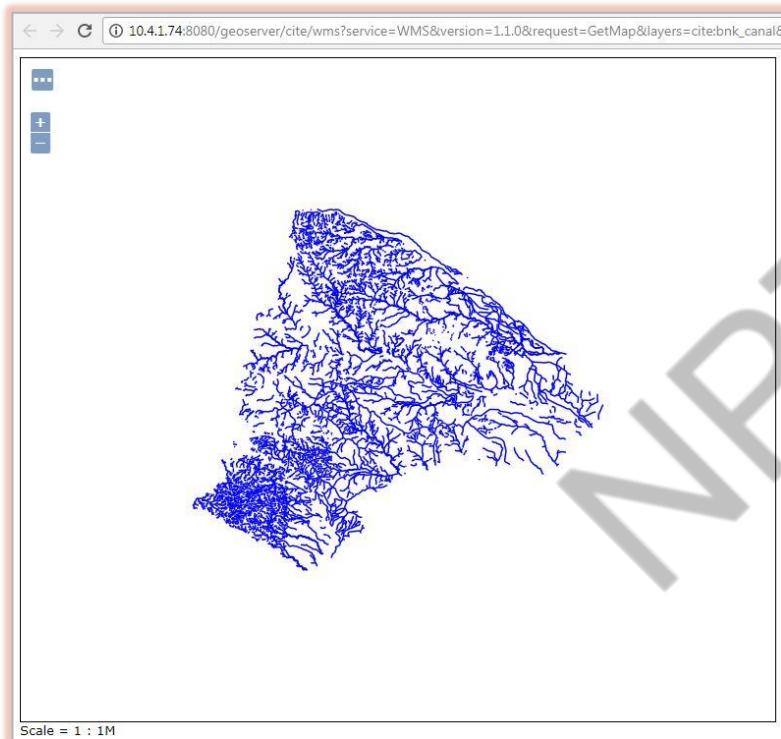
Shortest Path Calculation
CSE, IIT Kharagpur

Service Integration for Query in Cloud

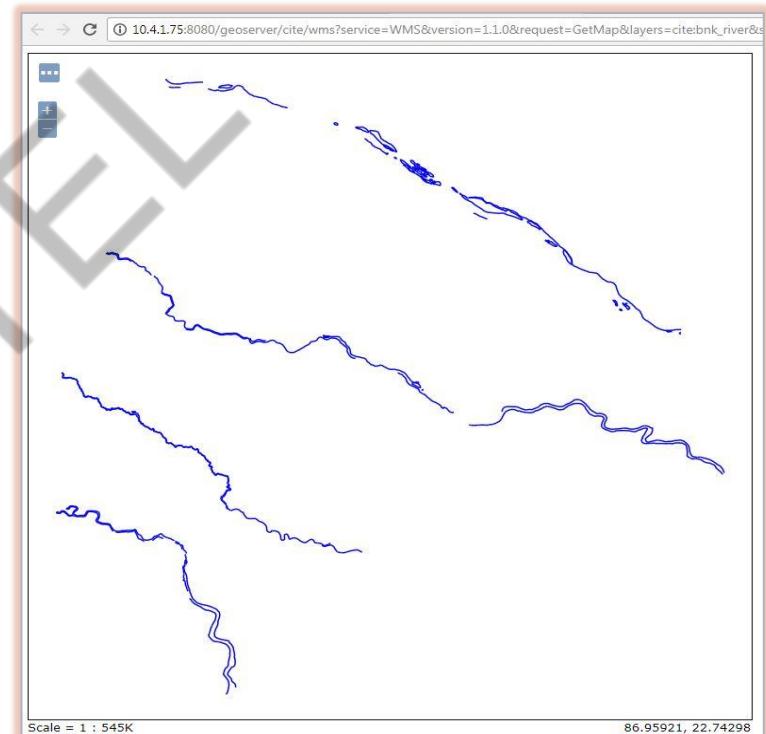


Shortest Path Calculation
CSE, IIT Kharagpur

Service Integration for Query in Cloud (Case Study 2)



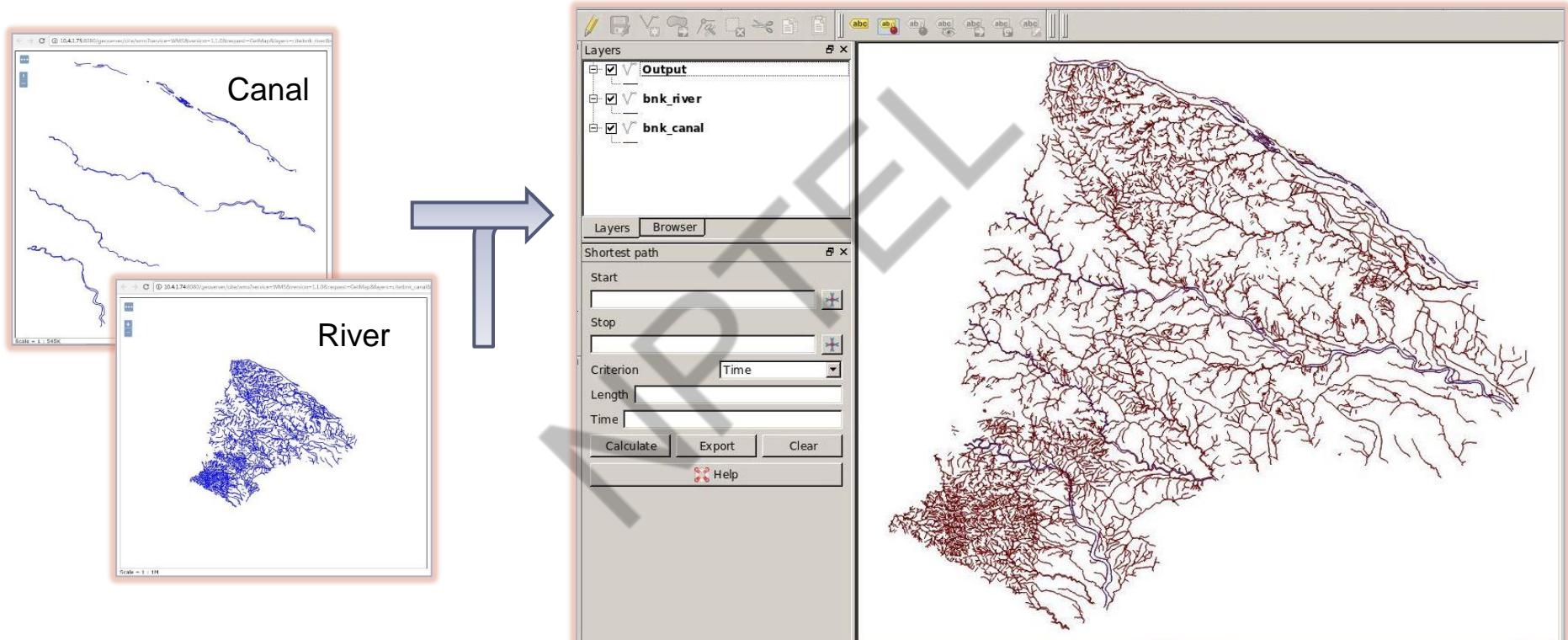
Canal



River



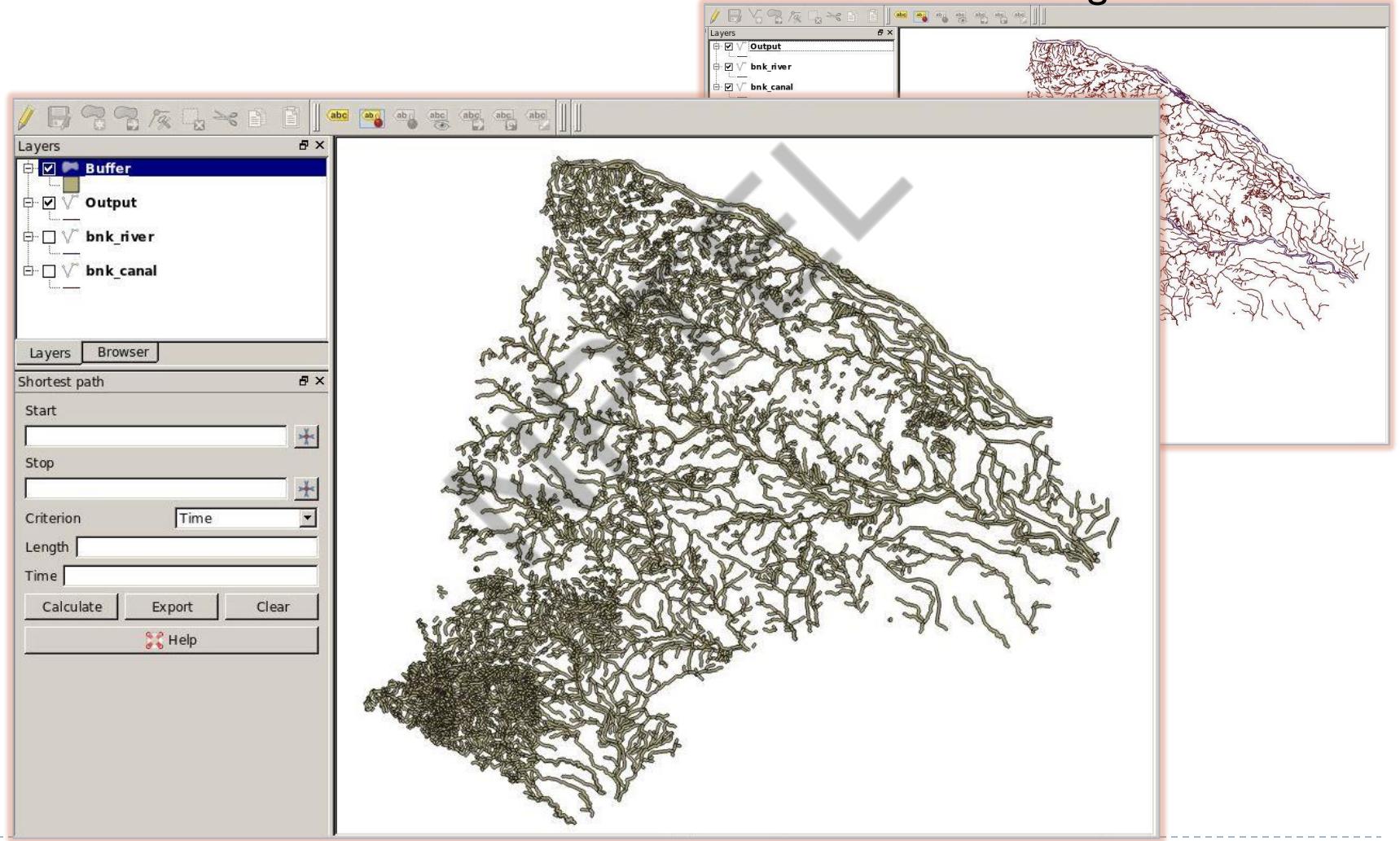
Service Integration for Query in Cloud



Merged Water Network

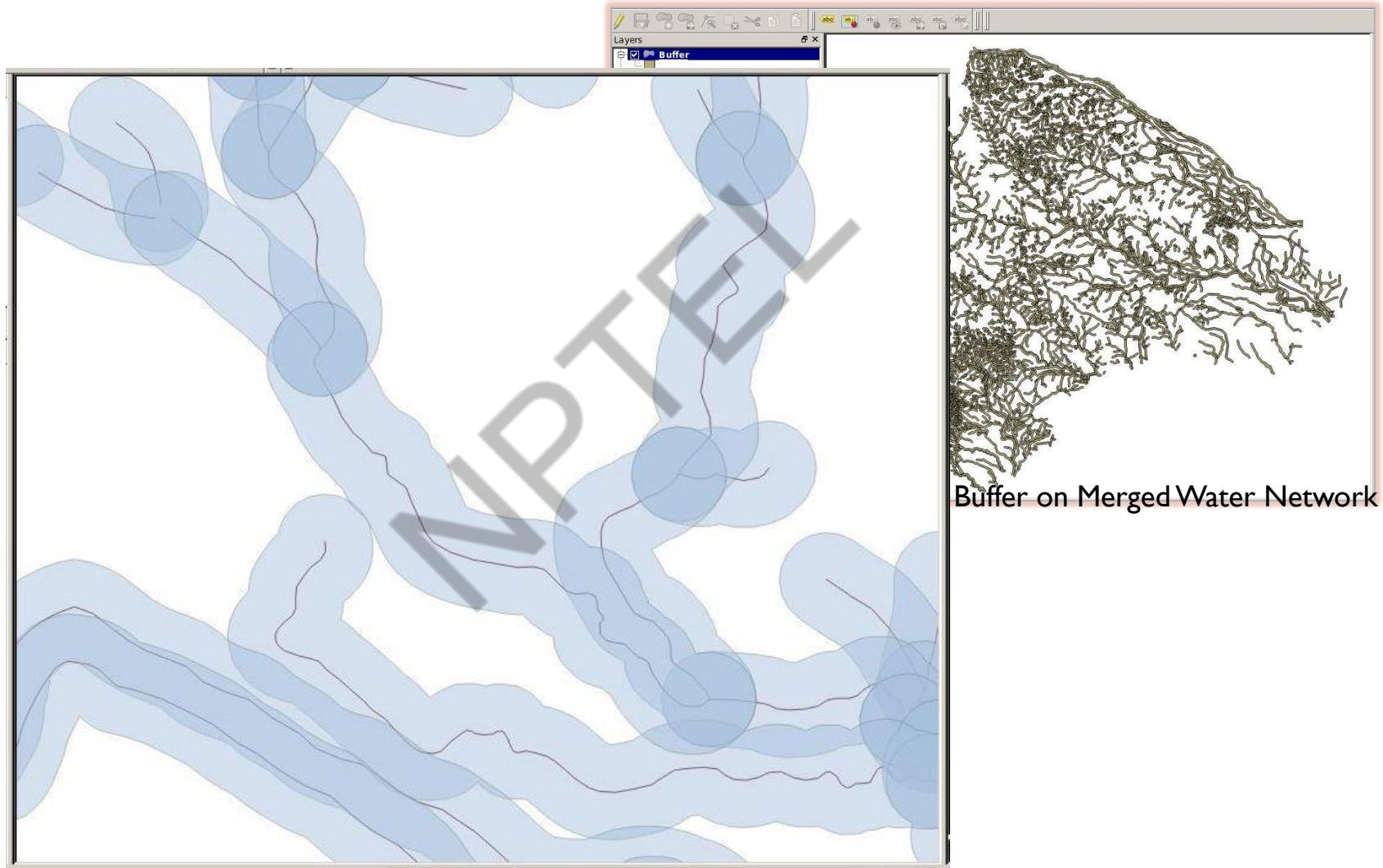
Service Integration for Query in Cloud

Merged Water Network



▶ Buffer on Merged Water Network

Service Integration for Query in Cloud



▶ Buffer on Merged Water Network (Zoomed)

Challenges in Geospatial Cloud

A large, semi-transparent watermark reading "NPTEL" diagonally across the center of the slide.

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

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



Geo-Cloud – Major Security Concern

- ▶ Multi-tenancy
- ▶ Lack of complete control - data, applications, services

NPTEL



Concerns

- ▶ 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



Thank You !

NPTEL





IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

CLOUD COMPUTING

Introduction to DOCKER Container

PROF. SOUMYA K. GHOSH

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
IIT KHARAGPUR

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/>

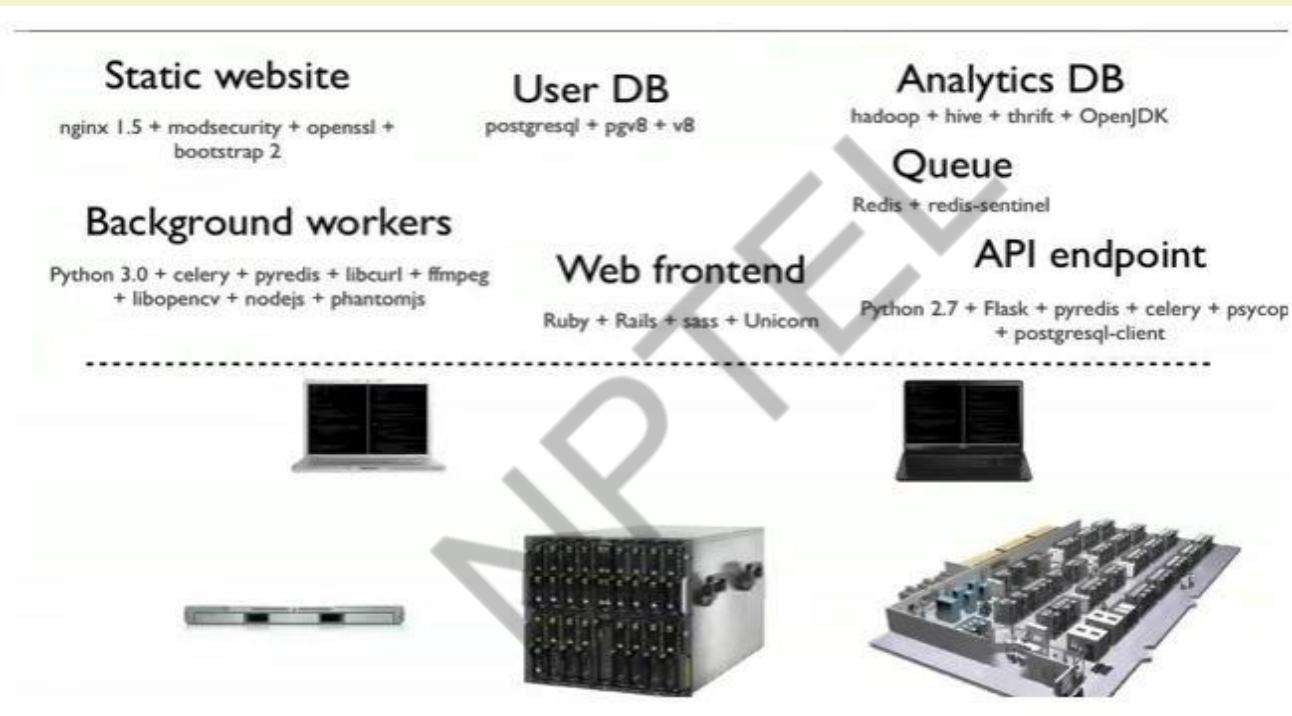


IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Infrastructure and Software Stack



Ref: Internet/YouTube



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Goal: Interoperability

Static website	?	?	?	?	?	?	?
Web frontend	?	?	?	?	?	?	?
Background workers	?	?	?	?	?	?	?
User DB	?	?	?	?	?	?	?
Analytics DB	?	?	?	?	?	?	?
Queue	?	?	?	?	?	?	?

Ref: Internet/YouTube



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

“Shipping”



Ref: Internet/YouTube



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

“Shipping”



Ref: Internet/YouTube

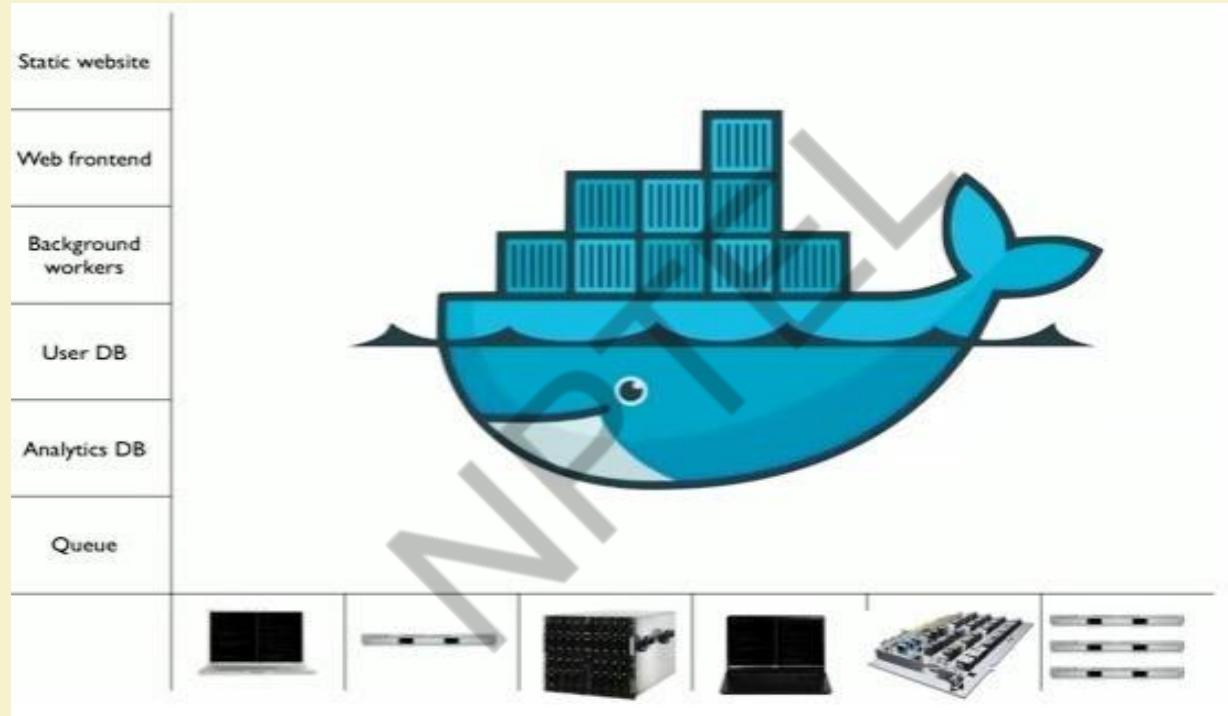


IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

“Docker”



Ref: Internet/YouTube



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

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
NPTEL ONLINE
CERTIFICATION COURSES

Docker – Components

- Docker for Mac – It allows one to run Docker containers on the Mac OS.
- Docker for Linux – It allows one to run Docker containers on the Linux OS.
- Docker for Windows – It allows one to run Docker containers on the Windows OS.
- Docker Engine – It is used for building Docker images and creating Docker containers.
- Docker Hub – This is the registry which is used to host various Docker images.
- Docker Compose – This is used to define applications using multiple Docker containers.

Ref: <https://www.tutorialspoint.com/docker/>



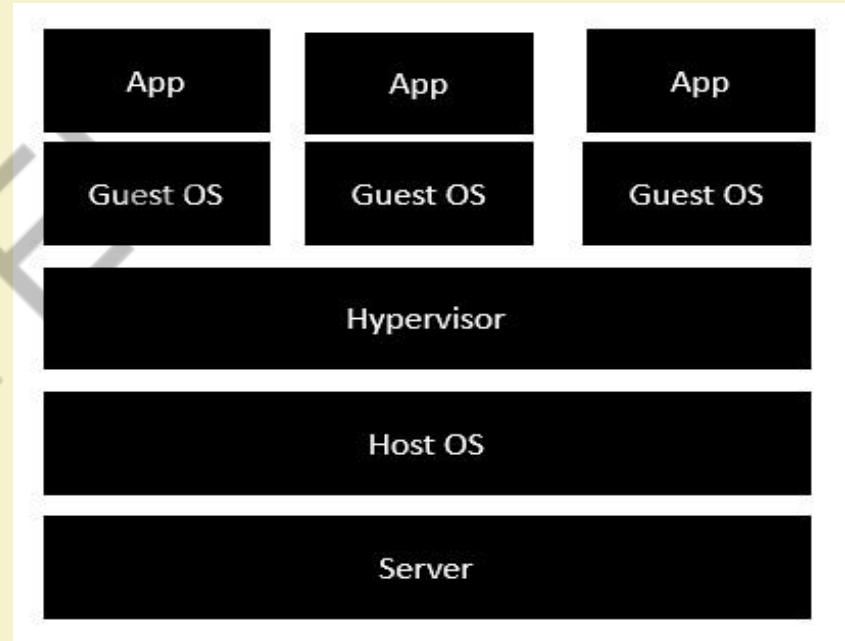
IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

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.



Ref: <https://www.tutorialspoint.com/docker/>



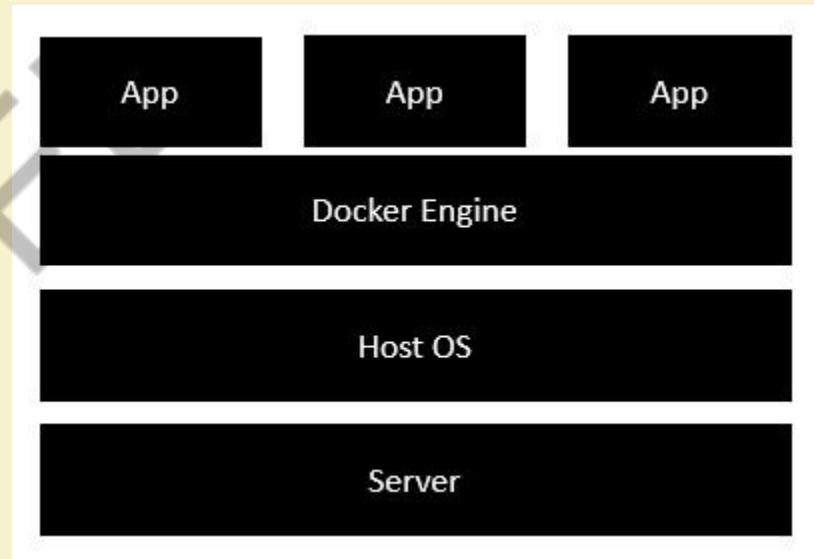
IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

Docker – Architecture

- Server is the physical server that is used to host multiple virtual machines.
- Host OS is the base machine such as Linux or Windows.
- Docker engine is used to run the operating system which earlier used to be virtual machines as Docker containers.
- All of the Apps now run as Docker containers.



Ref: <https://www.tutorialspoint.com/docker/>

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

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/>

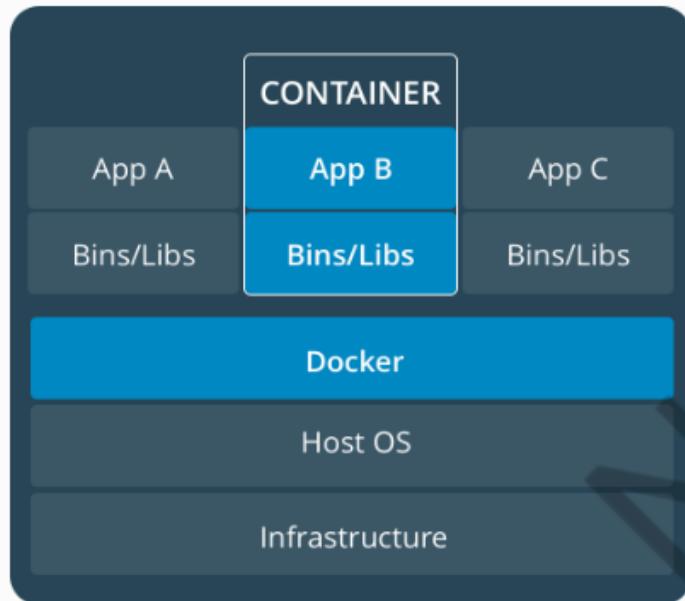


IIT KHARAGPUR

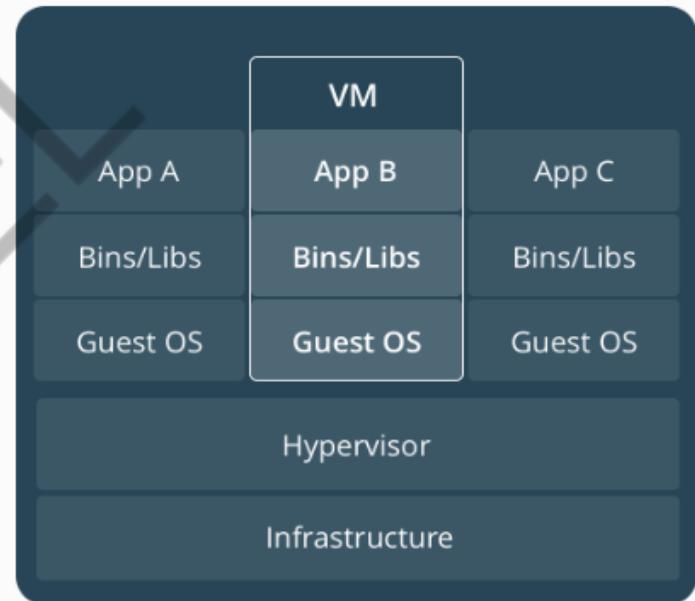


NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Containers and Virtual Machines



Container



VM

Ref: <https://www.docker.com/>



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Virtual Machines and Containers

- **Virtual machines** run guest operating systems - the OS layer in each box.
- Resource intensive, and the resulting disk image and application state is an entanglement of OS settings, system-installed dependencies, OS security patches, and other easy-to-lose, hard-to-replicate ephemera.
- **Containers** can share a single kernel, and the only information that needs to be in a container image is the executable and its package dependencies, which never need to be installed on the host system.
- These processes run like native processes, and can be managed individually
- Because they contain all their dependencies, there is no configuration entanglement; a containerized app “runs anywhere”

Ref: <https://www.docker.com/>

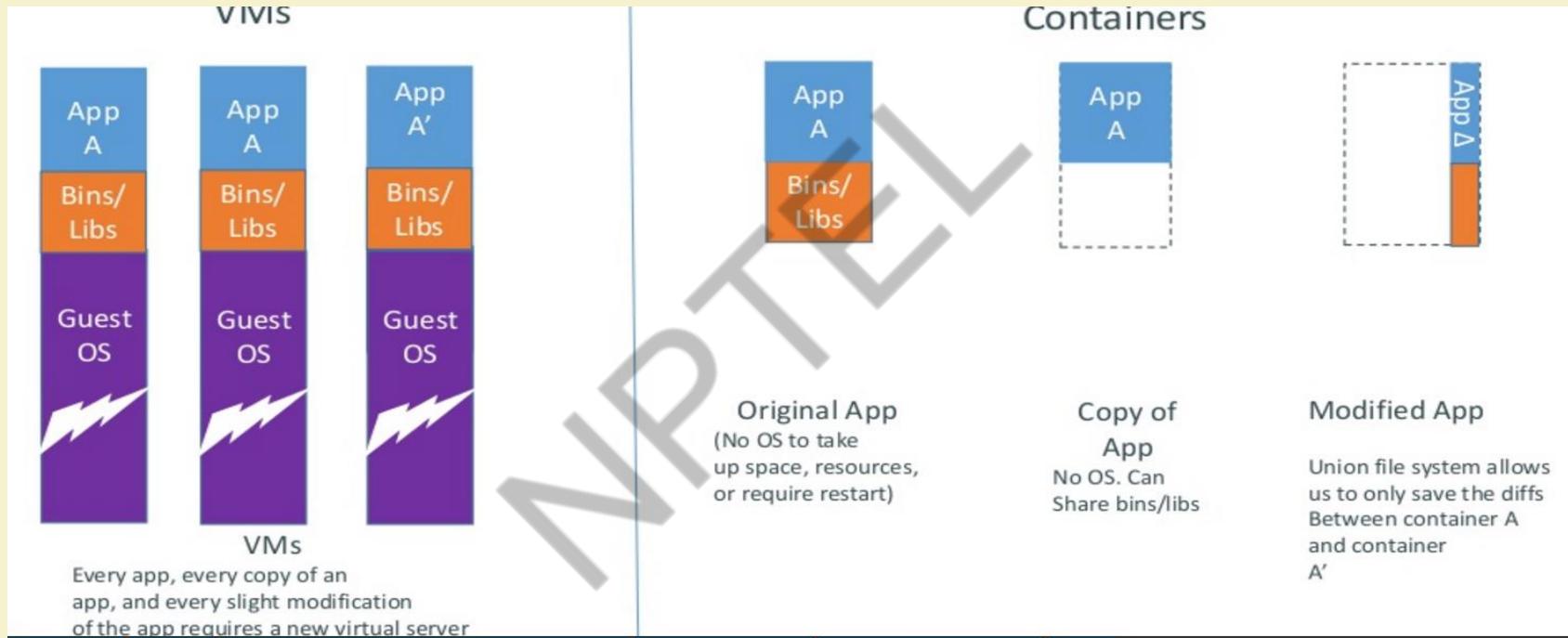


IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

Docker containers are lightweight



Ref: Internet

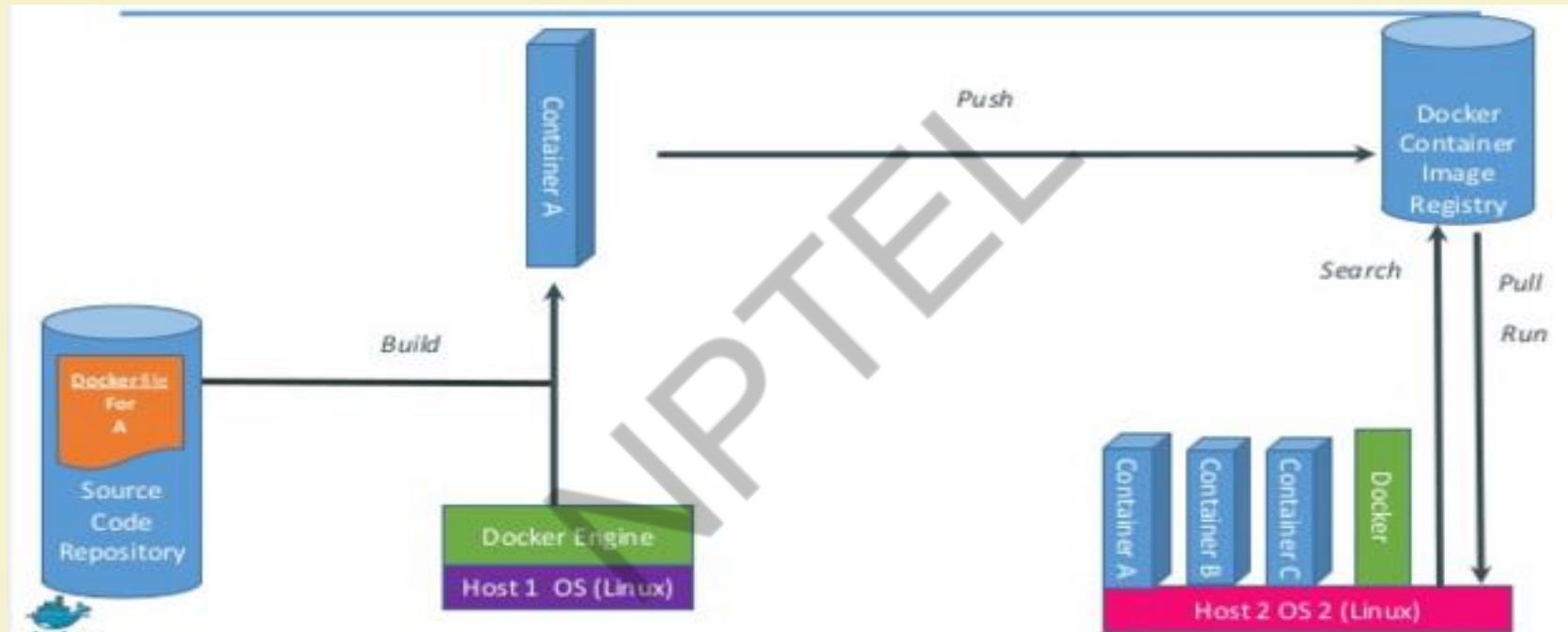


IIT KHARAGPUR



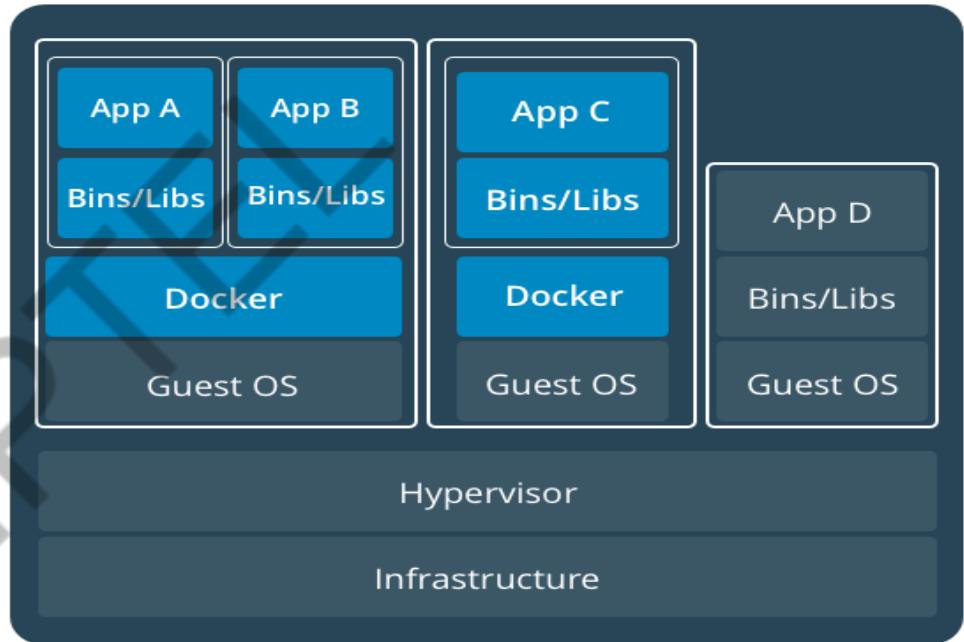
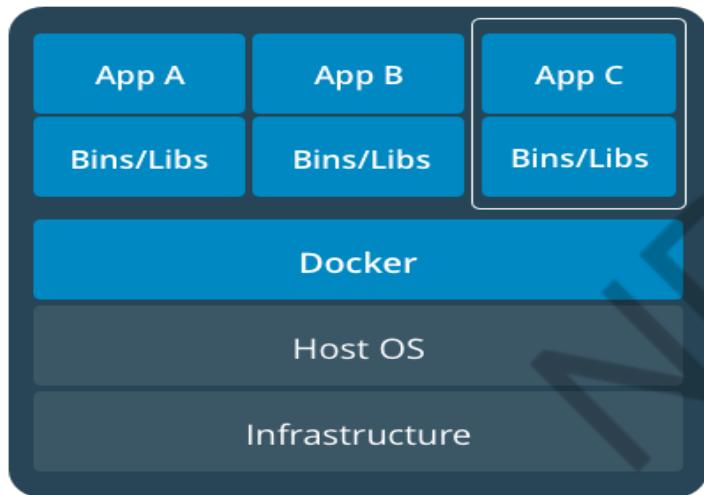
NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

How does Docker work



Source: Internet

Containers and Virtual Machines Together



Ref: <https://www.docker.com/>

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/>

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
NPTEL ONLINE
CERTIFICATION COURSES

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

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
NPTEL ONLINE
CERTIFICATION COURSES

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
NPTEL ONLINE
CERTIFICATION COURSES

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

Thank You!

NPTEL



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES



IIT KHARAGPUR



NPTEL

NPTEL ONLINE
CERTIFICATION COURSES

CLOUD COMPUTING

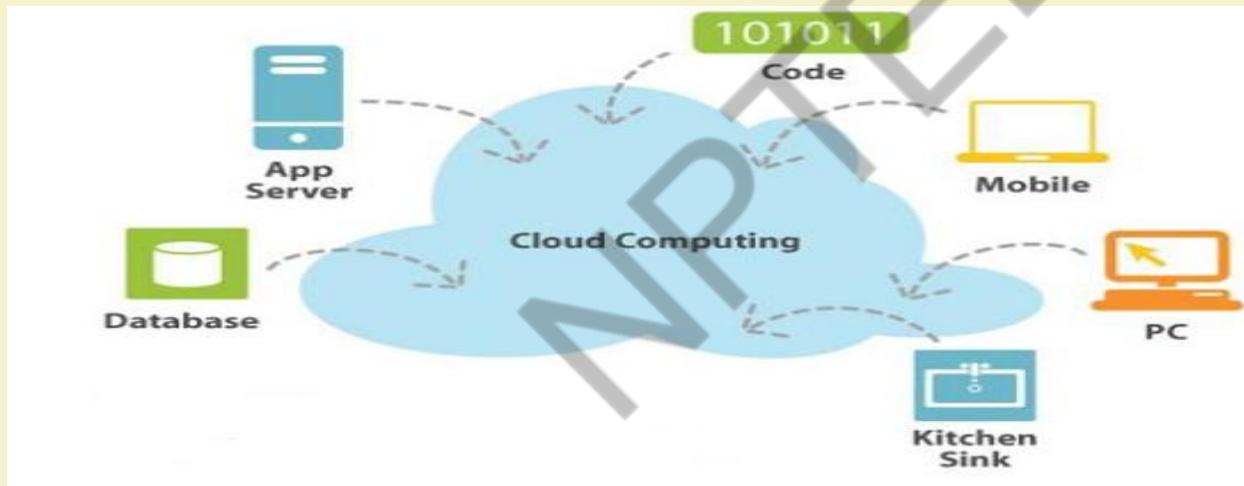
Green Cloud

PROF. SOUMYA K. GHOSH

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
IIT KHARAGPUR

Cloud Computing

Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources like networks, servers, storage, applications, and services.



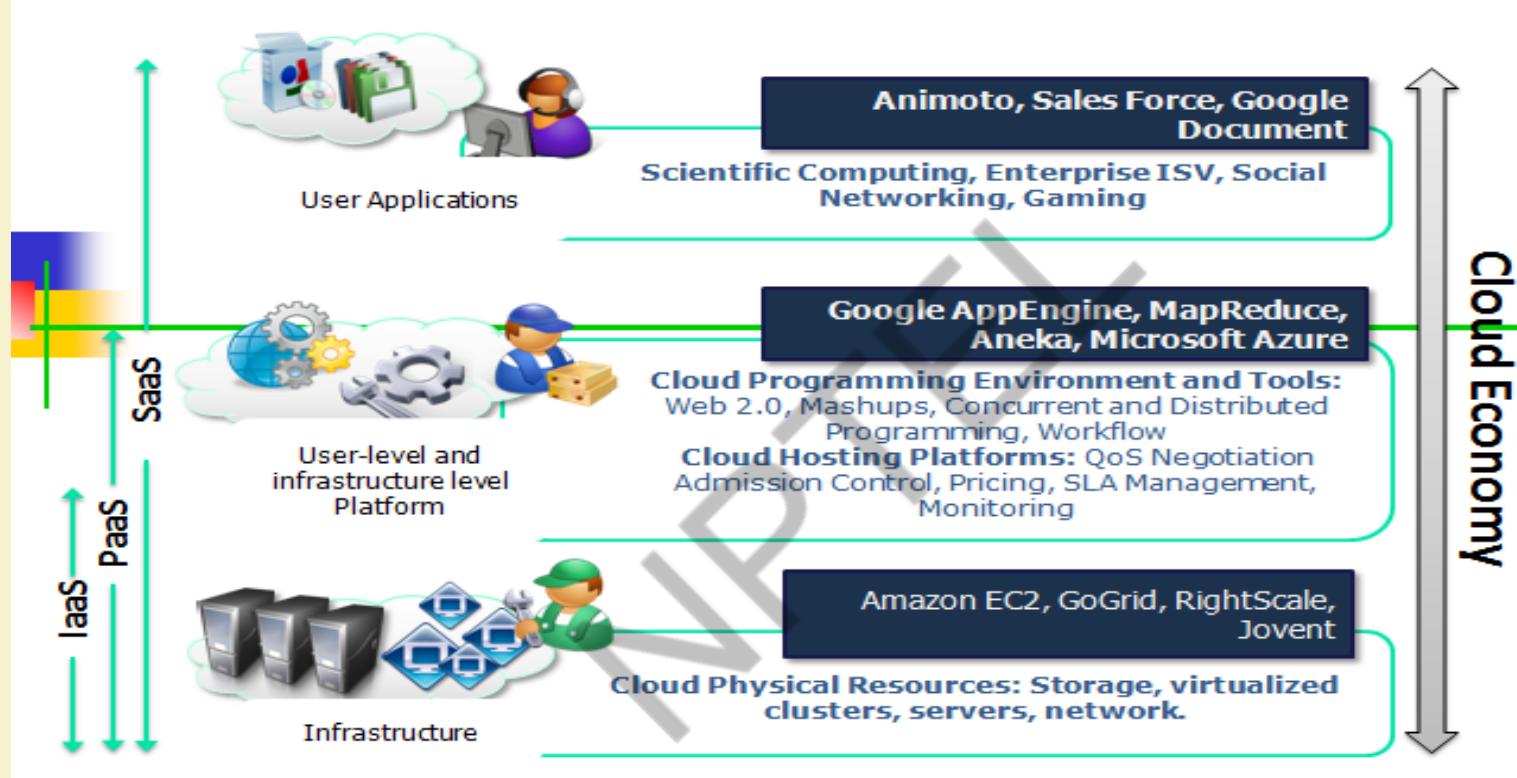
Source: Internet



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES



Source: Internet



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

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



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

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 – 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

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

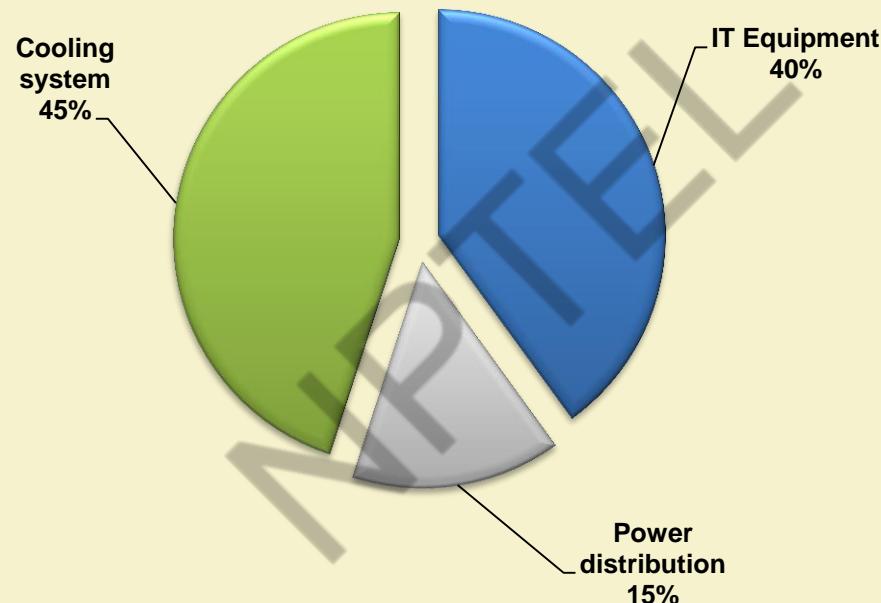


IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

Typical Data Center Energy Consumption

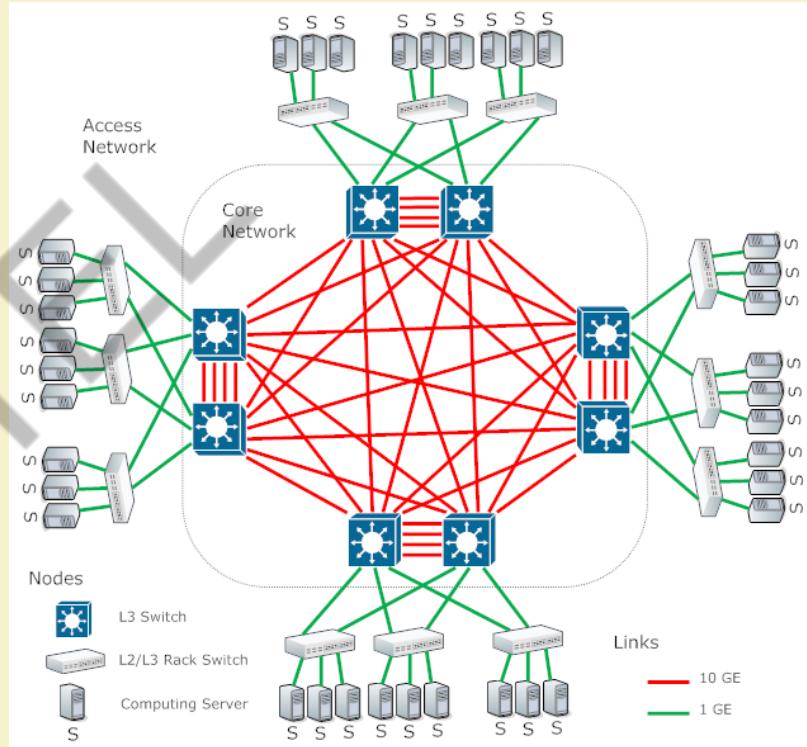


Ref: Dzmitry Kliazovich, University of Luxembourg

DC Architecture - Past

Two-tier DC architecture

- Access and Core layers
- 1 GE and 10 GE links
- Full mesh core network
- Load balancing using ICMP



Ref: Dzmitry Kliazovich, University of Luxembourg



IIT KHARAGPUR

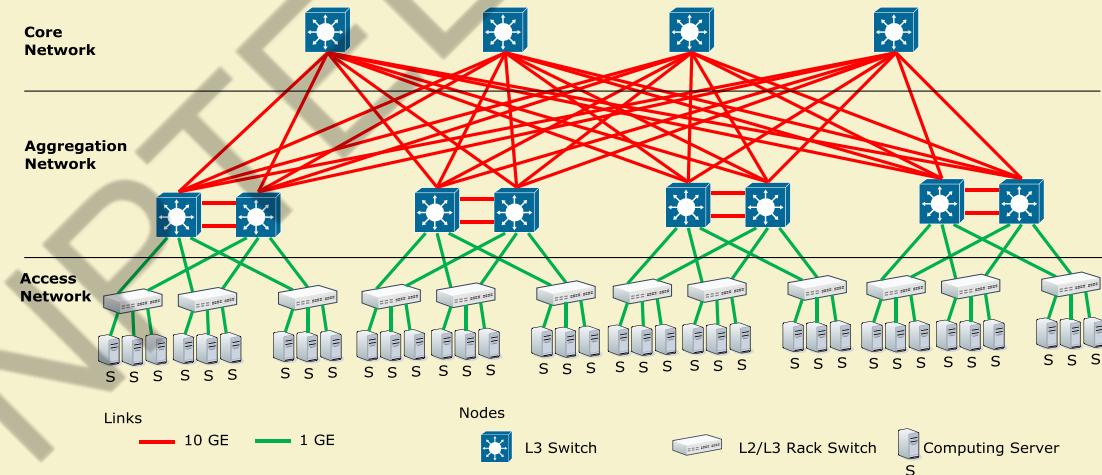


NPTEL ONLINE
CERTIFICATION COURSES

DC Architecture - Present

Three-tier DC architecture

- Most Widely Used Nowadays
- Access, Aggregation, and Core layers
- Scales to over 10,000 servers

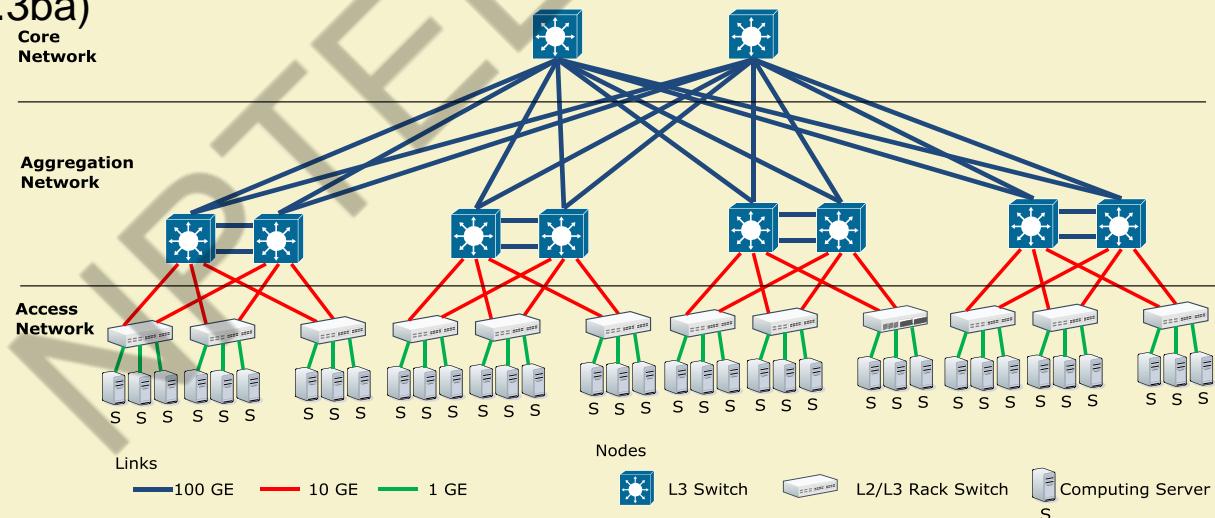


Ref: Dzmitry Kliazovich, University of Luxembourg

DC Architecture - Present

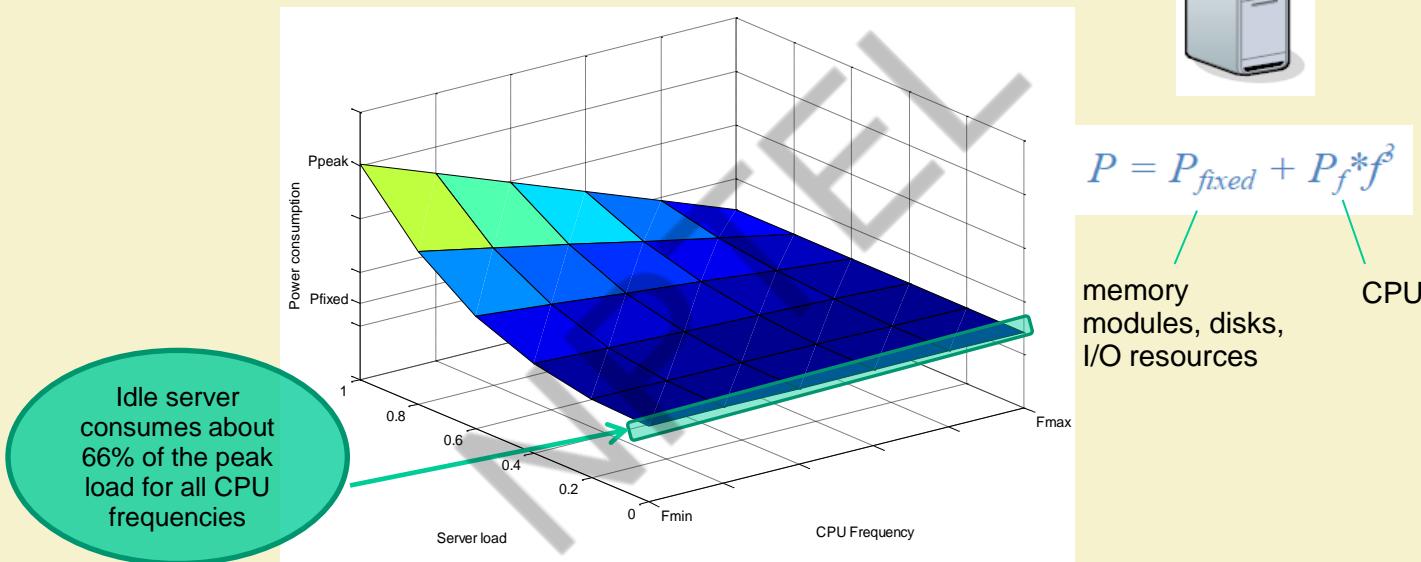
Three-tier High-Speed architecture

- Increased core network bandwidth
- 2-way ECMP load balancing
- 100 GE standard (IEEE 802.3ba)



Ref: Dzmitry Kliazovich, University of Luxembourg

DC Server Energy Model



Ref: Dzmitry Kliazovich, University of Luxembourg

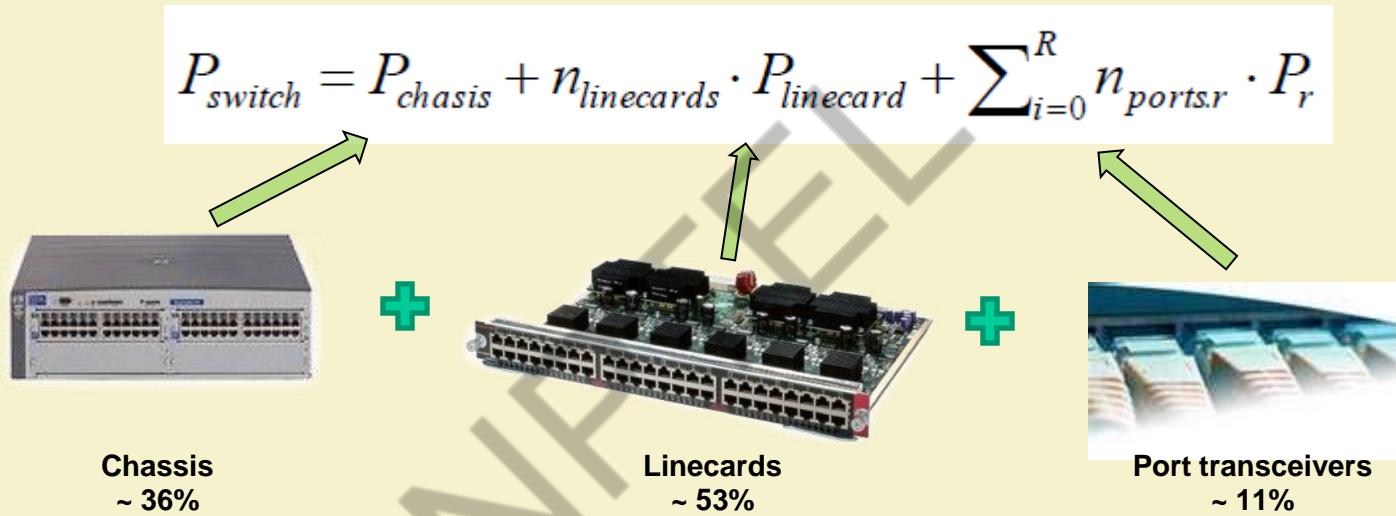


IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

DC Network Switches' Energy Model



Ref: Dzmitry Kliazovich, University of Luxembourg



IIT KHARAGPUR



NPTEL
ONLINE
CERTIFICATION COURSES

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



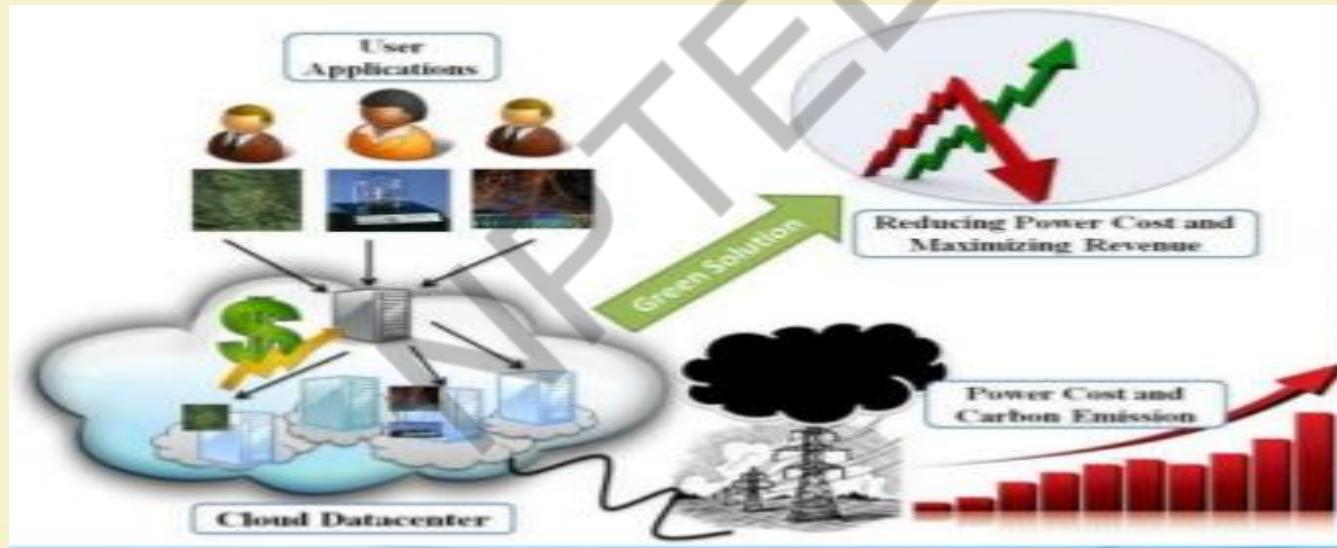
IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Performance <-> Energy Efficiency

As energy costs are increasing while availability decreases, there is a need to shift focus from optimizing data center resource management for pure performance alone to optimizing for energy efficiency while maintaining high service level performance.



Source: Internet



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

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

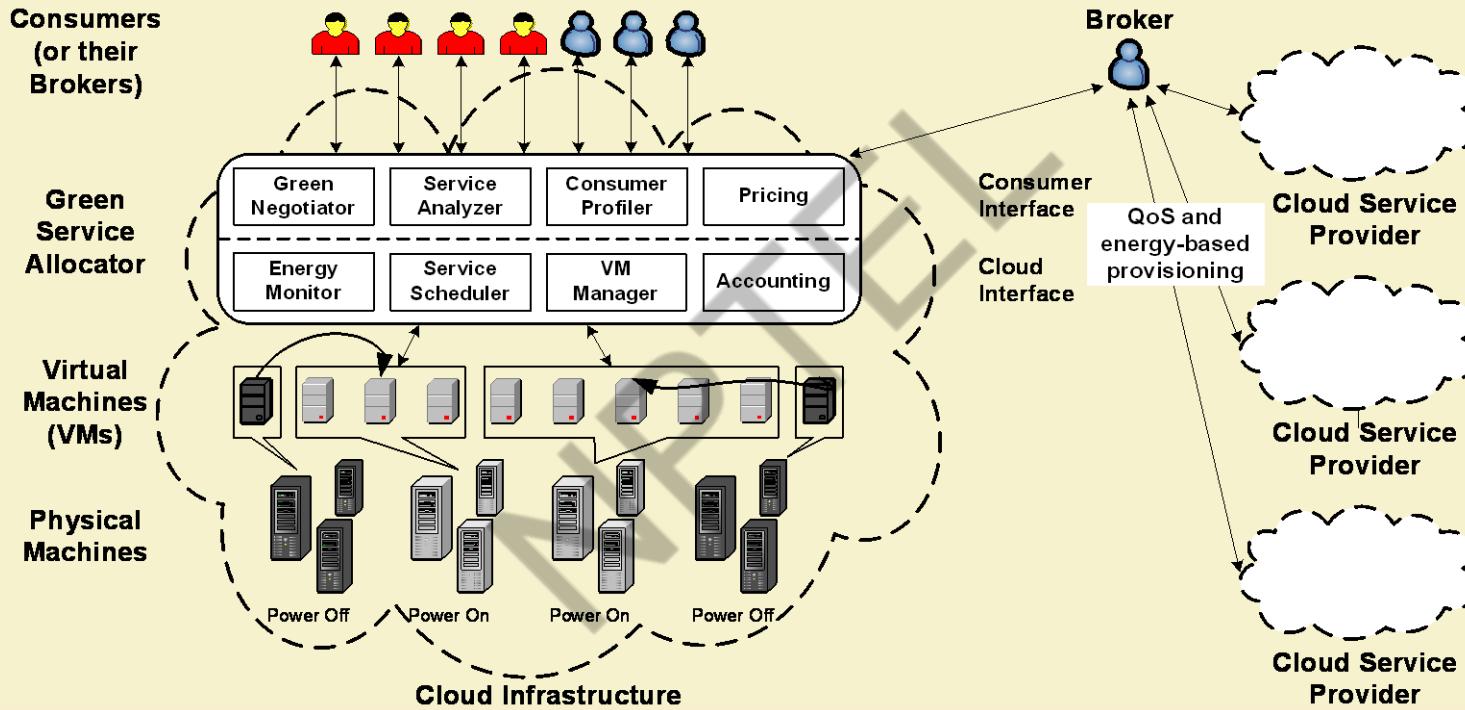


IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

A Typical Green Cloud Architecture



Source: Internet



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

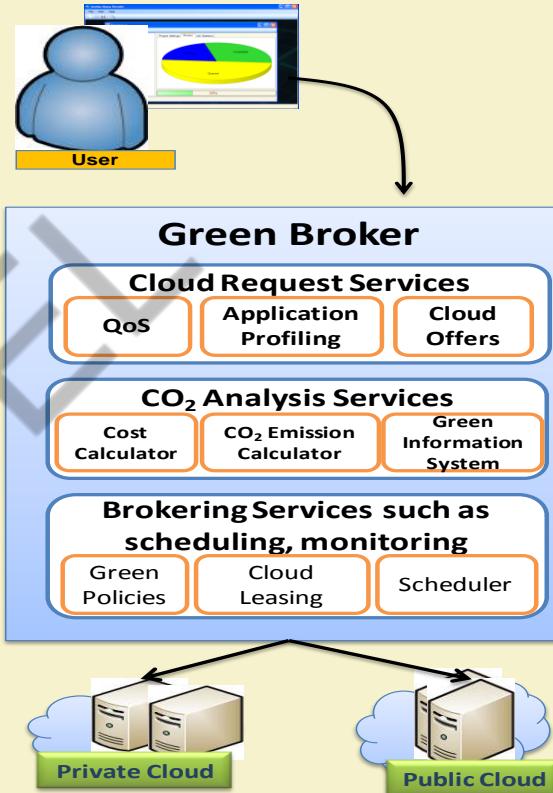
Green Broker

A typical Cloud broker

- Lease Cloud services
- Schedule applications

Green Broker

- 1st layer: Analyze user requirements
- 2nd layer: Calculates cost and carbon footprint of services
- 3rd layer: Carbon aware scheduling



Source: Internet

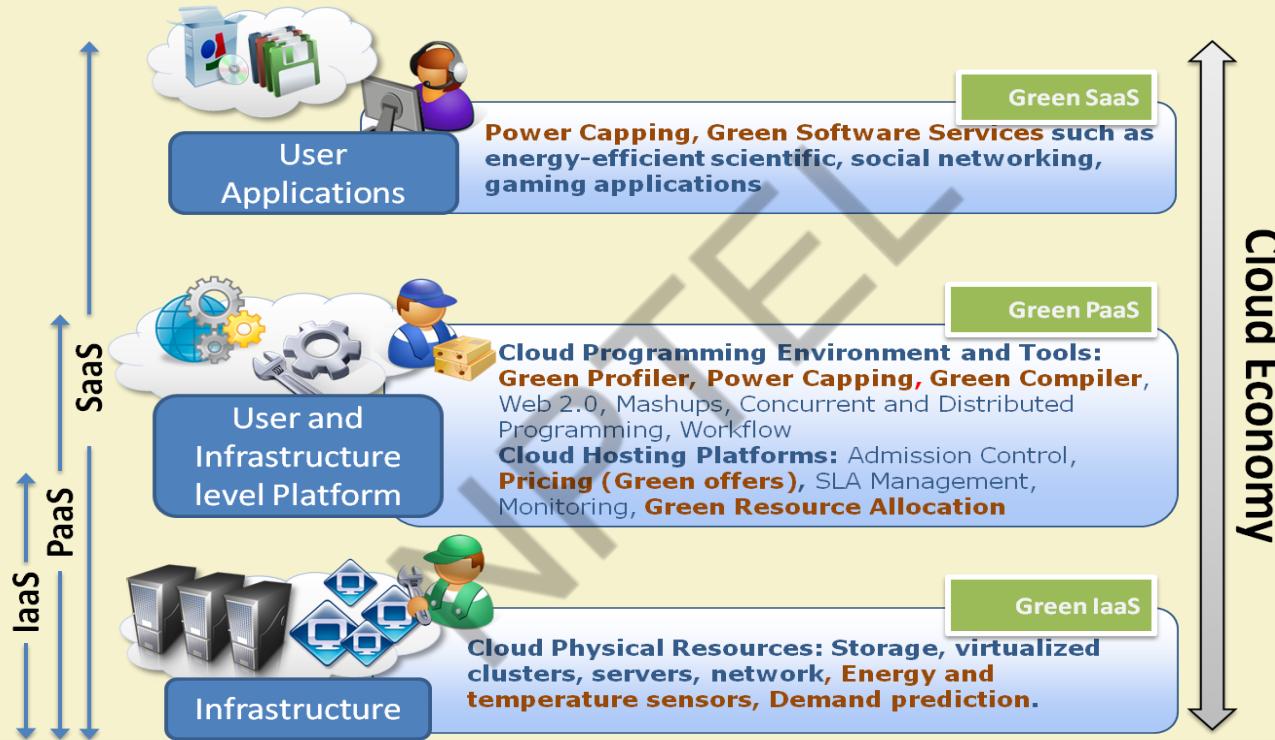


IIT KHARAGPUR



NPTEL
ONLINE
CERTIFICATION COURSES

Green Middleware



Source: Internet



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

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

Source: Internet



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Conclusions

- Clouds are essentially Data Centers hosting application services offered on a subscription basis. However, they consume high energy to maintain their operations.
=> high operational cost + environmental impact
- Presented a Carbon Aware Green Cloud Framework to improve the carbon footprint of Cloud computing.
- Open Issues: Lots of research to be carried out for Maximizing Efficiency of Green Data Centers and Developing Regions to benefit the most.

Source: Internet



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

Thank You!

NPTEL



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

CLOUD COMPUTING

Sensor Cloud Computing

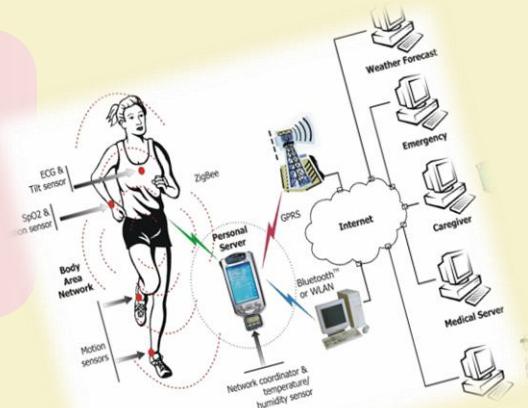
Prof. Soumya K Ghosh

Department of Computer Science and Engineering
IIT KHARAGPUR

Motivation



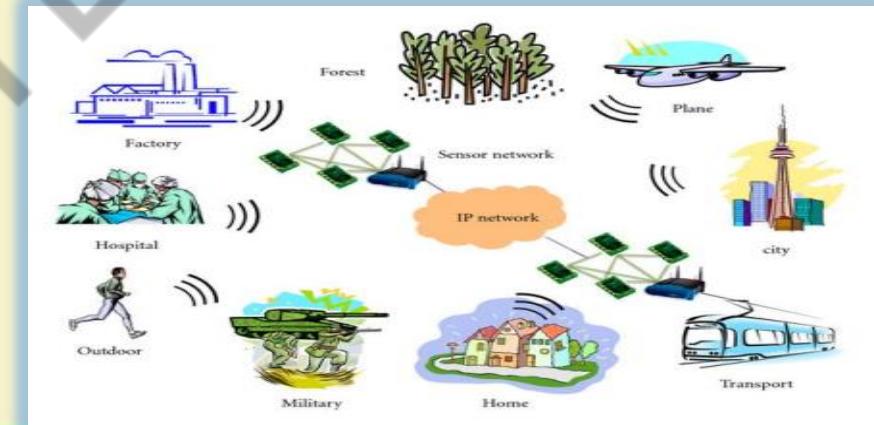
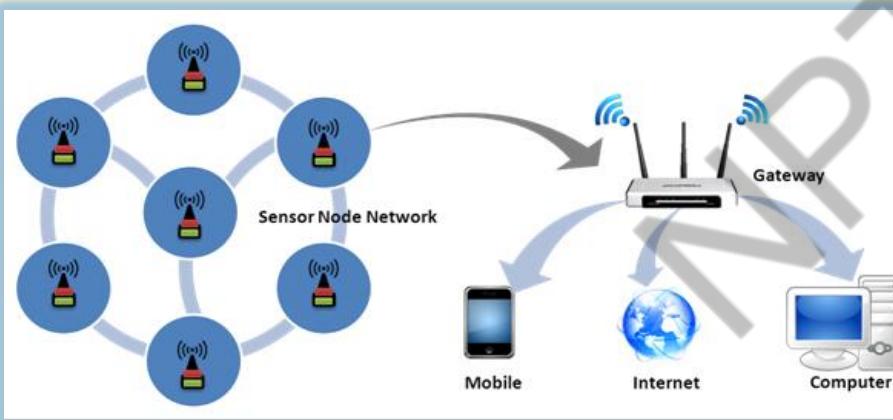
- Increasing adoption of sensing technologies (e.g., RFID, cameras, mobile phones)
- Internet has become a source of real time information (e.g., through blogs, social networks, live forums) for events happening around us



- 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

Wireless Sensor Network (WSNs)

- Seamlessly couples the physical environment with the digital world
- Sensor nodes are small, low power, low cost, and provide multiple functionalities
 - Sensing capability, processing power, memory, communication bandwidth, battery power.
- In aggregate, sensor nodes have substantial data acquisition and processing capability
- Useful in many application domains – Environment, Healthcare, Education, Defense, Manufacturing, Smart Home, etc.



IIT KHARAGPUR



NPTEL
ONLINE
CERTIFICATION COURSES

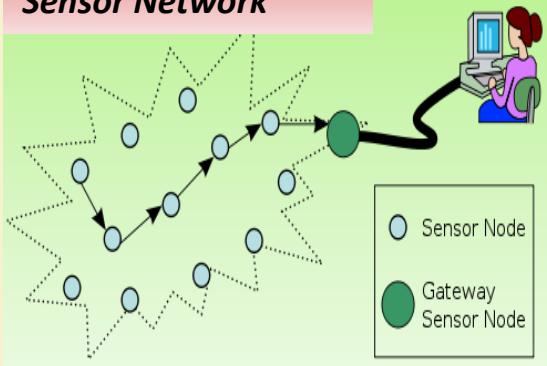
Limitations of Sensor Networks

- 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.

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

Sensor Network



What is missing?

Computing Platform

Applications

Cloud Storage

Social Networks

Codes

Cloud Server

Mobile Computing

Cloud Security



Cloud Economics

Services



IIT KHARAGPUR



NPTEL
ONLINE
CERTIFICATION COURSES

1. Lets go to the mountain peak!



A Motivating Scenario!



6. Your friend is at nearby restaurant.. Go catch up with her!



5. Menus of restaurants and recommended foods!



4. Take pictures of restaurants and send images



2. Sounds Good!

- I. Please take your lunch as you appear hungry!
- II. Carry drinking water – Water at that region is contaminated
- III. Use anti-UV skin cream

3. Map to nearest food outlets

Few insight from the example!

- Cell phone records the tourist's gestures and activates applications such as camera, microphone, etc.
- Cell phone produces very swift responses in real time after:
 - Processing geographical data
 - Acquiring tourist's physiological data from wearable physiological
 - Sensors (blood sugar, precipitation, etc.) and cross-comparing it with his medical records
 - Speech recognition
 - Image processing of restaurant's logos and accessing their internet-based profiles
 - Accessing tourist's social network profiles to find out his friends

Fact : the cell phone cannot perform so much tasks !

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.

What is Sensor Cloud Computing?

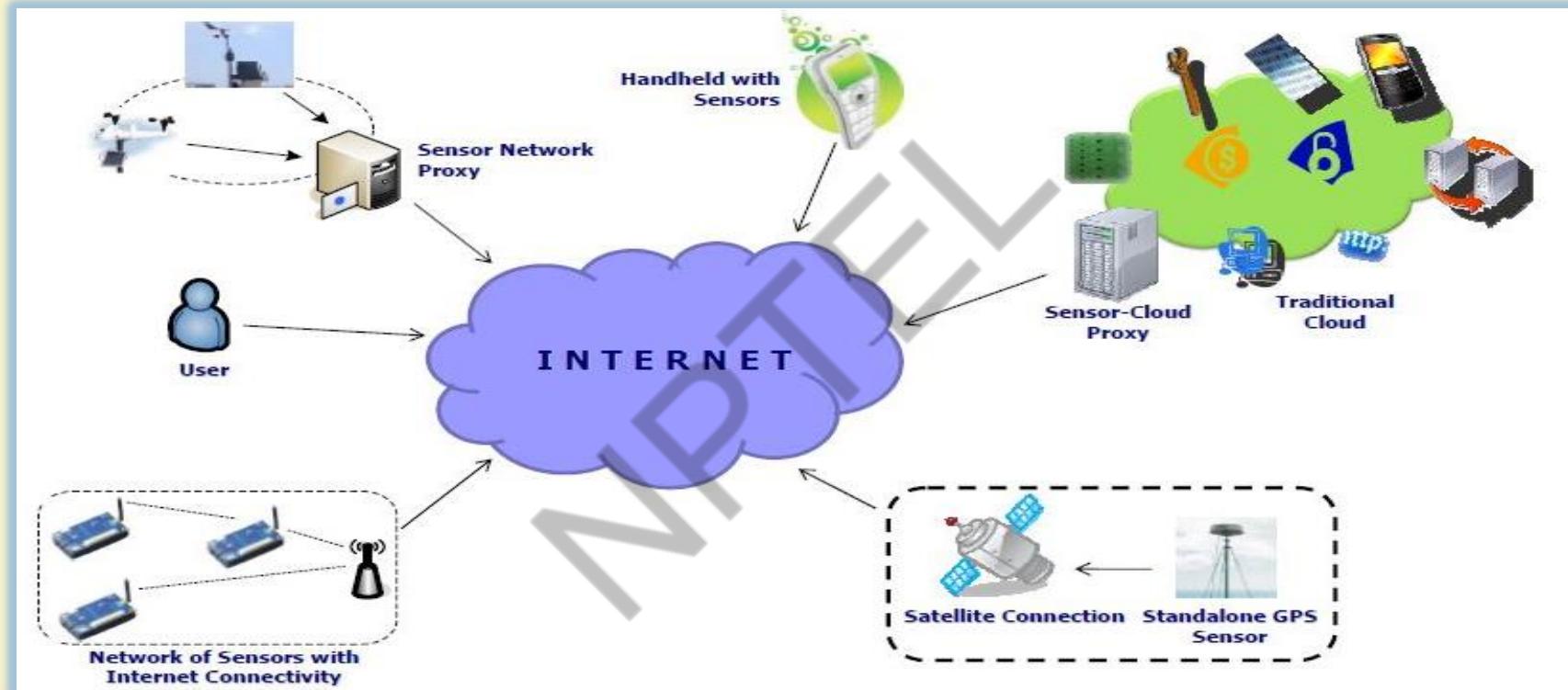
An infrastructure that allows truly pervasive computation using sensors as interface between physical and cyber worlds, the data-compute clusters as the cyber backbone and the internet as the communication medium

- 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.

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 processing.
- Various sensor nodes spread in a huge geographical area, to connect together and be employed 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.

Overview of Sensor-Cloud Framework



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Overview of Sensor-Cloud Framework

Sensor-Cloud Proxy

- Interface between sensor resources and the cloud fabric.
- Manages sensor network connectivity between the sensor resources and the cloud.
- Exposes sensor resources as cloud services.
- Manages sensor resources via indexing services.
- Uses cloud discovery services for resource tracking.
- Manages sensing jobs for programmable sensor networks.
- Manages data from sensor networks
 - Data format conversion into standard formats (e.g. XML)
 - Data cleaning and aggregation to improve data quality
 - Data transfer to cloud storage
- Sensor-cloud proxy can be virtualized and lives on the cloud !

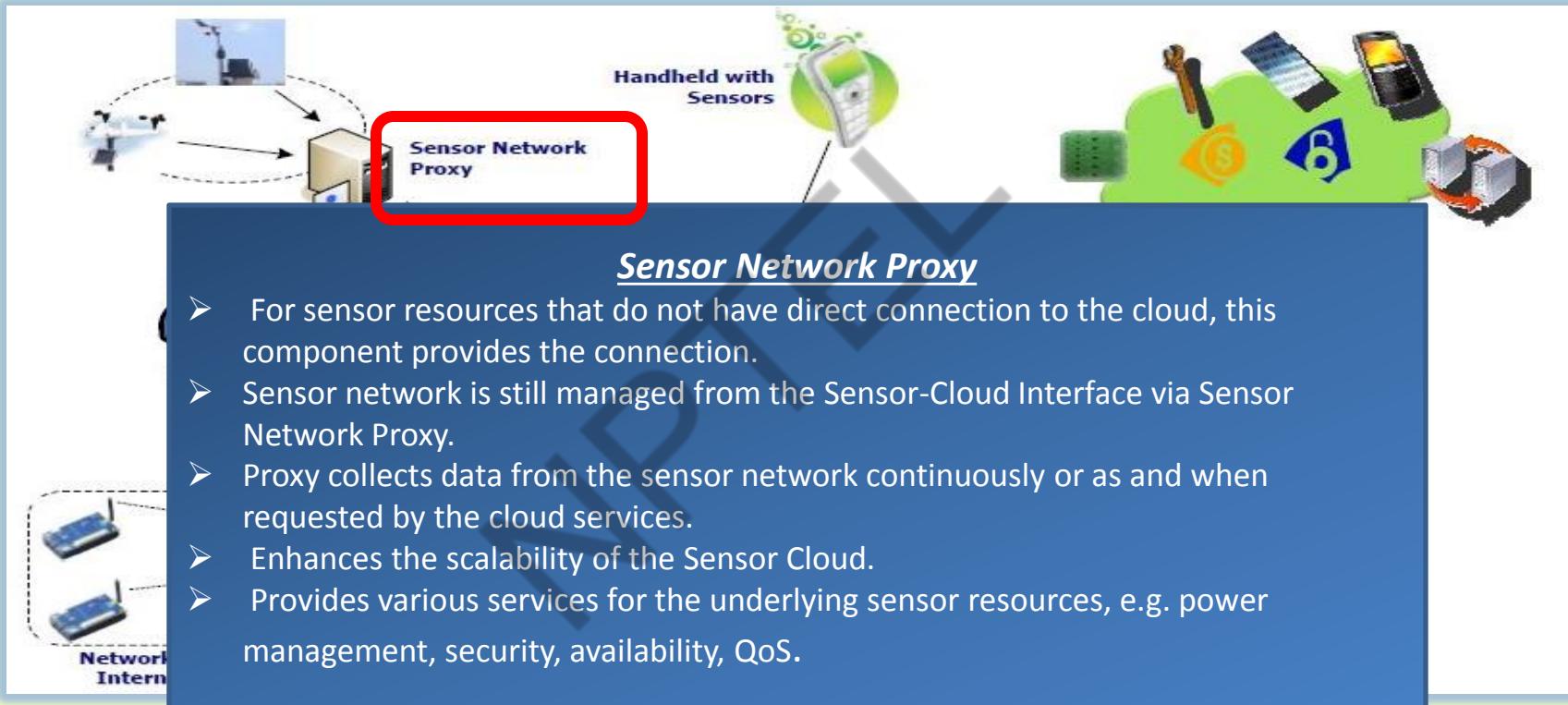


IIT KHARAGPUR



NPTEL
ONLINE
CERTIFICATION COURSES

Overview of Sensor-Cloud Framework



IIT KHARAGPUR

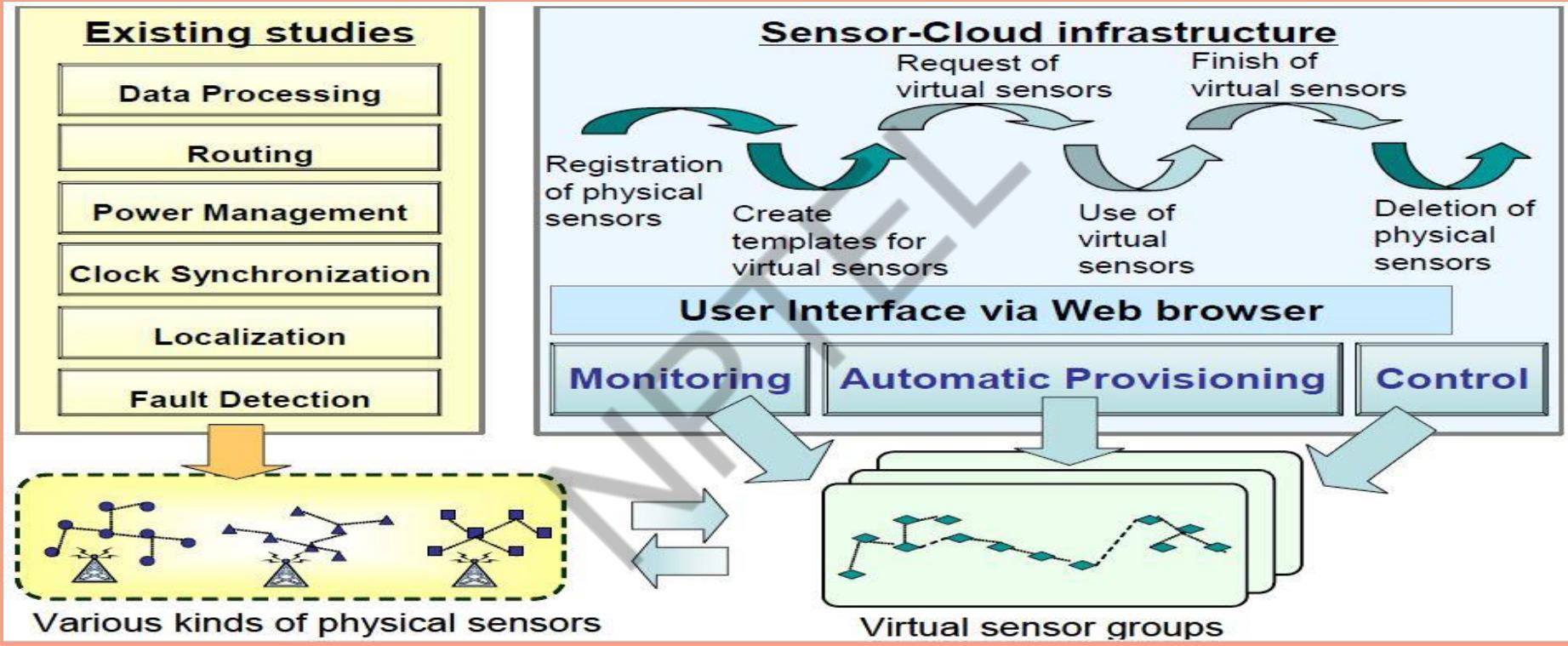


NPTEL ONLINE
CERTIFICATION COURSES

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.

Overview of Sensor Cloud Infrastructure



IIT KHARAGPUR

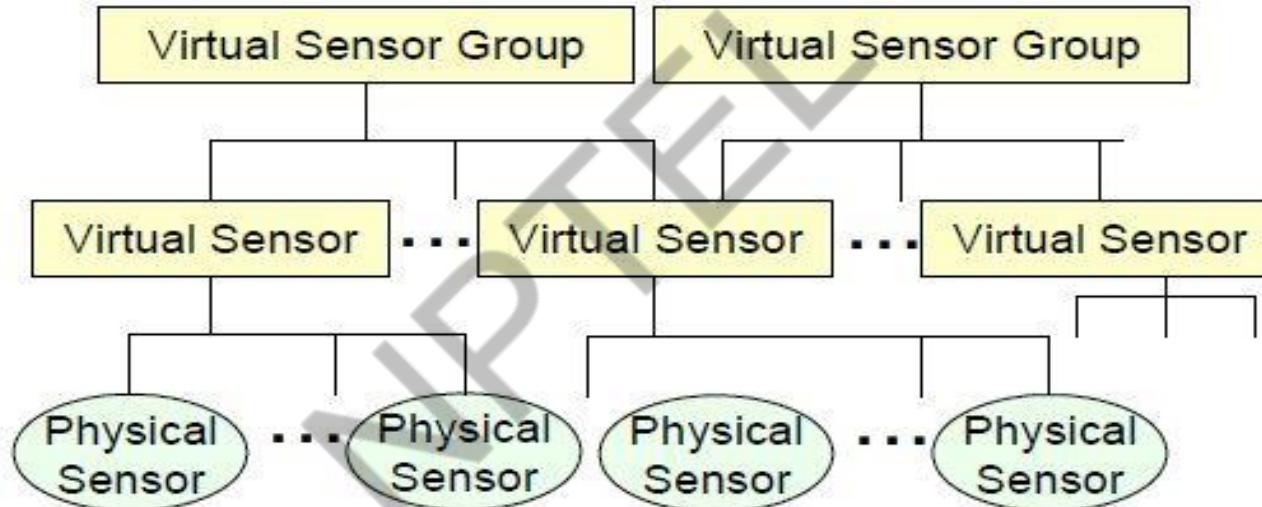


NPTEL
ONLINE
CERTIFICATION COURSES

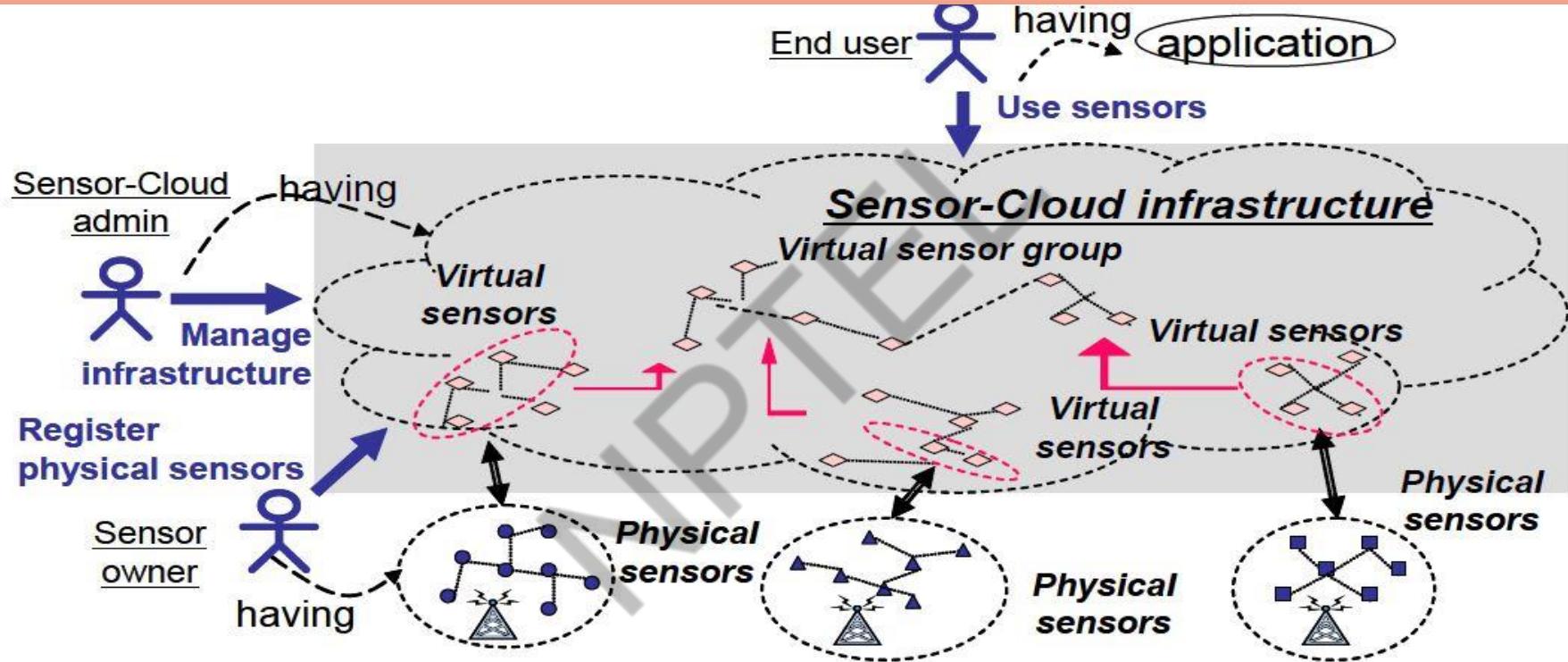
Madoka et al. "Sensor-Cloud Infrastructure Physical Sensor Management with Virtualized Sensors on Cloud Computing"

Virtual Sensors?

- A virtual sensor is an emulation of a physical sensor that obtains its data from underlying physical sensors.
- Virtual sensors are used to overcome the limitation of physical sensors in terms of location tracking, time and cost.
- In wireless sensor networks, virtual sensors are used to reduce the complexity of the system.
- To overcome the limitation of physical sensors, virtual sensors are used.
- The virtual sensors contain metadata about the physical sensors and the user currently holding that virtual sensor.



Relationship among Actors and Sensor Cloud Infrastructure



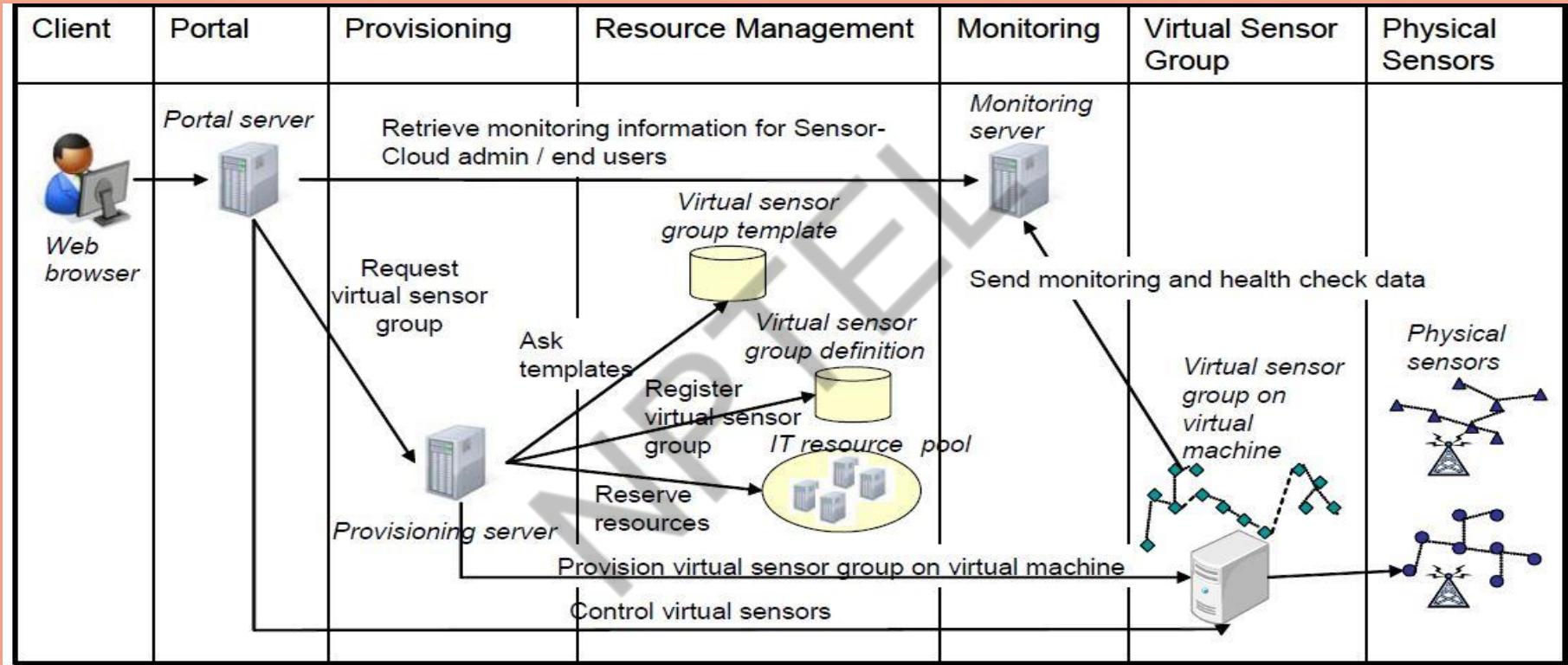
IIT KHARAGPUR



NPTEL
ONLINE
CERTIFICATION COURSES

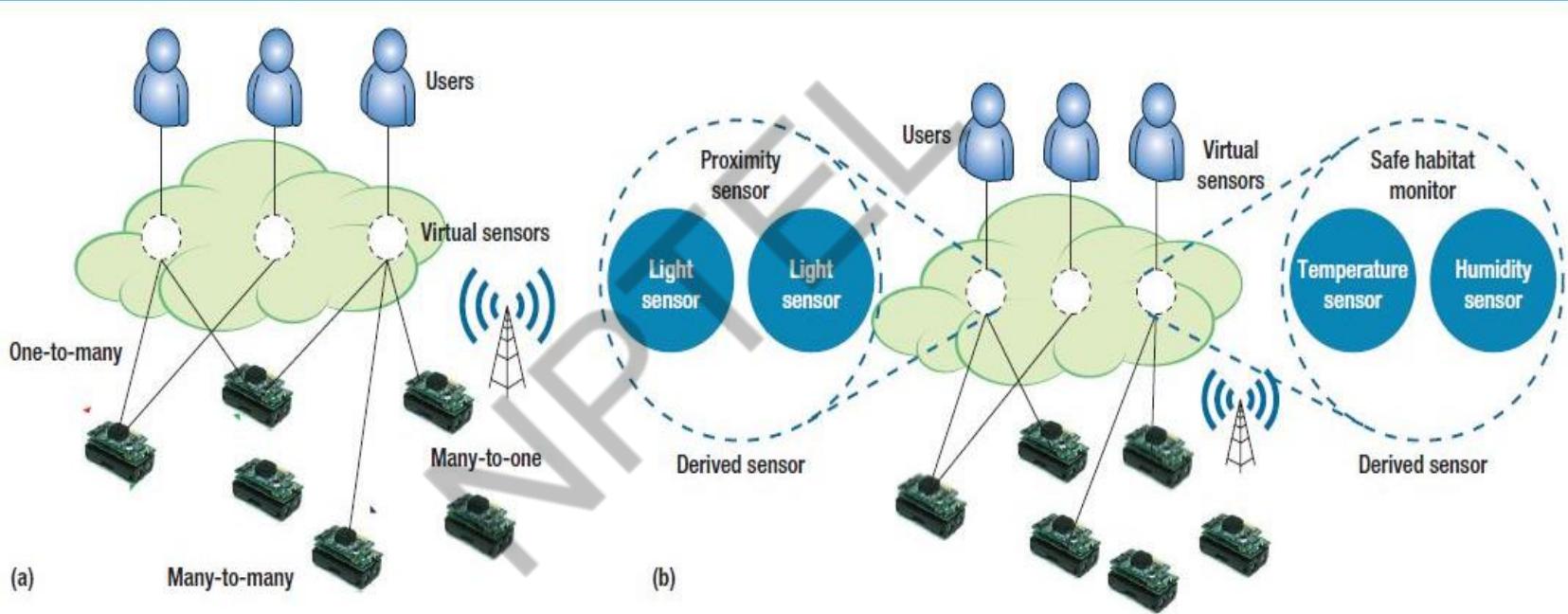
Madoka et al. "Sensor-Cloud Infrastructure Physical Sensor Management with Virtualized Sensors on Cloud Computing"

System Architecture of Sensor Cloud Infrastructure



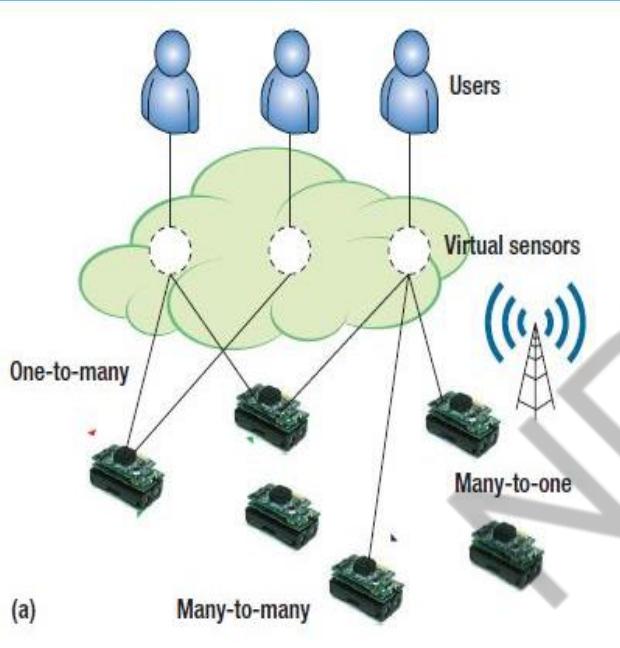
Virtual Sensor Configurations

(a) one-to-many, many-to-one, and many-to-many, and (b) derived



Virtual Sensor Configurations

(a) one-to-many, many-to-one, and many-to-many, and (b) derived



One to Many Configurations:

- In this configuration, one physical sensor corresponds to many virtual sensors.
- Although individual users own the virtual image, the underlying physical sensor is shared among all the virtual sensors accessing it.
- The middleware computes the physical sensor's sampling duration and frequency by taking into account all the users; it re-evaluates the duration and frequency when new users join or existing users leave the system.



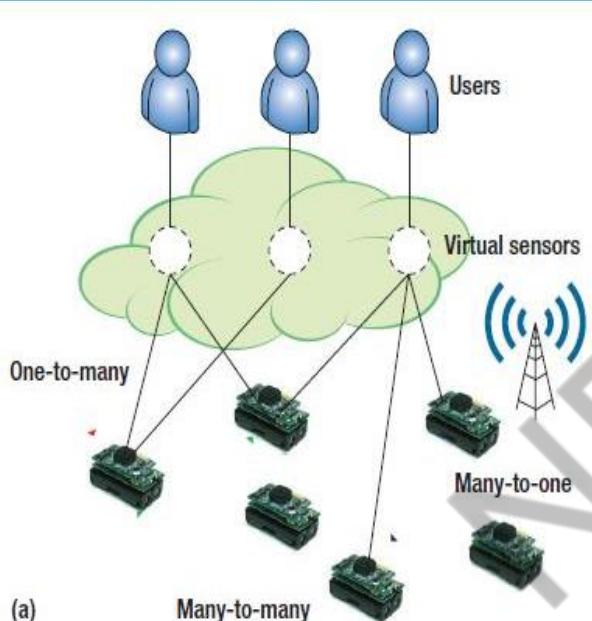
IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Virtual Sensor Configurations

(a) one-to-many, many-to-one, and many-to-many, and (b) derived



Many to One Configurations:

- In this configuration, the geographical area is divided into regions and each region can have one or more physical sensors and sensor networks.
- When a user requires aggregated data of specific phenomena from a region, all underlying WSNs switch on with the respective phenomena enabled, and the user has access to the aggregated data from these WSNs



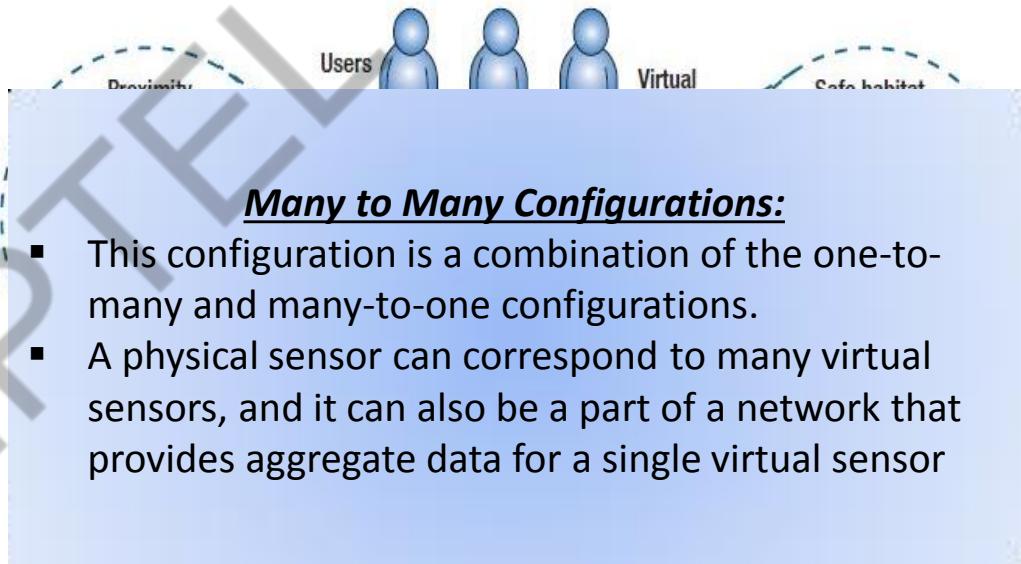
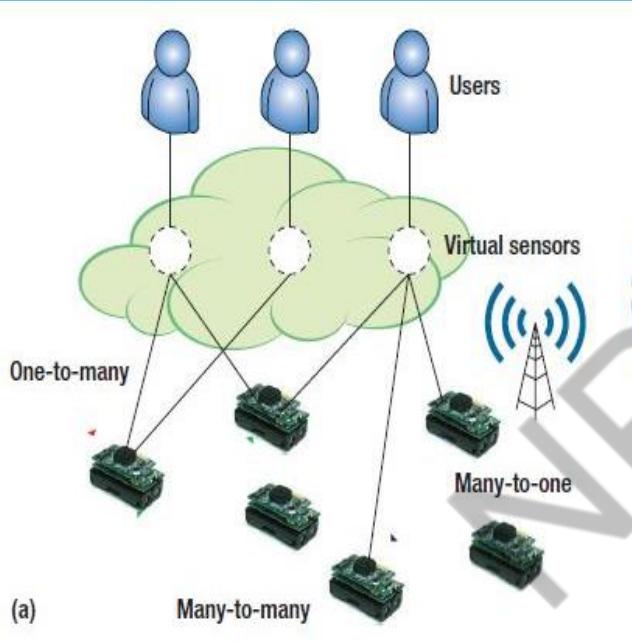
IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Virtual Sensor Configurations

(a) one-to-many, many-to-one, and many-to-many, and (b) derived

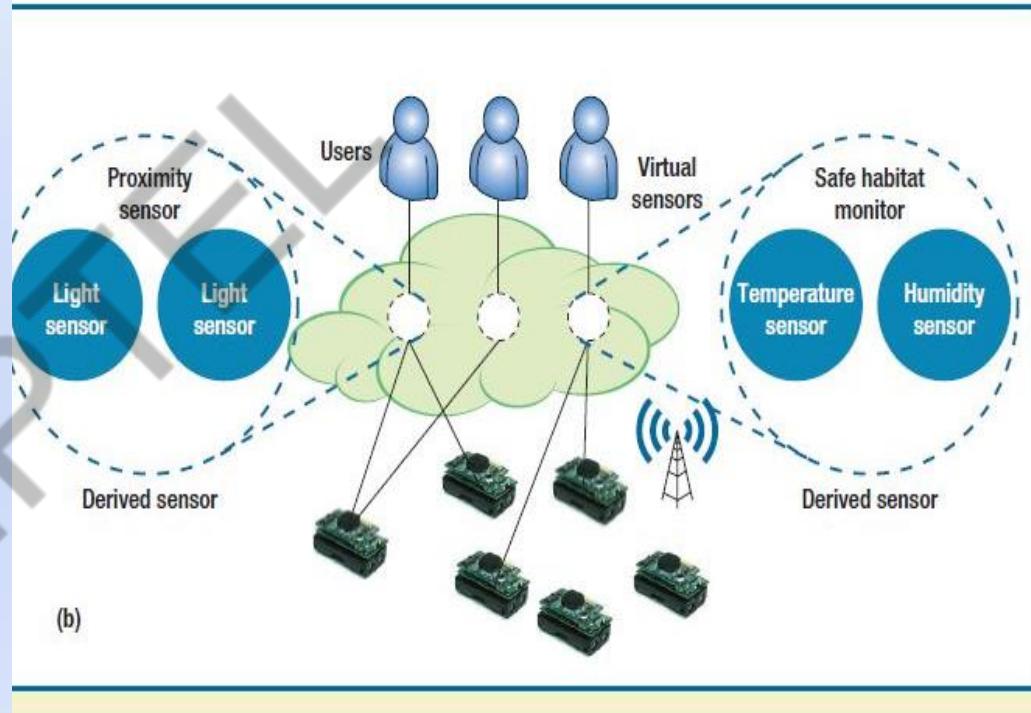


Virtual Sensor Configurations

(a) one-to-many, many-to-one, and many-to-many, and (b) derived

Derived:

- A derived configuration refers to a versatile configuration of virtual sensors derived from a combination of multiple physical sensors.
- This configuration can be seen as a generalization of the other three configurations, though, the difference lies in the types of physical sensors with which a virtual sensor communicates.
- While in the derived configuration, the virtual sensor communicates with multiple sensor types; in the other three configurations, the virtual sensor communicates with the same type of physical sensors.
- Derived sensors can be used in two ways: first, to virtually sense complex phenomenon and second, to substitute for sensors that aren't physically deployed.



(b)

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.

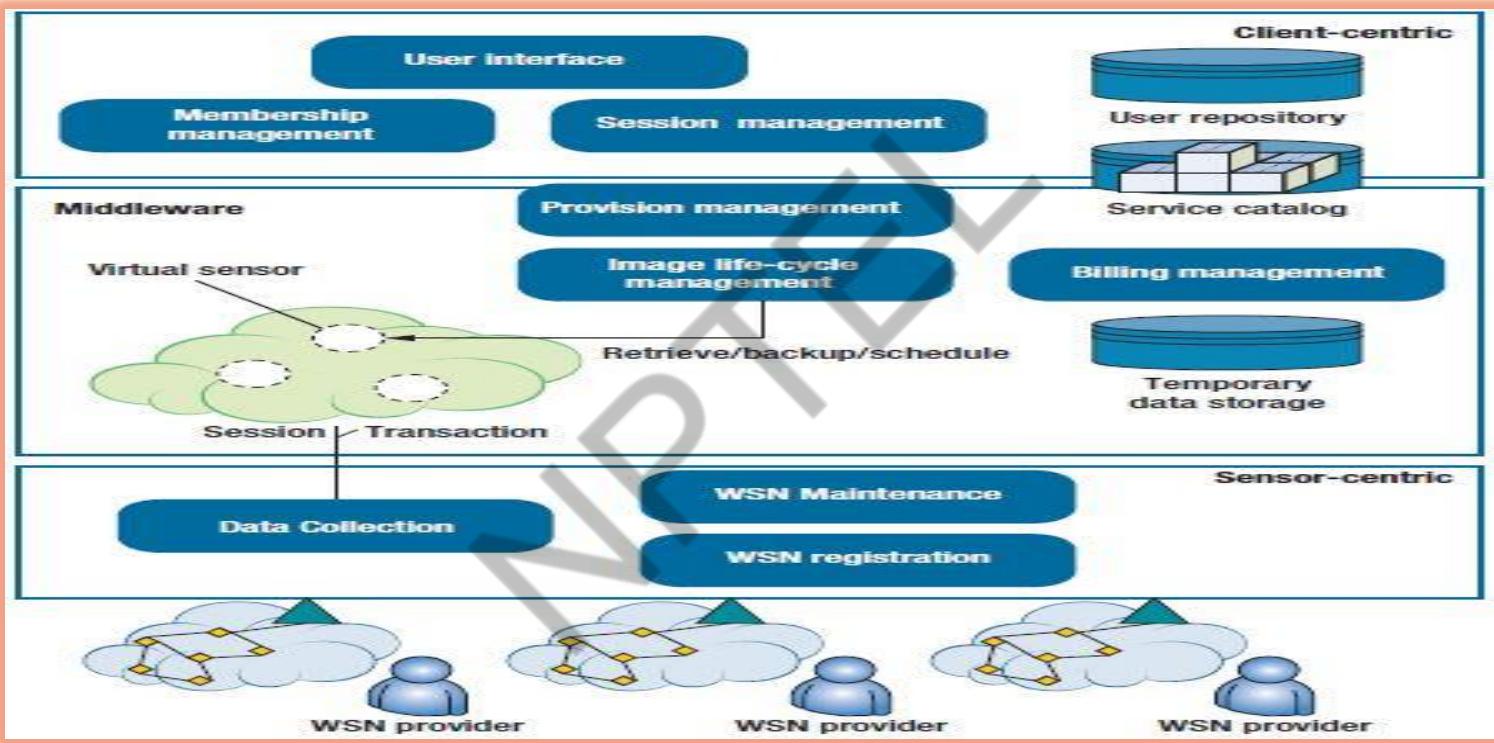


IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

A Layered Sensor Cloud Architecture



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

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.

References

- Beng, Lim Hock. "Sensor cloud: Towards sensor-enabled cloud services." *Intelligent Systems Center Nanyang Technological University* (2009)
- <http://www.ntu.edu.sg/intellisys>
- Sanjay et al. "Sensor Cloud: A Cloud of Virtual Sensors" , *IEEE Software*, 2014
- **Madoka et al.** "**Sensor-Cloud Infrastructure** Physical Sensor Management with Virtualized Sensors on Cloud Computing"

Thank You!



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

CLOUD COMPUTING

IoT Cloud

Prof. Soumya K Ghosh

Department of Computer Science and Engineering
IIT KHARAGPUR

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



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 technologies are often described under the umbrella term of the Internet of Things (IoT)”
 - th unique
to-human
 - 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*



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Source: Internet

More “*Things*” are being connected!

- Home/daily-life devices
- Business
- Public infrastructure
- Health-care and so on...

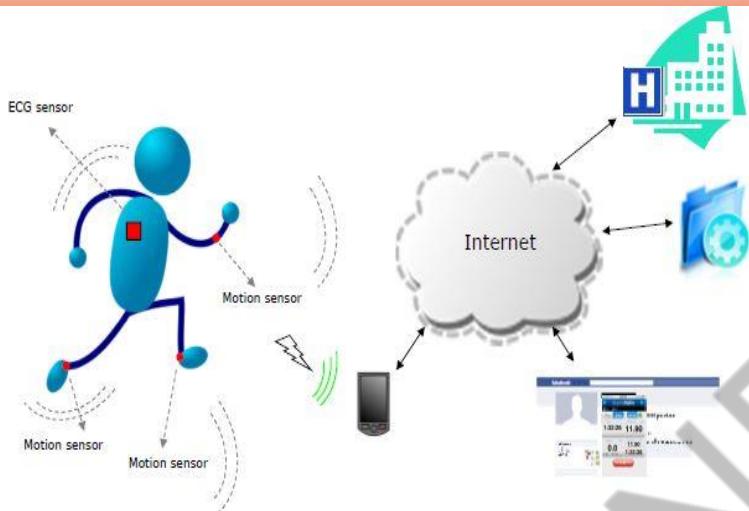


IIT KHARAGPUR

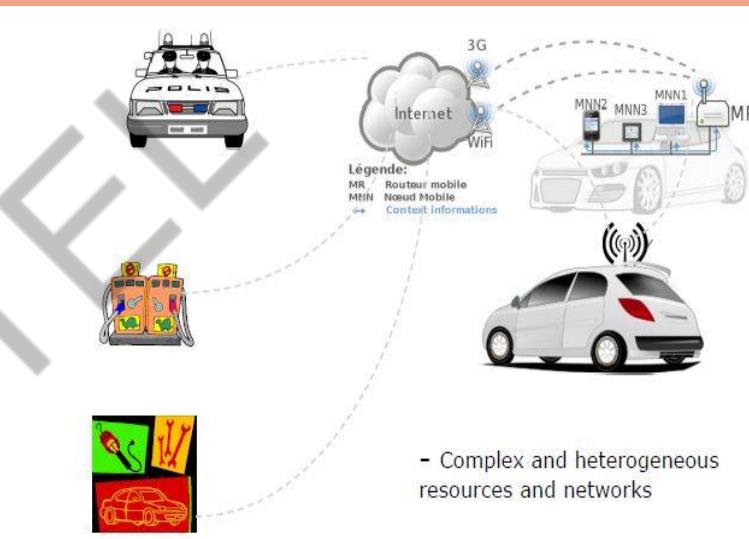


NPTEL
ONLINE
CERTIFICATION COURSES

Any time, Any place connectivity for Anyone and Anything!

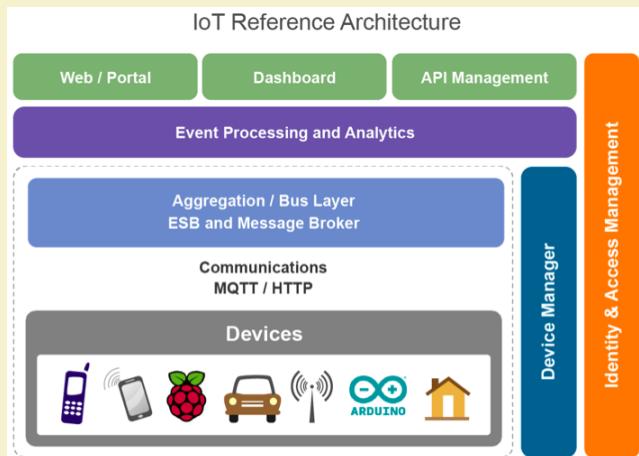


“People” Connecting to “Things”!



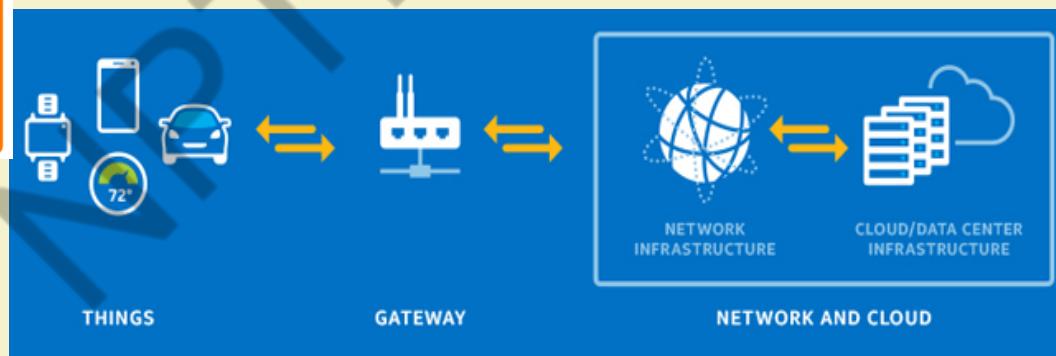
- Complex and heterogeneous resources and networks

Basic IoT Architecture



An IoT platform has basically three building blocks

- Things
- Gateway
- Network and Cloud



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.

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



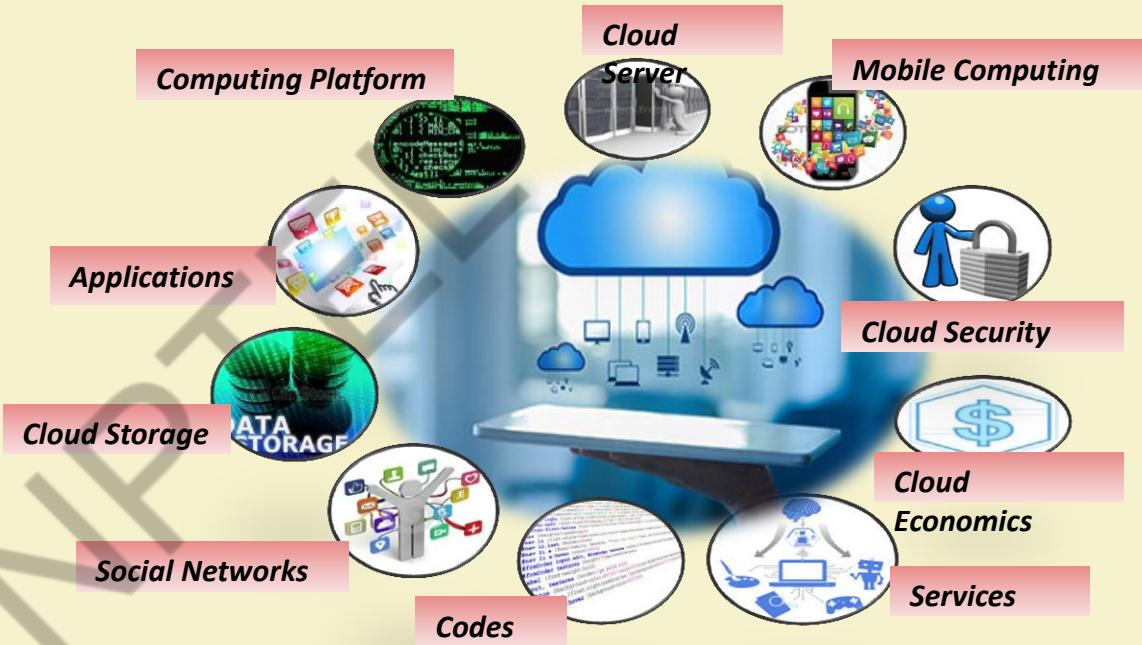
IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

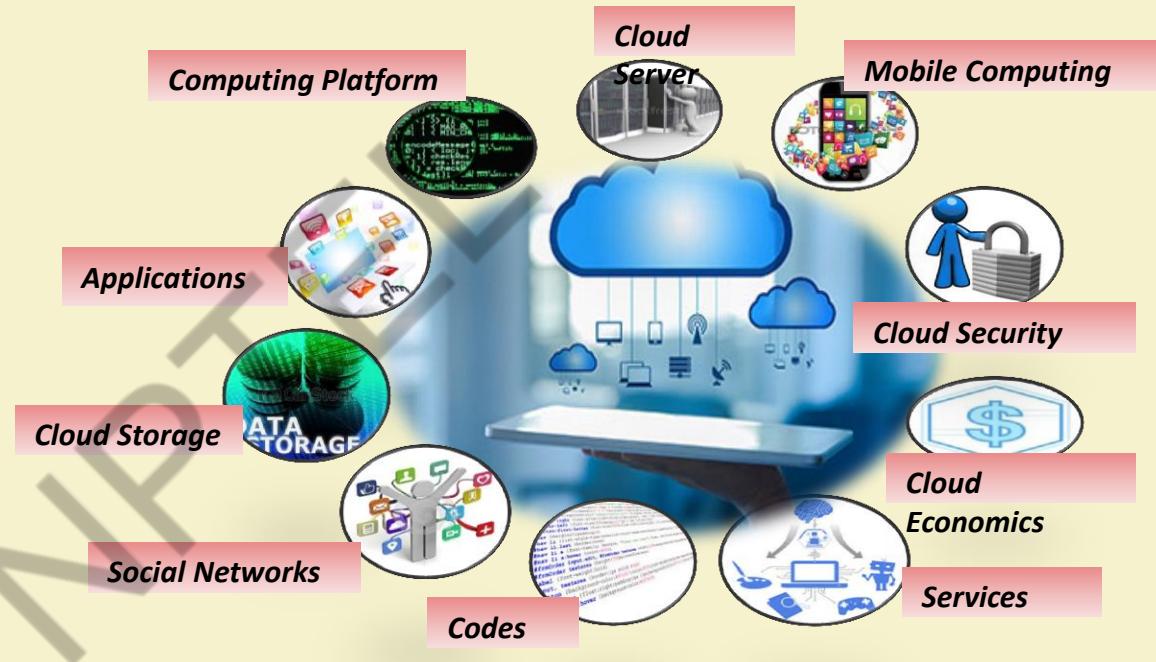
Cloud Computing!

- Cloud computing enables companies and applications, which are system infrastructure dependent, to be infrastructure-less.
- Cloud infrastructure offers “**pay-as-used and on-demand**” services
- Clients can offload their data and applications on cloud for storage and processing



Cloud Computing!

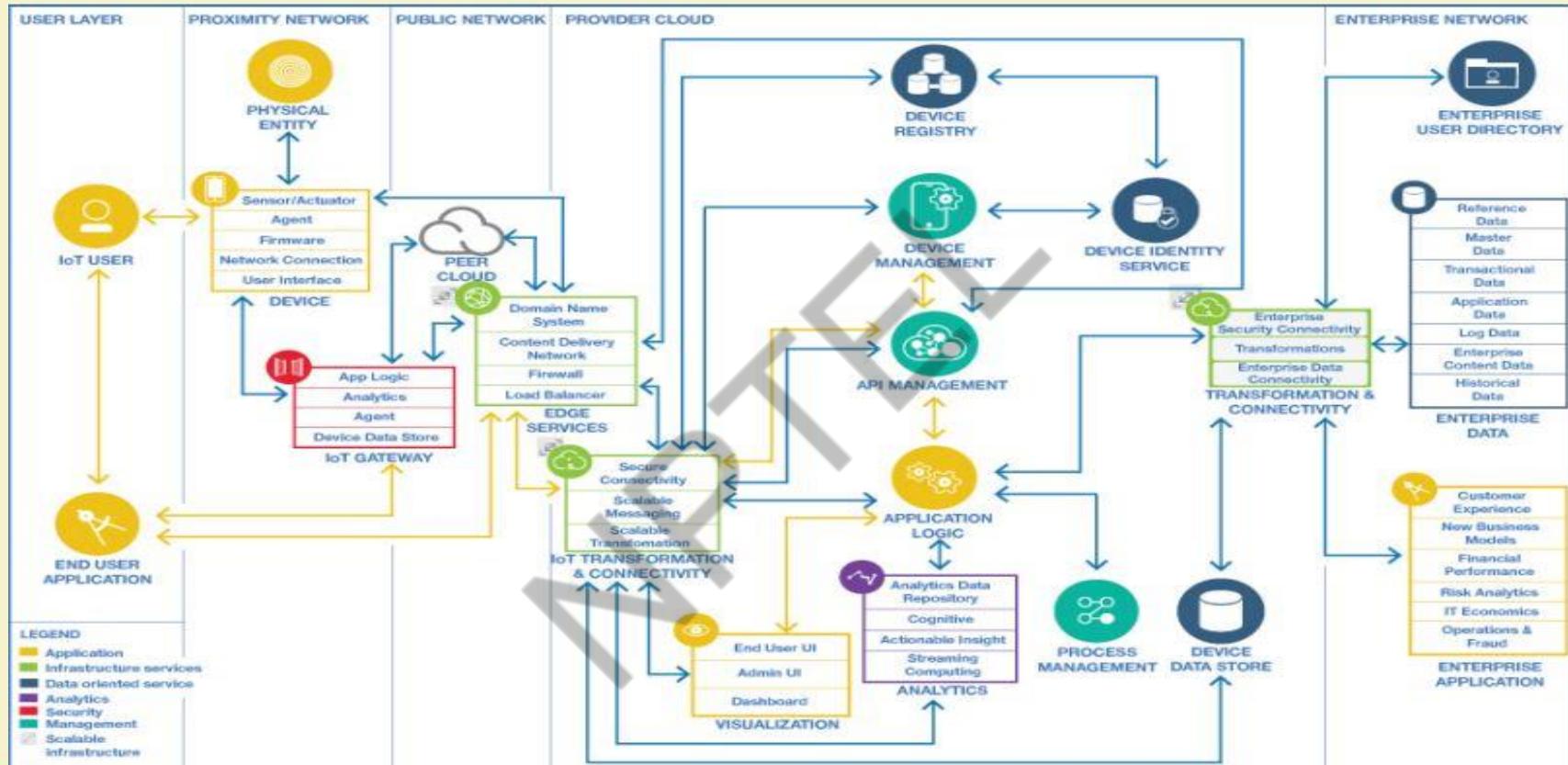
- It enables services to be used without any understanding of the infrastructure.
- Cloud computing works using economies of scale
- Data and services are stored remotely but accessible from “anywhere”.



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.
- An 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
- The 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

Cloud Components for IoT

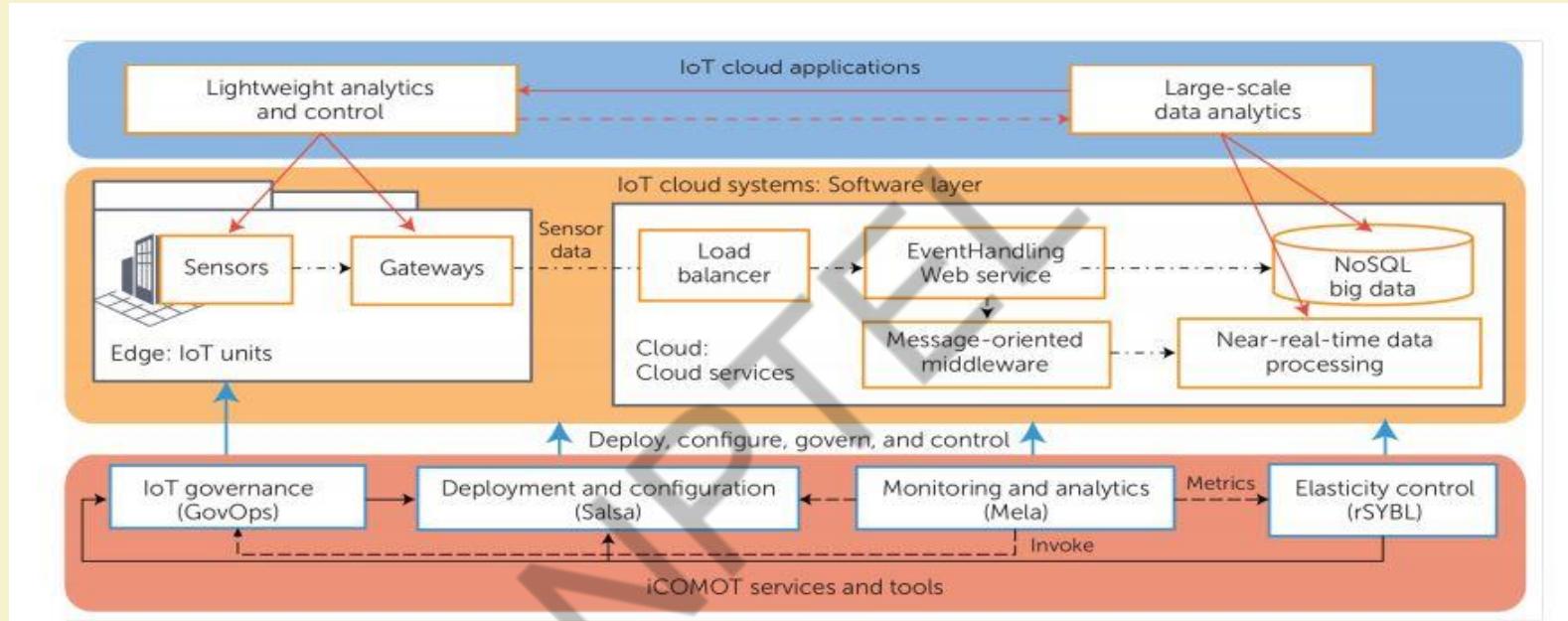


IIT KHARAGPUR



NPTEL
ONLINE
CERTIFICATION COURSES

iCOMOT: An IoT Cloud System



Top layer represents typical IoT applications executed across IoT and Clouds.

Middle layer represents the software layer as an IoT cloud system built on top of various types of cloud services and IoT elements.

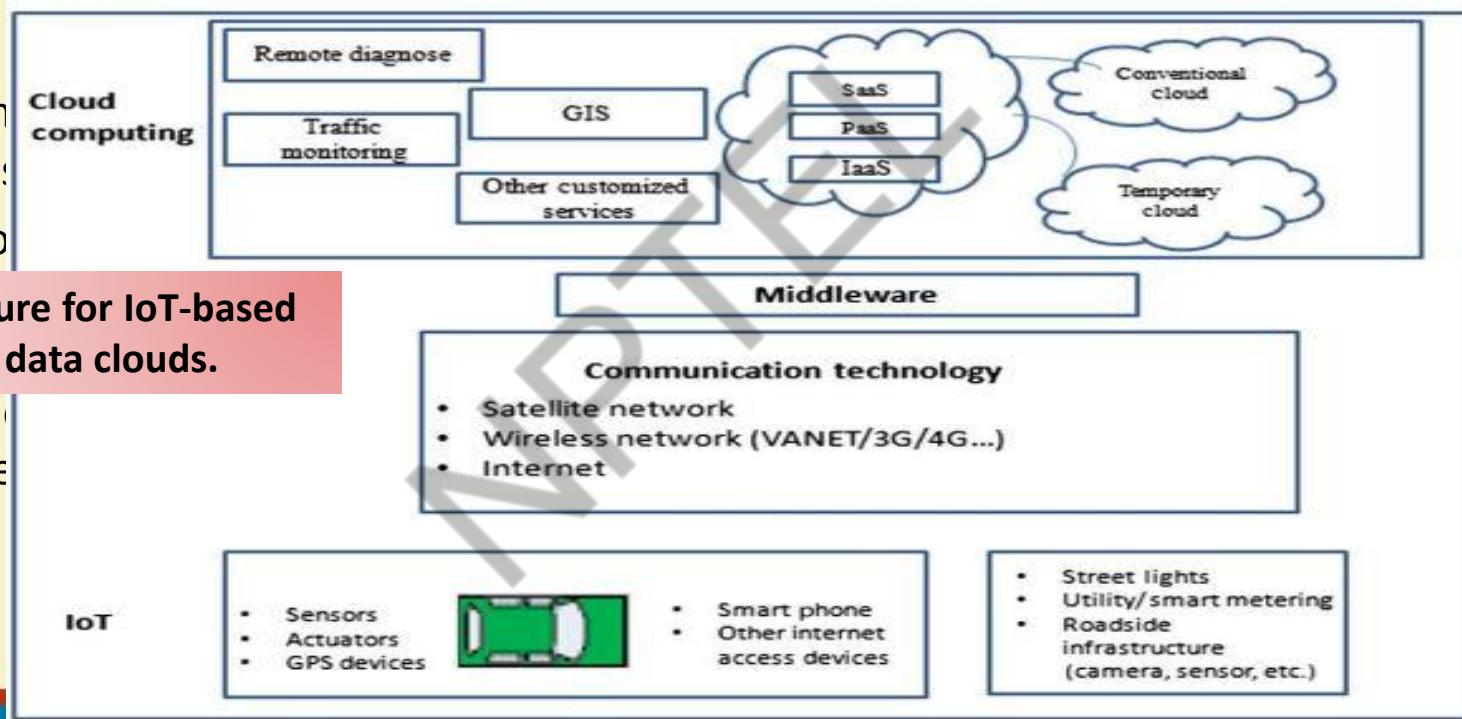
Bottom layer shows different tools and services from iCOMOT that can be used to monitor, control, and configure the software layer.

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

Motivating example: *Developing Vehicular Data Cloud Services in the IoT Environment*

- The pronounced trend of transitioning to cloud computing
- A no



IIT KHARAGPUR



NPTEL
ONLINE
CERTIFICATION COURSES

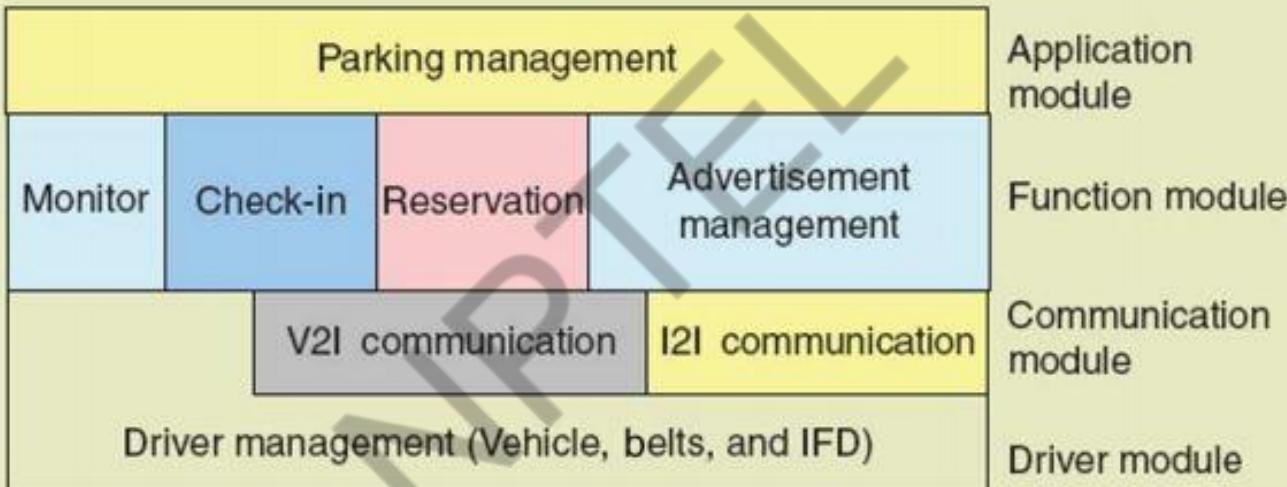
He, Wu, Gongjun Yan, and Li Da Xu. "Developing vehicular data cloud services in the IoT environment." *IEEE Transactions on Industrial Informatics* 10.2 (2014): 1587-1595.

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



Architecture for Intelligent Parking Cloud service

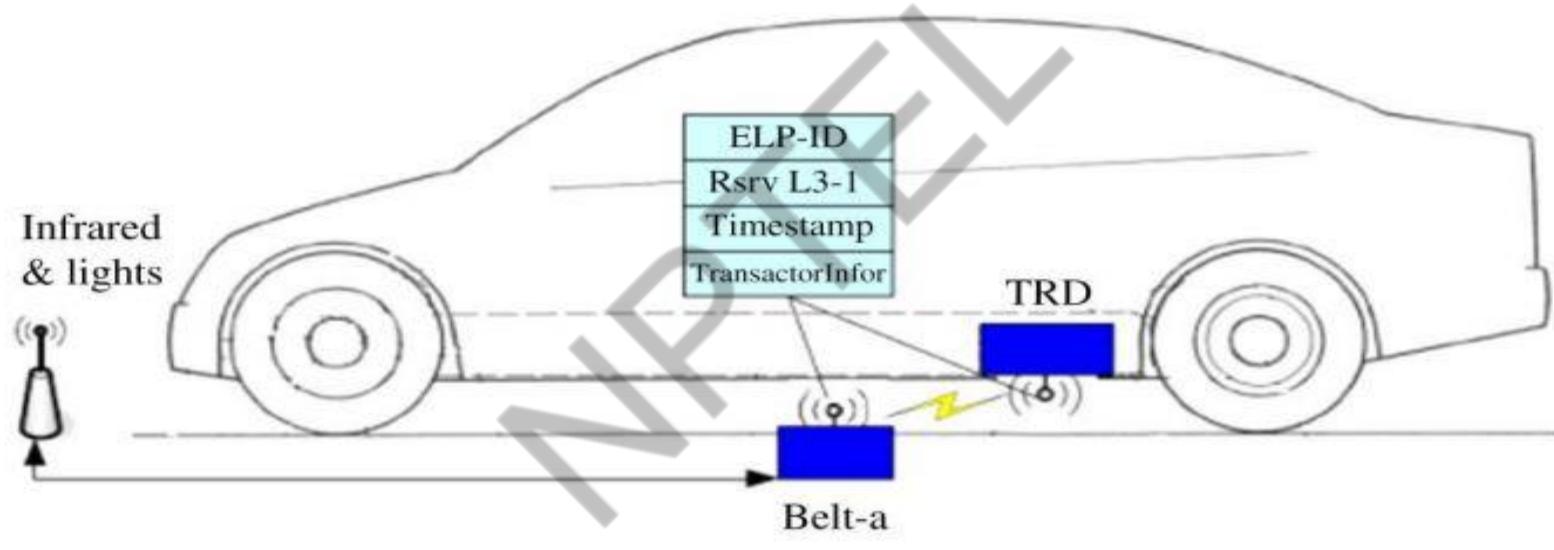


IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Vacancy detections by Sensors

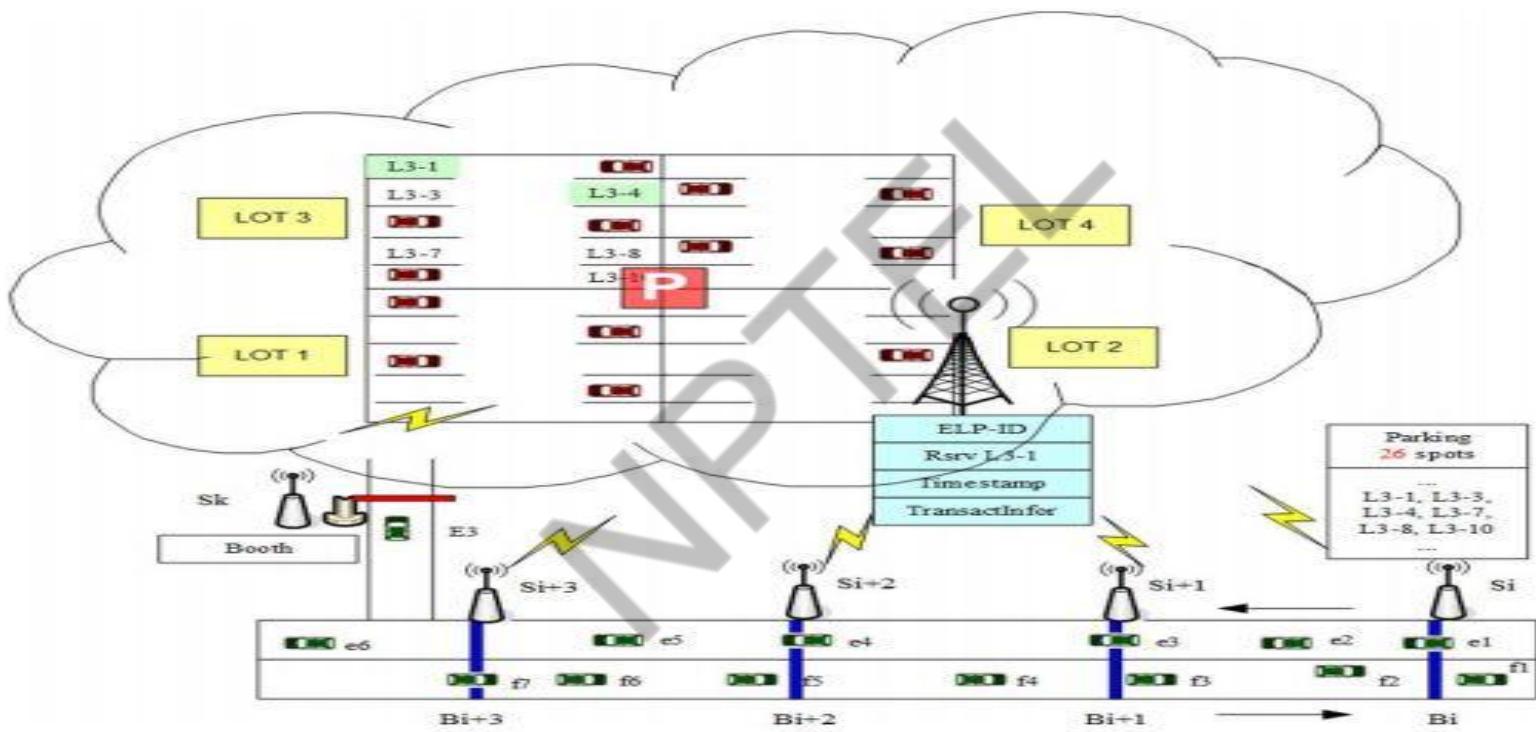


IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES

Parking cloud service



IIT KHARAGPUR



NPTEL
ONLINE
CERTIFICATION COURSES

Summary

- Internet of Things (IoT) is a dynamic and exciting area of IT. Many IoT systems are going to be created over the next few years, covering wide variety of areas, like domestic, commercial, industrial, health and government contexts
- IoT systems have several challenges, namely scale, speed, safety, security and privacy
- 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 Cloud Platform is an enabling paradigm to realize variety of services

References

- Cloud Standards Customer Council 2015, Cloud Customer Architecture for Big Data and Analytics, Version 1.1 <http://www.cloud-council.org/deliverables/CSCC-Customer-Cloud-Architecture-for-Big-Data-andAnalytics.pdf>
- He, Wu, Gongjun Yan, and Li Da Xu. "Developing vehicular data cloud services in the IoT environment." *IEEE Transactions on Industrial Informatics* 10.2 (2014): 1587-1595.
- H.-L. Truong et al., "iCOMOT: Toolset for Managing IoT Cloud Systems," demo, 16th IEEE Int'l Conf. Mobile Data Management, 2015
- Truong, Hong-Linh, and Schahram Dustdar. Principles for engineering IoT cloud systems." *IEEE Cloud Computing* 2.2 (2015): 68-76

Thank You!

NPTEL



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES



IIT KHARAGPUR



NPTEL

NPTEL ONLINE
CERTIFICATION COURSES

CLOUD COMPUTING

Course Summary and Research Areas

PROF. SOUMYA K. GHOSH

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

IIT KHARAGPUR

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

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.



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

Cloud Computing – Research Areas



IIT KHARAGPUR



NPTEL
ONLINE
CERTIFICATION COURSES

Cloud Infrastructure and Services

- Cloud Computing Architectures
 - Storage and Data Architectures
 - Distributed and Cloud Networking
 - Infrastructure Technologies
-
- IaaS, PaaS, SaaS
 - Storage-as-a-Service
 - Network-as-a-Service
 - Information-as-a-Service



IIT KHARAGPUR



NPTEL
ONLINE
CERTIFICATION COURSES

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
NPTEL ONLINE
CERTIFICATION COURSES

Cloud Security

- Data Privacy
- Access Control
- Identity Management
- Side Channel Attacks
- Security-as-a-Service

NPTEL



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

Performance, Scalability, Reliability

- Performance of cloud systems and Applications
- Cloud Availability and Reliability
- Micro-services based architecture

NPTEL



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

Systems Software and Hardware

- Virtualization Technology
- Service Composition
- Cloud Provisioning Orchestration
- Hardware Architecture support for Cloud Computing



IIT KHARAGPUR



NPTEL
ONLINE
CERTIFICATION COURSES

Data Analytics in Cloud

- Analytics Applications
- Scientific Computing and Data Management
- Big data management and analytics
- Storage, Data, and Analytics Clouds

NPTEL



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

Cloud Computing – Service Management

- Services Discovery and Recommendation
- Services Composition
- Services QoS Management
- Services Security and Privacy
- Semantic Services
- Service Oriented Software Engineering



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

Cloud and Other Technologies

- Fog Computing
- IoT Cloud
- Container Technology

NPTEL



IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

Thank You!

NPTEL



IIT KHARAGPUR



NPTEL
NPTEL ONLINE
CERTIFICATION COURSES



NPTEL ONLINE CERTIFICATION COURSES

Cloud Computing

Prof. Soumya K Ghosh

**Department of Computer Science
and Engineering**

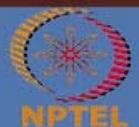
Module 09: Cloud Computing Paradigm

Lecture 41: Cloud-Fog Computing - Overview

CONCEPTS COVERED

- Cloud-Fog Paradigm - Overview
- Cloud-Fog-Edge/IoT
- Case Study: Health Cloud-Fog Framework
- Case Study: Performance

NPTEL



KEYWORDS

- Cloud Computing
- Fog Computing
- Edge, IoT, Sensors
- Performance analysis

NPTEL

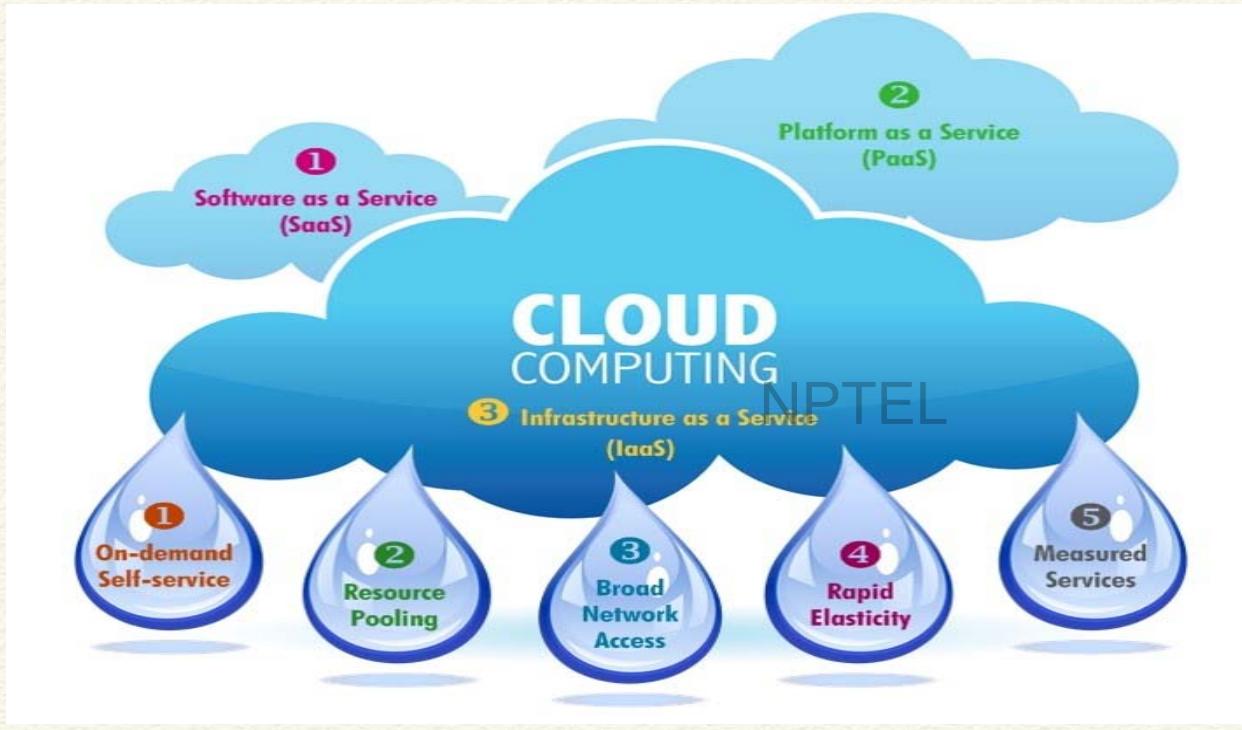


Cloud – Fog Computing Paradigm

NPTEL



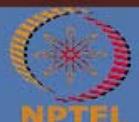
Cloud Computing: “Anything”-as-a-Service



Fog Computing

- Fog computing 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
- Vision of fog computing is to enable applications on billions of connected devices to run directly at the network edge.

NPTEL



Cloud vs Fog Computing

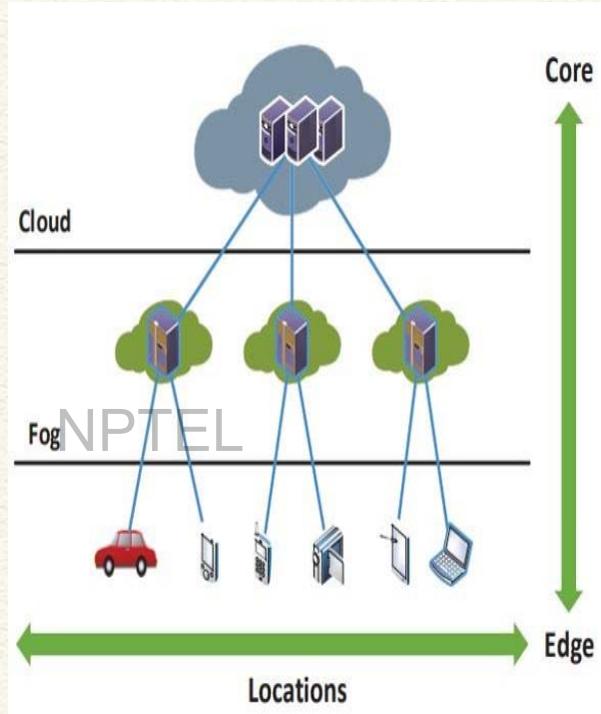
Requirement	Cloud computing	Fog computing
Latency	high	low
Delay jitter	High	Very low
Location of server nodes	With in internet	At the edge of local n/w
Distance between the client and server	Multiple hops	One hop
Security	Undefined	Can be defined
Attack on data enrouter	High probability	Very Less probability
Location awareness	No	Yes

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

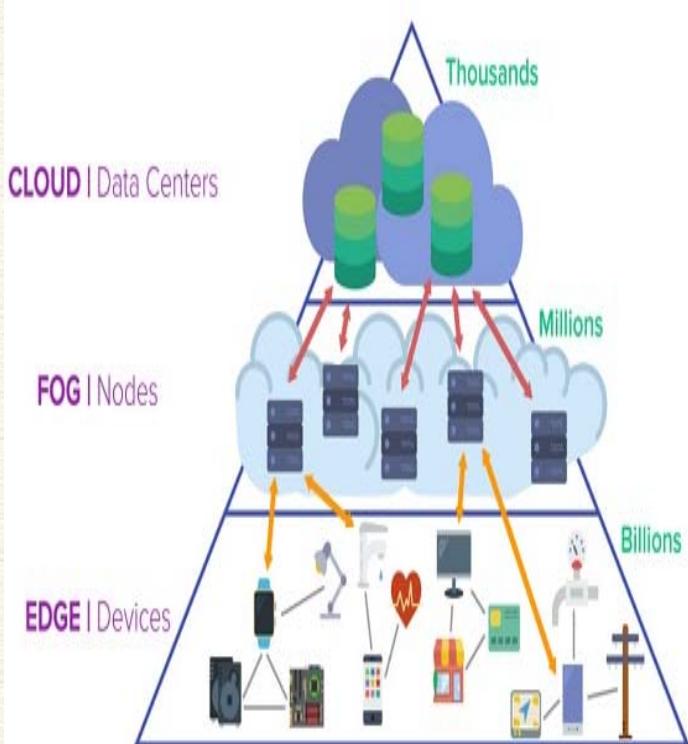


Cloud-Fog-Edge Computing

- Bringing intelligence down from the cloud close to the ground/ end-user.
- Cellular base stations, Network routers, WiFi Gateways will be capable of running applications.
- End devices, like sensors, are able to perform basic data processing.
- Processing close to devices lowers response time, enabling real-time applications.



Cloud – Fog – Edge Computing



Source: <https://www.learntechnology.com/network/fog-computing/>

Cloud Issues Limitations

- Latency
- Large volume of data being generated.
- Bandwidth requirement
- Not designed for volume, variety and velocity of data generated by IoT devices

IoT Device (Edge) Issues

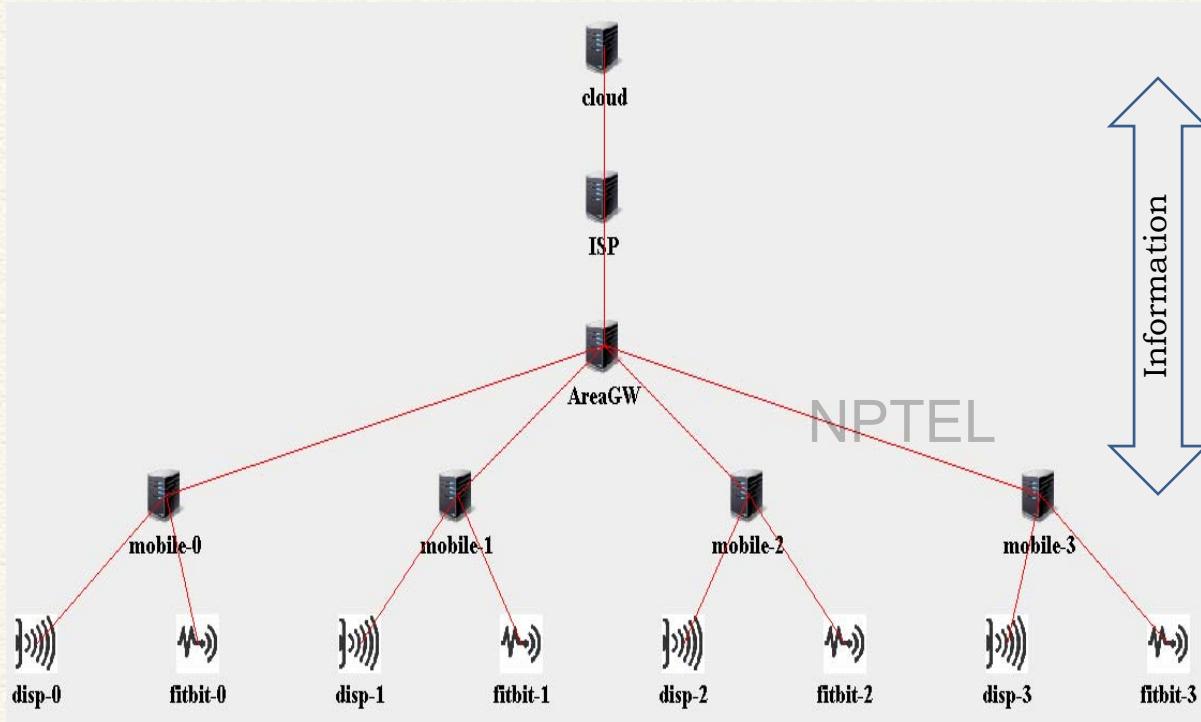
- Processing
- Storage
- Power requirement

Fog layer

- Much lesser **latency** permits usage in Real-time applications
- Less **network congestion**
- Reduced **cost of execution** at cloud
- More of data location awareness
- Better handling of colossal data generated by sensors



Case Study: Health Cloud-Fog-Edge



NPTEL
Information

- Level0: Cloud
- Level1: ISP
- Level2: AreaGW
- Level3: Mobile
- Level4: IoT devices



Case Study: Health Cloud-Fog-Edge

Device Configuration

Device	MIPS	RAM	Up Bw	Down Bw	Level	Cost/MIPS	Busy Power	Idle Power
Cloud	44800	40000	100	10000	0	0.01	16*103	16*83.25
ISP	2800	4000	10000	10000	1	0	107.339	83.4333
AreaGW	2800	4000	10000	10000	2	0	107.339	83.4333
Mobile	350	1000	10000	270	3	0	87.53	82.44

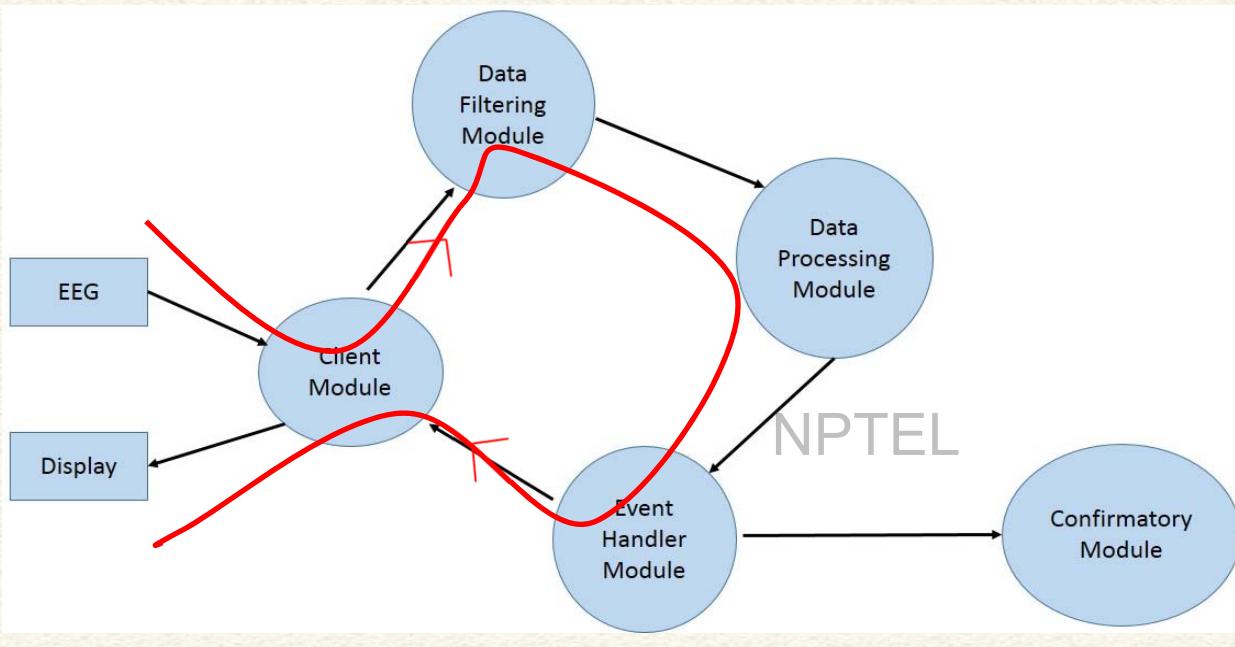
Latency

Source	Destination	Latency
Fitbit	Mobile	1
Mobile	Area GW	2
Area GW	ISP GW	2
ISP GW	Cloud	100
Mobile	Display	1

NPTEL



Case Study: Health Cloud-Fog-Edge



Typical Application Components and Flow



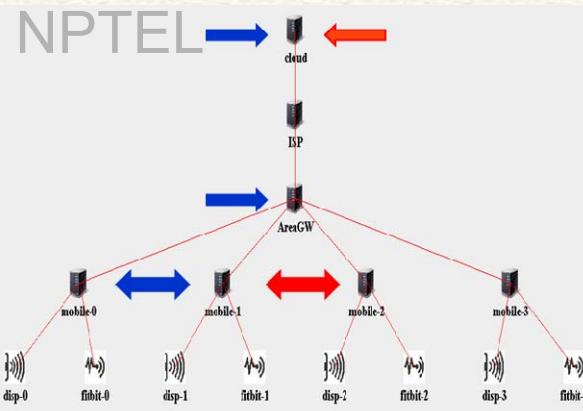
Case Study: Simulation of Health Cloud-Fog Model

Placement obtained for different Application Modules for Fog and Cloud architecture:

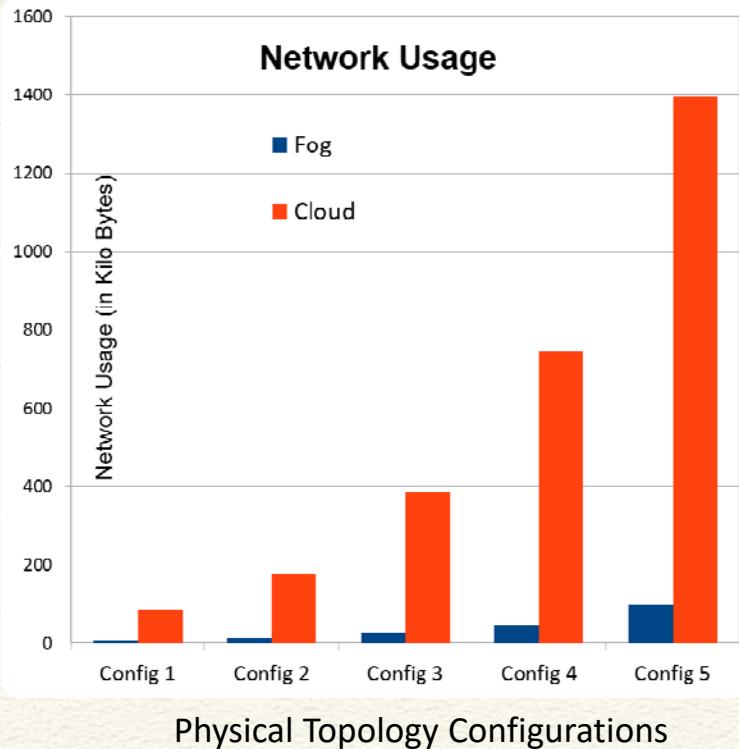
Application Module	Placement in Fog based Model	Placement in Cloud based Model
Client Module	Mobile	Mobile
Data Filtering Module	Area Gateway	Cloud
Data Processing Module	Area Gateway	Cloud
Event Handler Module	Area Gateway	Cloud
Confirmatory Module	Cloud	Cloud

Simulation of different configurations in iFogSim:

Configuration	No. of AreaGW	Total No. of Users
1	1	4
2	2	8
3	4	16
4	8	32
5	16	64



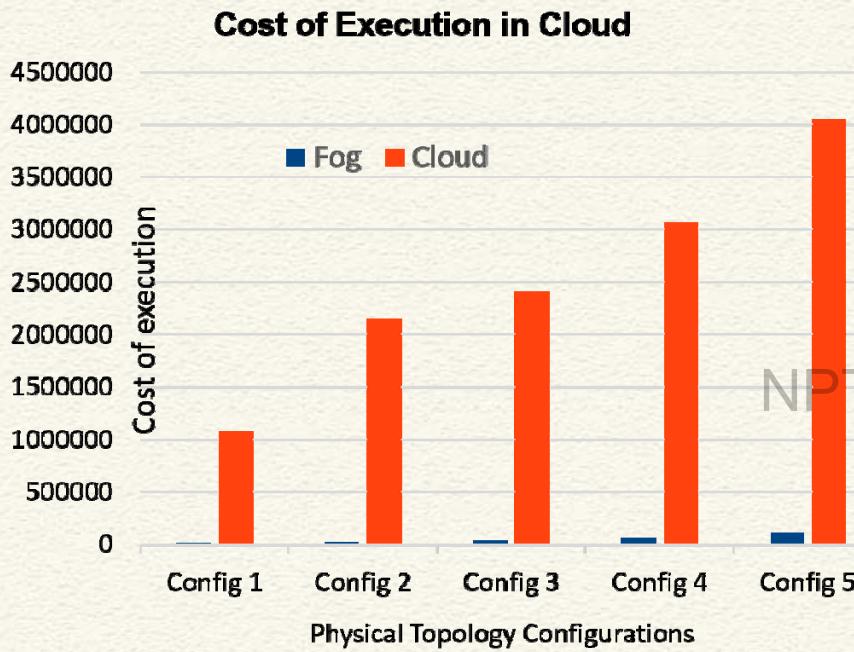
Case Study: Performance Analysis - Network Usage



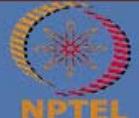
- Network usage is very low in case of Fog architecture as only for few positive cases, the Confirmatory module residing on Cloud is accessed.
- In case of Cloud based architecture, the usage is high as all modules are now on Cloud.



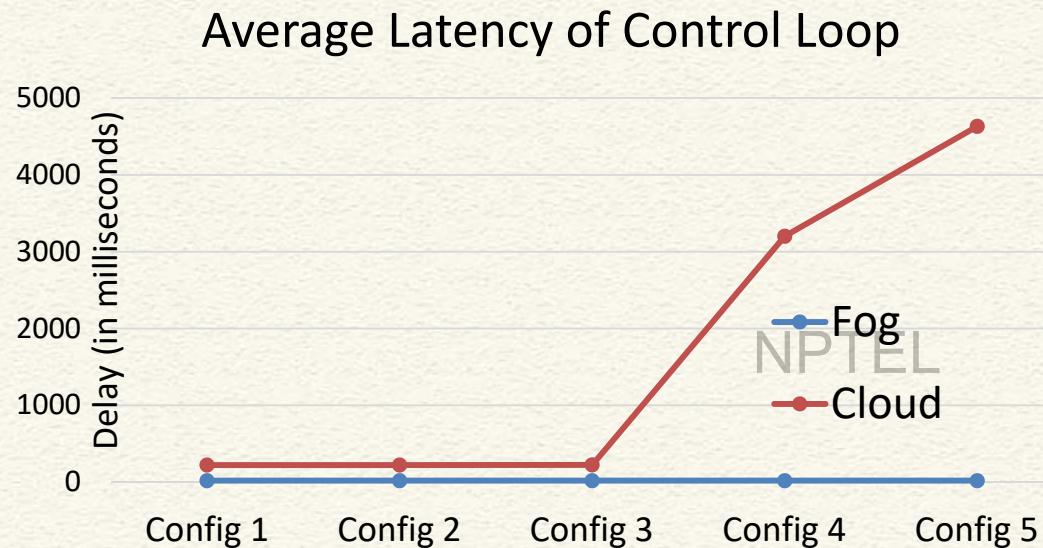
Case Study: Performance Analysis – Execution Cost



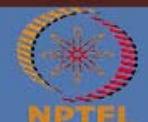
- Only the resources on Cloud incur cost, other resources are owned by the organization.
- More processing at Cloud leads to higher costs in case of Cloud based architecture.



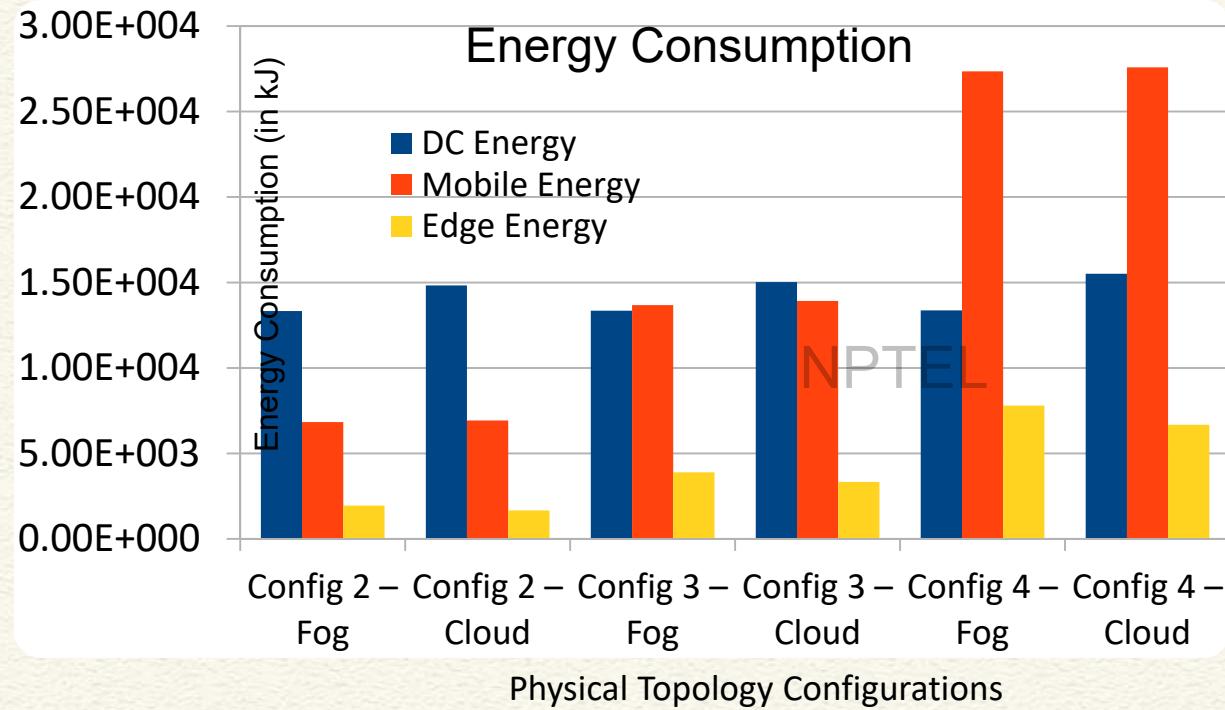
Case Study: Performance Evaluation - Latency



- In this case:
- Latency is fixed in case of Fog architecture as the application modules which form part of the control loop are located at Area Gateway itself.
 - The modules are located at the Datacenter in case of Cloud based architecture.

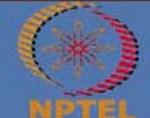
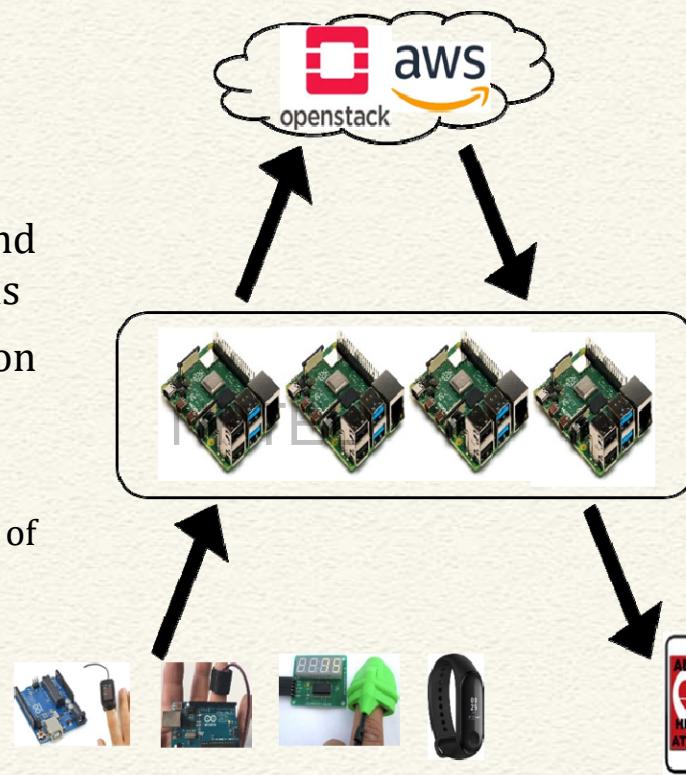


Case Study: Energy Consumptions



Case Study: Prototype Implementation

- Lab based setup:
 - Raspberry Pi as Fog Devices
 - AWS as Cloud
- Use different dataset and customize formulae for analysis
- Changes in Resource Allocation Policy in terms of:
 - Customized physical devices
 - Customized requirements of Application Modules
 - Module Placement policy



REFERENCES

- Cisco White Paper. 2015. Fog Computing and the Internet of Things: Extend the Cloud to Where the Things Are.
- Gupta H, Vahid Dastjerdi A, Ghosh SK, Buyya R. iFogSim: A toolkit for modeling and simulation of resource management techniques in the Internet of Things, Edge and Fog computing environments. *Softw Pract Exper.* 2017;47:1275-296.
<https://doi.org/10.1002/spe.2509>
- Mahmud, Md & Buyya, Rajkumar. (2019). Modeling and Simulation of Fog and Edge Computing Environments Using iFogSim Toolkit: Principles and Paradigms. 10.1002/9781119525080.ch17
- Mahmud, Md and Koch, Fernando and Buyya, Rajkumar. (2018). Cloud-Fog Interoperability in IoT-enabled Healthcare Solutions. 10.1145/3154273.3154347. In proceedings of 19th International Conference on Distributed Computing and Networking, January 4–7, 2018, Varanasi, India. ACM, NewYork, NY, USA.

NPTEL



*Thank
you*



NPTEL





NPTEL ONLINE CERTIFICATION COURSES

Cloud Computing

Prof. Soumya K Ghosh

**Department of Computer Science
and Engineering**

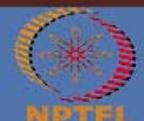
Module 09: Cloud Computing Paradigm

Lecture 42: Resource Management - I

CONCEPTS COVERED

- Cloud-Fog Paradigm – Resource Management Issues
- Service Placement Problem

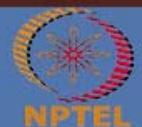
NPTEL



KEYWORDS

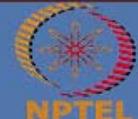
- Cloud Computing
- Fog - Edge Computing
- Resource Management
- Service Placement

NPTEL



Resource Management - I

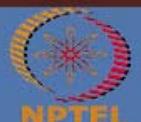
NPTEL



Challenges in “Cloud-only” scenario

- Processing IoT applications directly in the cloud may not be the most efficient solution for each IoT scenario, especially for time-sensitive applications.
- A promising alternative is to use fog and edge computing, which address the issue of managing the large data bandwidth needed by end devices.
- These paradigms impose to process the large amounts of generated data close to the data sources rather than in the cloud.
- One of the considerations of cloud-based IoT environments is resource management, which typically revolves around resource allocation, workload balance, resource provisioning, task scheduling, and QoS to achieve performance improvements.

NPTEL



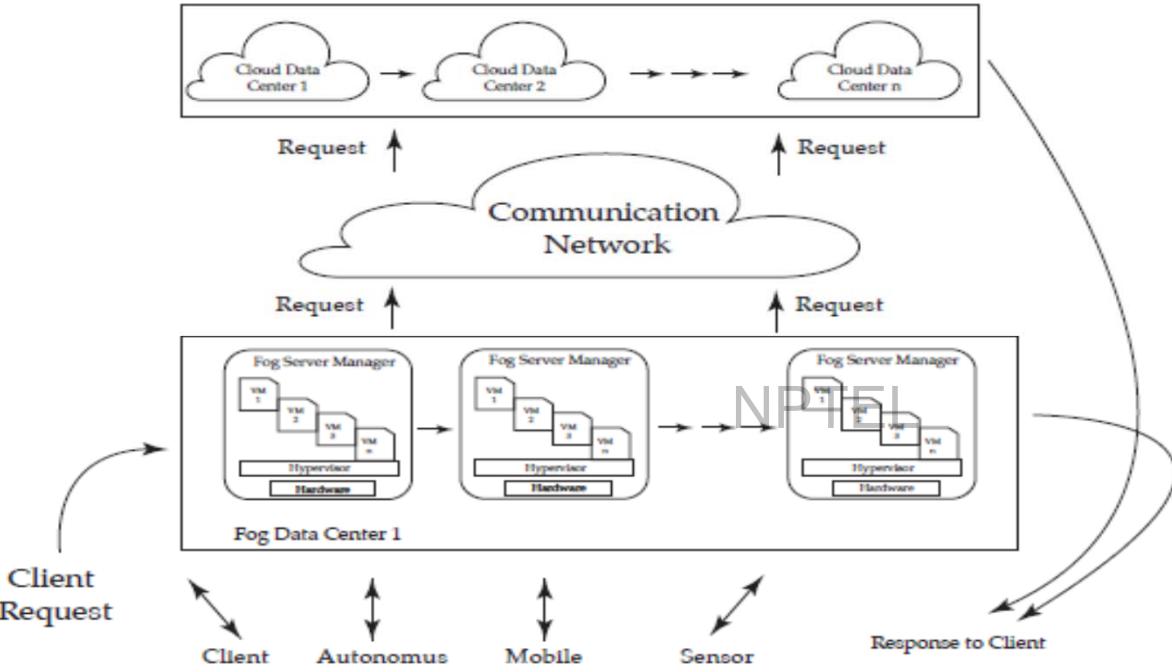
Fog-Edge to support Cloud Computing

- Latency issue: May involve transport of data from each single sensor to a data center over a network, process these data, and then send instructions to actuators.
- Fog and edge computing may aid cloud computing in overcoming these limitations.
- Fog computing and edge computing are no substitutes for cloud computing as they do not completely replace it.
- Oppositely, the three technologies can work together to grant improved latency, reliability, and faster response times.

NPTEL



Cloud-Fog Paradigm



Ref: Agarwal, S.; Yadav, S.; Yadav, A.K. An efficient architecture and algorithm for resource provisioning in fog computing. *Int. J. Inf. Eng. Electronic Bus. (IJIEEB)* 2016, 8, 48–61.



Fog-Edge to support Cloud Computing

- Cloud–fog environment model, typically, is composed of three layers: a client layer (edge), a fog layer, and a cloud layer.
- Fog layer is to accomplish the requirement of resources for clients.
- If there is no/limited availability of resources in the fog layer, then the request is passed to the cloud layer.
- Main functional components of this model are:
 - Fog server manager employs all the available processors to the client.
 - Virtual machines (VMs) operate for the fog data server, process them, and then deliver the results to the fog server manager.
 - Fog servers contain fog server manager and virtual machines to manage requests by using a 'server virtualization technique'.

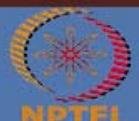
NPTEL



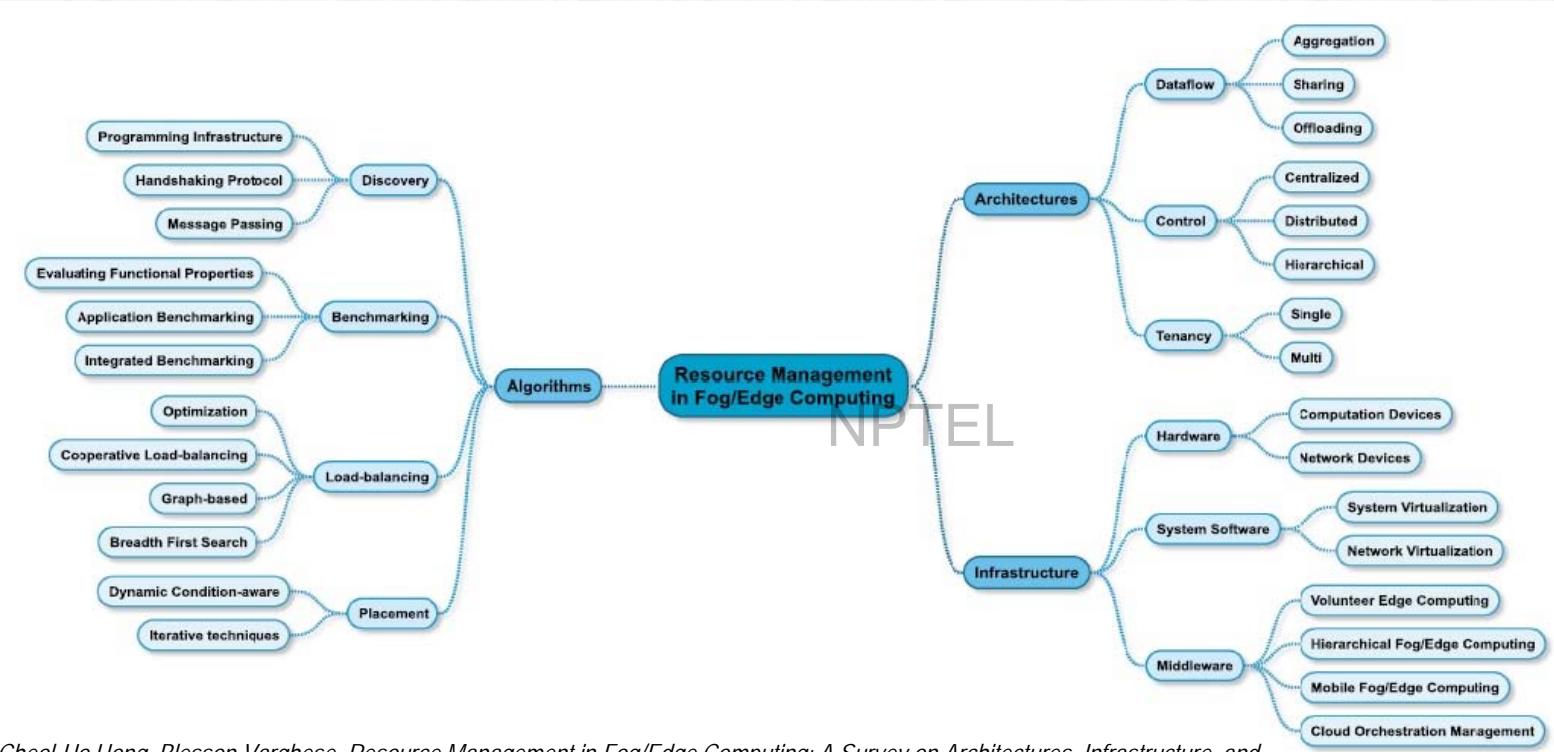
Fog-Edge to support Cloud Computing

- Trend is to decentralize some of the computing resources available in large Cloud data centers by distributing them towards the edge of the network closer to the end-users and sensors
- Resources may take the form of either (i) dedicated “micro” data centers that are conveniently and safely located within public/private infrastructure or (ii) Internet nodes, such as routers, gateways, and switches that are augmented with computing capabilities.
- *A computing model that makes use of resources located at the edge of the network is referred to as “edge computing”.*
- *A model that makes use of both edge resources and the cloud is referred to as “fog computing”*

NPTEL



Resource Management in Cloud-Fog-Edge



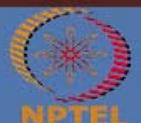
Cheol-Ho Hong, Blesson Varghese, Resource Management in Fog/Edge Computing: A Survey on Architectures, Infrastructure, and Algorithms, ACM Computing Surveys, Vol 52(5), October 2019, pp 1-37.



Resource Management Approaches

- **Architectures** - the architectures used for resource management in fog/edge computing are classified on the basis of data flow, control, and tenancy
- **Infrastructure** - The infrastructure for fog/edge computing provides facilities composed of hardware and software to manage the computation, network, and storage resources for applications utilizing the fog/edge.
- **Algorithms** - There are several underlying algorithms used to facilitate fog/edge computing.

NPTEL

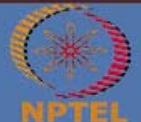


Resource Management Approaches - Architectures

Architectures

Data Flow Control Tenancy

- **Data Flow:** Based on the direction of movement of workloads and data in the computing ecosystem. For example, workloads could be transferred from the user devices to the edge nodes or alternatively from cloud servers to the edge nodes.
- **Control:** Based on how the resources are controlled in the computing ecosystem. For example, a single controller or central algorithm may be used for managing a number of edge nodes. Alternatively, a distributed approach may be employed.
- **Tenancy:** Based on the support provided for hosting multiple entities in the ecosystem. For example, either a single application or multiple applications could be hosted on an edge node.



Resource Management Approaches - Infrastructure

Infrastructure

Hardware

System Software

Middleware

- **Hardware:** Fog/ edge computing exploits small-form-factor devices such as network gateways, WiFi Access Points (APs), set-top boxes, small home servers, edge ISP servers, cars, and even drones as compute servers for resource efficiency. also utilizes commodity products such as desktops, laptops, and smartphones.
- **System Software:** Runs directly on fog/edge hardware resources such as the CPU, memory, and network devices. It manages resources and distributes them to the fog/edge applications. E.g., operating systems and virtualization software.
- **Middleware:** Runs on an operating system and provides complementary services that are not supported by the system software. The middleware coordinates distributed compute nodes and performs deployment of VMs or containers to each fog/edge node.

NPTEL



Resource Management Approaches – Algorithms

Algorithms

Discovery

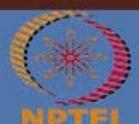
Benchmarking

Load Balancing

Placement

- **Discovery:** Identifying edge resources so workloads from the clouds or from user devices/sensors can be deployed on them
- **Benchmarking:** Capturing the performance (of entities like, CPU, storage, network, etc.) of a computing system
- **Load balancing:** As edge data centers are deployed across the network edge, the issue of distributing tasks using an efficient load-balancing algorithm has gained significant attention. Typical techniques are, namely, optimization techniques, cooperative load balancing, graph-based balancing, and using breadth-first search.
- **Placement:** Addresses the issue of place incoming computation tasks on suitable fog/edge resources, considering the availability of resources in the fog/edge layer and the environmental changes. Dynamic condition-aware techniques and iterative techniques.

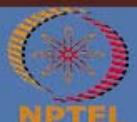
NPTEL



REFERENCES

- Cheol-Ho Hong, Blessen Varghese, Resource Management in Fog/Edge Computing: A Survey on Architectures, Infrastructure, and Algorithms, ACM Computing Surveys, Vol 52(5), October 2019, pp 1–37.
- Farah Aït Salaht, Frédéric Desprez, and Adrien Lebre. 2020. An Overview of Service Placement Problem in Fog and Edge Computing. ACM Comput. Surv. 53, 3, Article 65 (June 2020), 35 pages.
- Agarwal, S.; Yadav, S.; Yadav, A.K. An efficient architecture and algorithm for resource provisioning in fog computing. Int. J. Inf. Eng. Electronic Bus. (IJIEEB) 2016, 8, 48–61.

NPTEL



*Thank
you*



NPTEL





NPTEL ONLINE CERTIFICATION COURSES

Cloud Computing

Prof. Soumya K Ghosh

**Department of Computer Science
and Engineering**

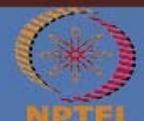
Module 09: Cloud Computing Paradigm

Lecture 43: Resource Management - II

CONCEPTS COVERED

- Cloud-Fog Paradigm – Resource Management Issues
- Service Placement Problem
- Service and Data Offloading
- Hardware and Software

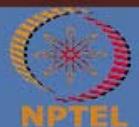
NPTEL



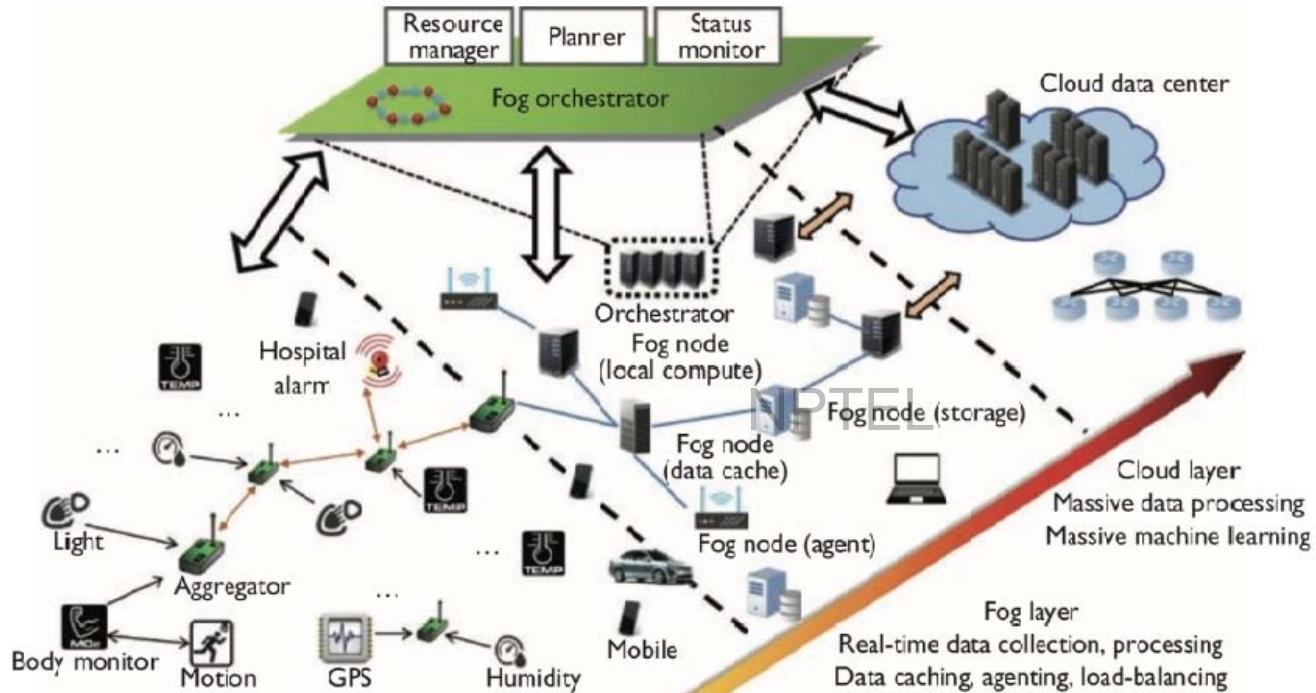
KEYWORDS

- Cloud Computing
- Fog - Edge Computing
- Resource Management
- Service Placement

NPTEL



Service Placement Problem in Fog and Edge Computing



Ref: Farah Ait Salaht, Frédéric Desprez, and Adrien Lebre. 2020. An Overview of Service Placement Problem in Fog and Edge Computing. *ACM Comput. Surv.* 53, 3, Article 65 (June 2020), 35 pages.

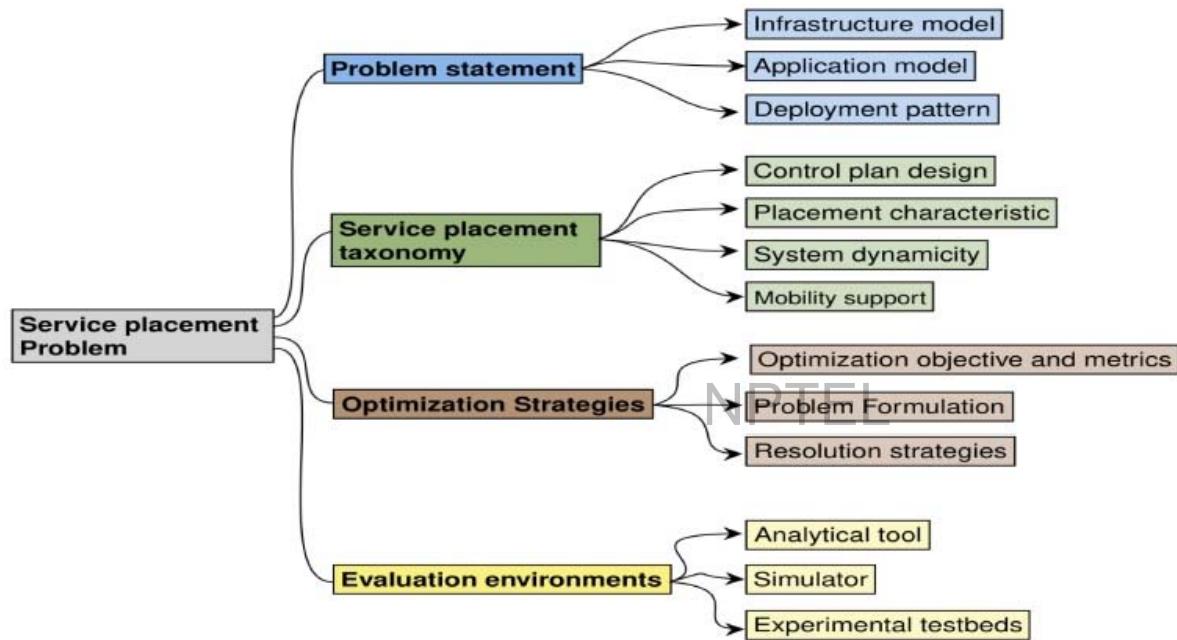


Service Placement Problem in Fog and Edge Computing

- Fog Computing is a highly virtualized platform that offers computational resources, storage, and control between end-users and Cloud servers.
- It is a new paradigm in which centralized Cloud coexists with distributed edge nodes and where the local and global analyses are performed at the edge devices or forwarded to the Cloud.
- Fog infrastructure consists of IoT devices (End layer), Fog Nodes, and at least one Cloud Data Center (Cloud layer), with following characteristics:
 - Location awareness and low latency
 - Better bandwidth utilization
 - Scalable
 - Support for mobility



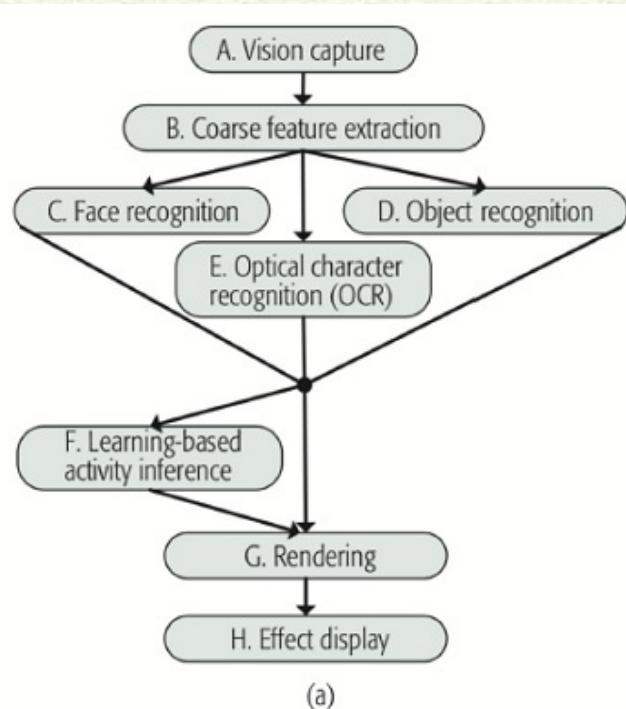
Service Placement Problem in Fog and Edge Computing



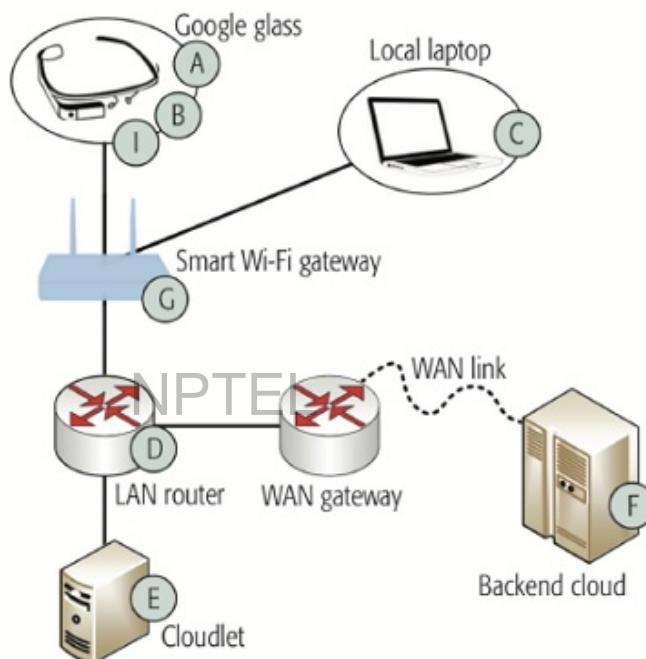
Ref: Farah Ait Salaht, Frédéric Desprez, and Adrien Lebre. 2020. An Overview of Service Placement Problem in Fog and Edge Computing. *ACM Comput. Surv.* 53, 3, Article 65 (June 2020), 35 pages.



Deployment (Application Placement) on Cloud-Fog-Edge framework

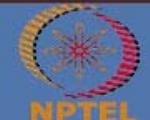


(a)



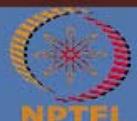
(b)

Ref: Farah Ait Salaht, Frédéric Desprez, and Adrien Lebre. 2020. An Overview of Service Placement Problem in Fog and Edge Computing. *ACM Comput. Surv.* 53, 3, Article 65 (June 2020), 35 pages.



Application Placement on Cloud-Fog-Edge framework

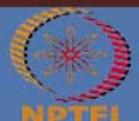
- Application placement problem defines a mapping pattern by which applications components and links are mapped onto an infrastructure graph (i.e., computing devices and physical edges)
- Application placement involves finding the available resources in the network (nodes and links) that satisfy the application(s) requirements, satisfy the constraints, and optimize the objective.
- For instance, respect the applications (services) requirements, not exceed the resource capacities, satisfy the locality constraints, minimize the energy consumed, and so on.
- Service providers have to take into account these constraints to, (i) limit the research space and, (ii) provide an optimum or near-optimum placement



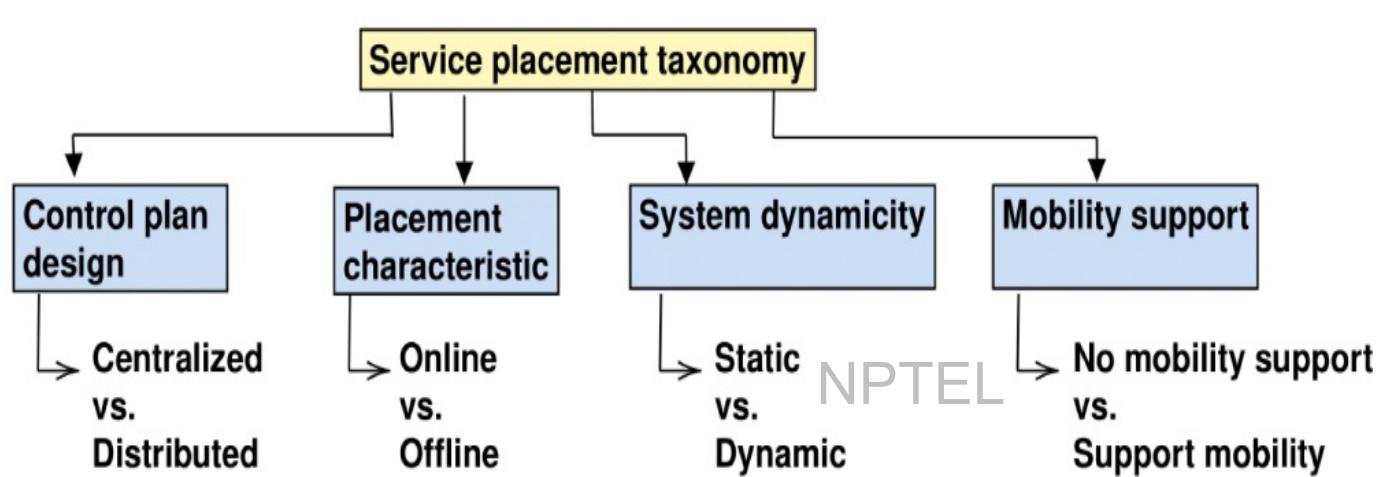
Application Placement - Constraints

- **Resource constraints:** An infrastructure node is limited by finite capabilities in terms of CPU, RAM, storage, bandwidth, etc. While placing application(s) (service components), the resource requirements need to be considered
- **Network constraints:** constraints such as latency, bandwidth, etc. and these constraints need to be considered when deploying applications.
- **Application constraints:**
 - Locality requirement: restricts certain services' executions to specific locations
 - Delay sensitivity: Some applications can specify a deadline for processing operation or deploying the whole application in the network

NPTEL



Service Placement Taxonomy



Ref: Farah Ait Salaht, Frédéric Desprez, and Adrien Lebre. 2020. An Overview of Service Placement Problem in Fog and Edge Computing. *ACM Comput. Surv.* 53, 3, Article 65 (June 2020), 35 pages.



Service Placement – Optimization Strategies

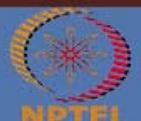
- Optimizing the service placement problem in a Cloud-Fog infrastructure can have several different objectives, with different formulations and diverse algorithm proposals.
 - Optimization Objective and Metrics :
 - Latency
 - Resource utilization
 - Cost
 - Energy consumption

NPTEL

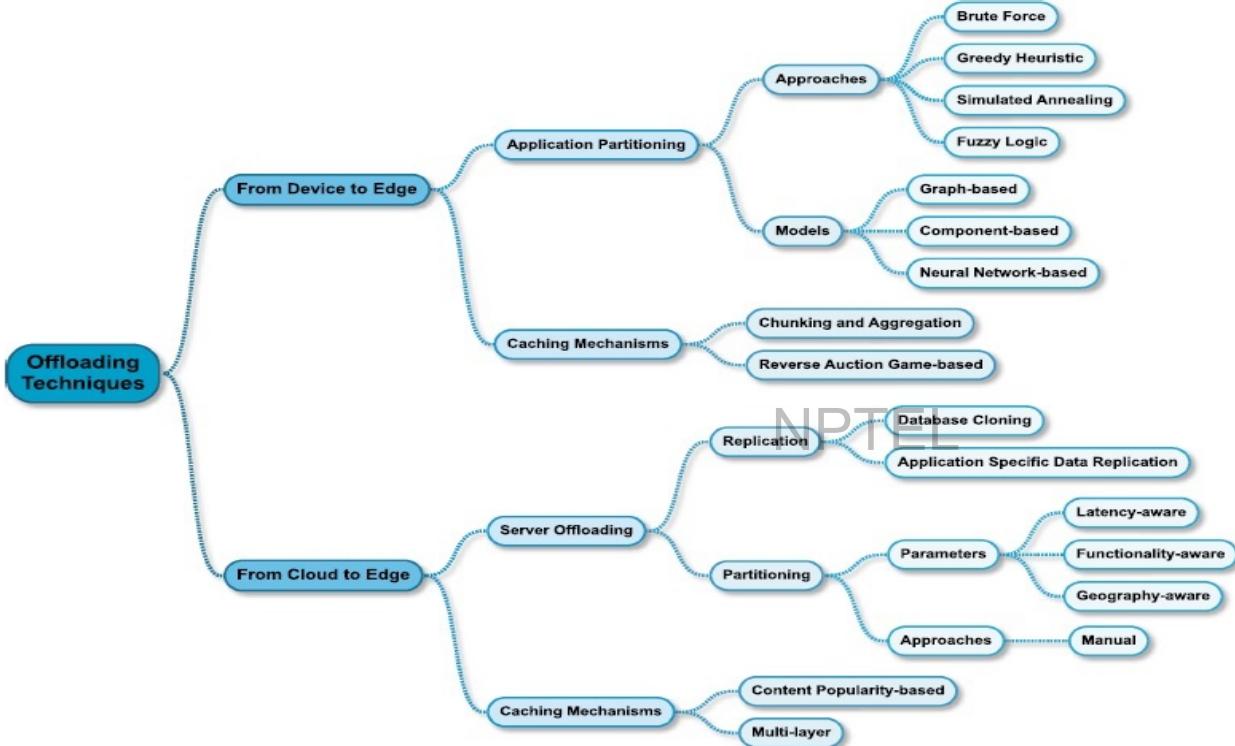


Offloading – Application and Data

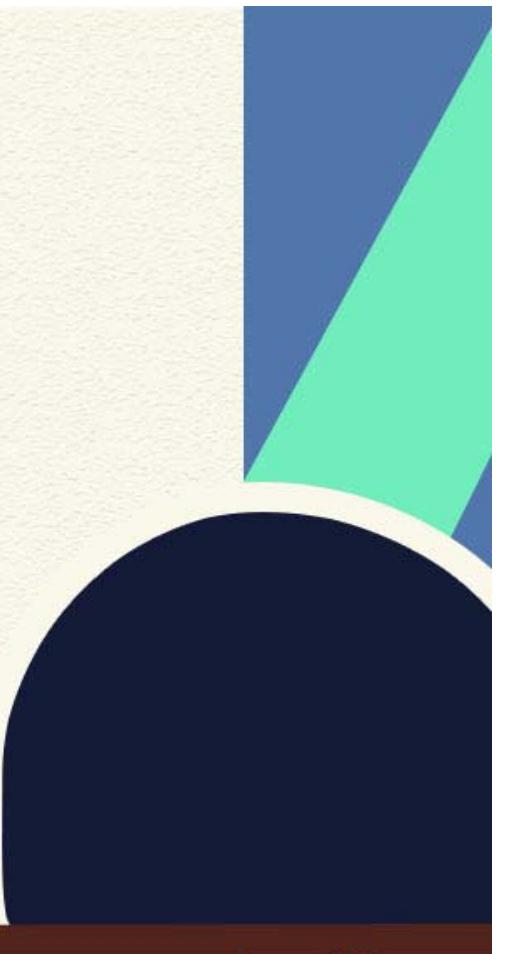
- Offloading is a technique in which a server, an application, and the associated data are moved onto the edge of the network.
- Augments the (i) computing requirements of individual or a collection of user devices, (ii) brings services in the cloud that process requests from devices closer to the source.
- *Offloading from User Device to Edge:* Augments computing in user devices by making use of edge nodes (usually a single hop away)
(i) Application partitioning, (ii) Caching mechanisms
- *Offloading from the Cloud to the Edge:* A workload is moved from the cloud to the edge.
(i) Server offloading, (ii) Caching mechanisms



Offloading – Application and Data (contd.)

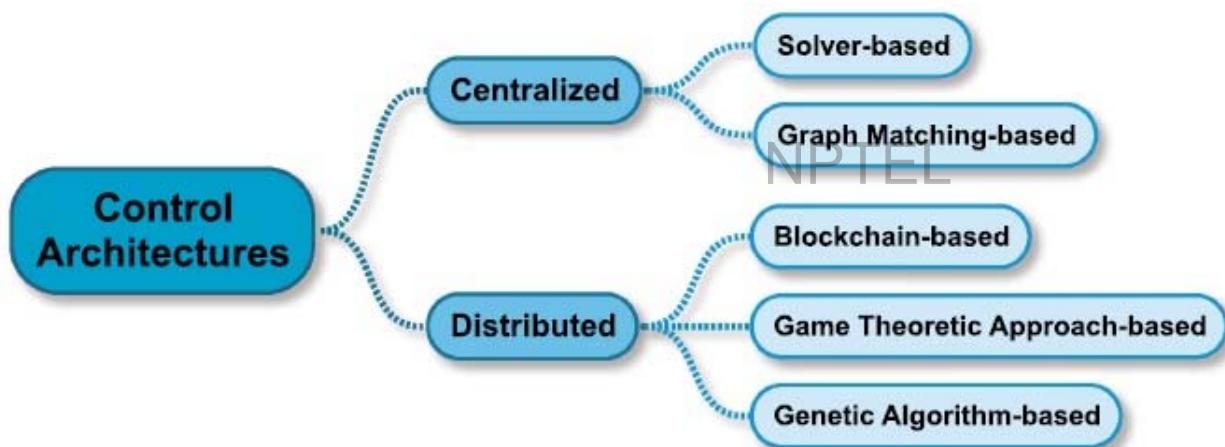


Ref: Farah Ait Salaht, Frédéric Desprez, and Adrien Lebre. 2020. An Overview of Service Placement Problem in Fog and Edge Computing. *ACM Comput. Surv.* 53, 3, Article 65 (June 2020), 35 pages.

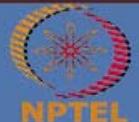


Control

- Centralized
- Distributed



Ref: Farah Ait Salaht, Frédéric Desprez, and Adrien Lebre. 2020. An Overview of Service Placement Problem in Fog and Edge Computing. *ACM Comput. Surv.* 53, 3, Article 65 (June 2020), 35 pages.



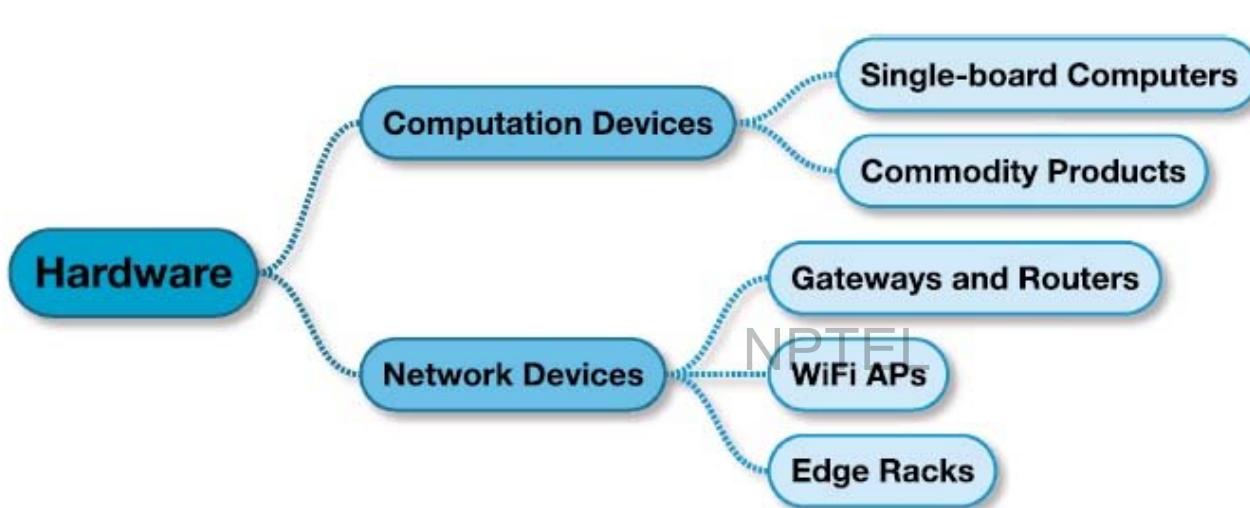
Hardware

- Fog/edge computing forms a computing environment that uses low-power devices, namely, mobile devices, routers, gateways, home systems.
- Combination of these small-form-factor devices, connected to network, enables a cloud computing environment that can be leveraged by a rich set of applications processing Internet of Things (IoT) and cyber-physical systems (CPS) data.

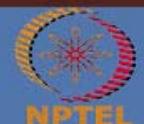
NPTEL



Hardware (contd..)



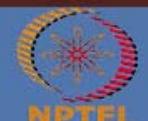
Ref: Farah Ait Salaht, Frédéric Desprez, and Adrien Lebre. 2020. An Overview of Service Placement Problem in Fog and Edge Computing. *ACM Comput. Surv.* 53, 3, Article 65 (June 2020), 35 pages.



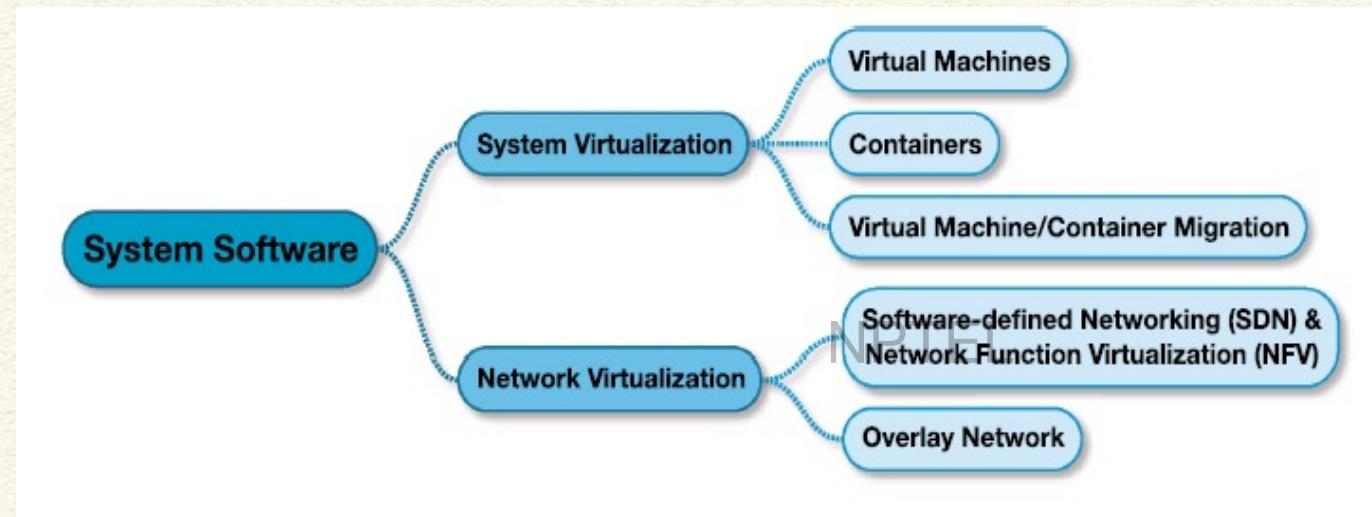
System Software

- System software for the fog/edge is a platform designed to operate directly on fog/edge devices
- Manage the computation, network, and storage resources of the devices.
- System software needs to support multi-tenancy and isolation, because fog/edge computing accommodates several applications from different tenants.
- Two categories
 - System Virtualization
 - Network Virtualization

NPTEL



System Software (contd..)

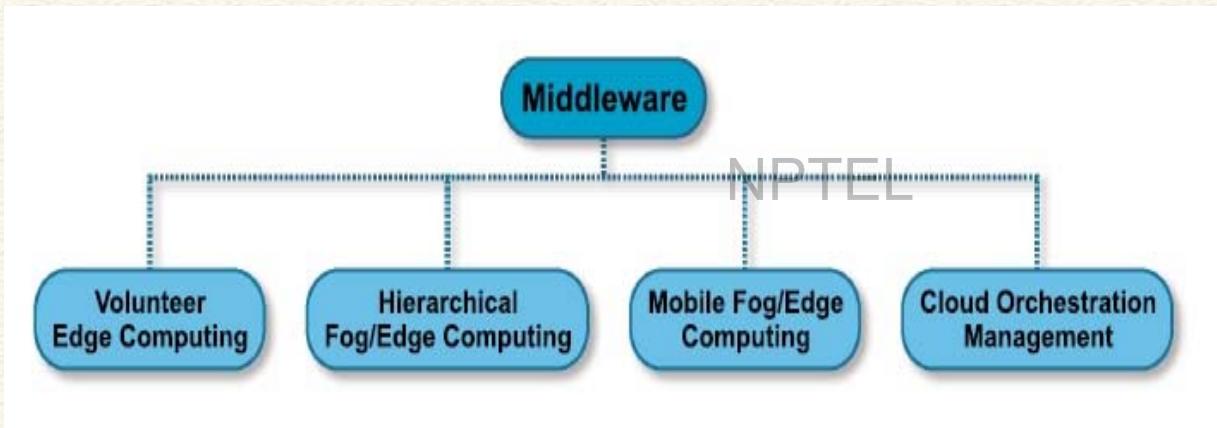


Ref: Farah Ait Salaht, Frédéric Desprez, and Adrien Lebre. 2020. An Overview of Service Placement Problem in Fog and Edge Computing. *ACM Comput. Surv.* 53, 3, Article 65 (June 2020), 35 pages.



Middleware

- Middleware provides complementary services to system software.
- Middleware in fog/edge computing provides performance monitoring, coordination and orchestration, communication facilities, protocols etc.



Ref: Farah Ait Salaht, Frédéric Desprez, and Adrien Lebre. 2020. An Overview of Service Placement Problem in Fog and Edge Computing. *ACM Comput. Surv.* 53, 3, Article 65 (June 2020), 35 pages.



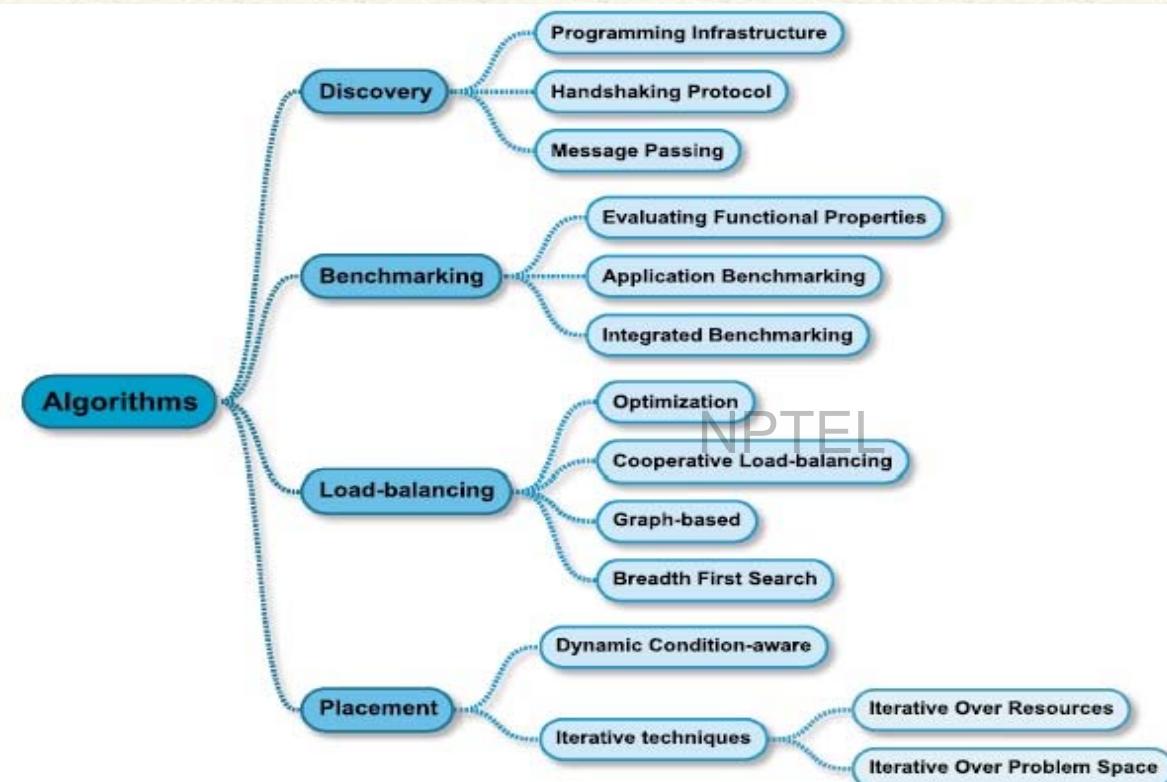
ALGORITHMS

- Algorithms used to facilitate fog/edge computing. Four major algorithms.
- *Discovery*: identifying edge resources within the network that can be used for distributed computation
- *Benchmarking*: capturing the performance of resources for decision-making to maximize the performance of deployments
- *Load-balancing*: distributing workloads across resources based on different criteria such as priorities, fairness etc.
- *Placement*: identifying resources appropriate for deploying a workload.

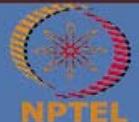
NPTEL



ALGORITHMS (contd..)



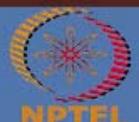
Ref: Farah Ait Salaht, Frédéric Desprez, and Adrien Lebre. 2020. An Overview of Service Placement Problem in Fog and Edge Computing. *ACM Comput. Surv.* 53, 3, Article 65 (June 2020), 35 pages.



REFERENCES

- Cheol-Ho Hong, Blessen Varghese, Resource Management in Fog/Edge Computing: A Survey on Architectures, Infrastructure, and Algorithms, ACM Computing Surveys, Vol 52(5), October 2019, pp 1–37.
- Farah Aït Salaht, Frédéric Desprez, and Adrien Lebre. 2020. An Overview of Service Placement Problem in Fog and Edge Computing. ACM Comput. Surv. 53, 3, Article 65 (June 2020), 35 pages.
- Agarwal, S.; Yadav, S.; Yadav, A.K. An efficient architecture and algorithm for resource provisioning in fog computing. Int. J. Inf. Eng. Electronic Bus. (IJIEEB) 2016, 8, 48–61.

NPTEL



*Thank
you*



NPTEL





NPTEL ONLINE CERTIFICATION COURSES

Cloud Computing

Prof. Soumya K Ghosh

**Department of Computer Science
and Engineering**

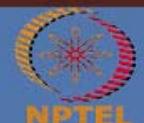
Module 09: Cloud Computing Paradigm

Lecture 44: Cloud Federation

CONCEPTS COVERED

- Cloud-Fog Paradigm – Resource Management Issues
- Service Placement Problem

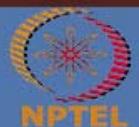
NPTEL



KEYWORDS

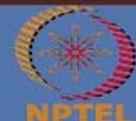
- Cloud Computing
- Fog - Edge Computing
- Resource Management
- Service Placement

NPTEL



Cloud Federation

NPTEL



Cloud Federation?

- A federated cloud (also called cloud federation) is the deployment and management of multiple external and internal cloud computing services to match business needs.
- A federation is the union of several smaller parts that perform a common action.

NPTEL

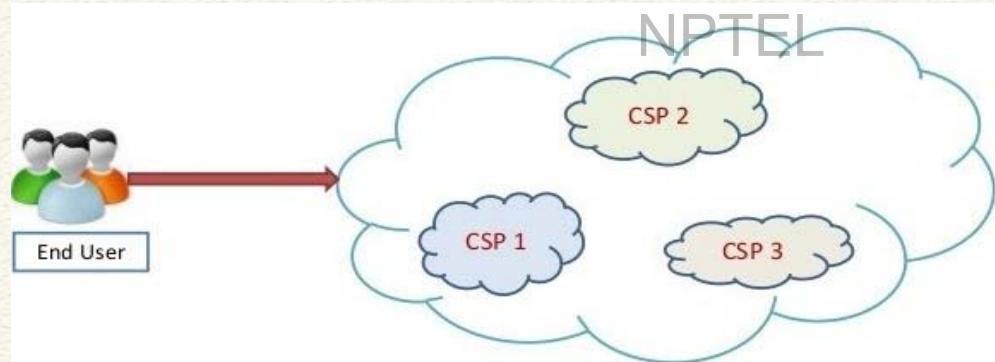
[Ref: <http://whatis.techtarget.com/definition/federated-cloud-cloud-federation>]



Cloud Federation?

Collaboration between Cloud Service Providers (CSPs) to achieve:

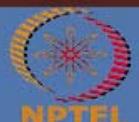
- Capacity utilization
- Inter-operability
- Catalog of services
- Insight about providers and SLA's



Federation - Motivation

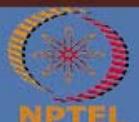
- Different CSPs join together to form a federation
- Benefits:
 - Maximize resource utilization
 - Minimize power consumption
 - Load balancing
 - Global utility
 - Expand CSP's global foot prints

NPTEL



Federation - Characteristics

- To overcome the current limitations of cloud computing such as service interruptions, lack of interoperability and degradation of services.
 - Many inter-cloud organizations have been proposed.
 - Cloud federation is an example of an inter-cloud organization.
-
- It is a inter-cloud organization with voluntary characteristics.
 - It should have maximum geographical separation.
 - Well defined marketing system and regulated federal agreement.
 - IT is an environment where multiple SP come together and share their resources.



Federation Architecture

- Cloud federation is associated with several portability and interoperability issues.
- Typical federation architectures: cloud bursting, brokering, aggregation, and multitier.
- These architectures can be classified according to the level of coupling or interoperation among the cloud instances involved, ranging from loosely coupled (with no or little interoperability among cloud instances) to tightly coupled (with full interoperability among cloud instances).



Loosely Coupled Federation

- Limited interoperation between CSPs / cloud instances.
Example: a private cloud complementing its infrastructure with resources from an external commercial cloud
- A CSP has little or no control over remote resources (for example, decisions about VM placement are not allowed), monitoring information is limited (for example, only CPU, memory, or disk consumption of each VM is reported), and there is no support for advanced features such as cross-site networks or VM migration.



Partially Coupled Federation

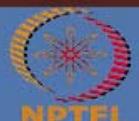
- Different CSPs (partner clouds) establish a contract or framework agreement stating the terms and conditions under which one partner cloud can use resources from another.
- This contract can enable a certain level of control over remote resources (for example, allowing the definition of affinity rules to force two or more remote VMs to be placed in the same physical cluster);
- May agree to the interchange of more detailed monitoring information (for example, providing information about the host where the VM is located, energy consumption, and so on);
- May enable some advanced networking features among partner clouds (for example, the creation of virtual networks across site boundaries).



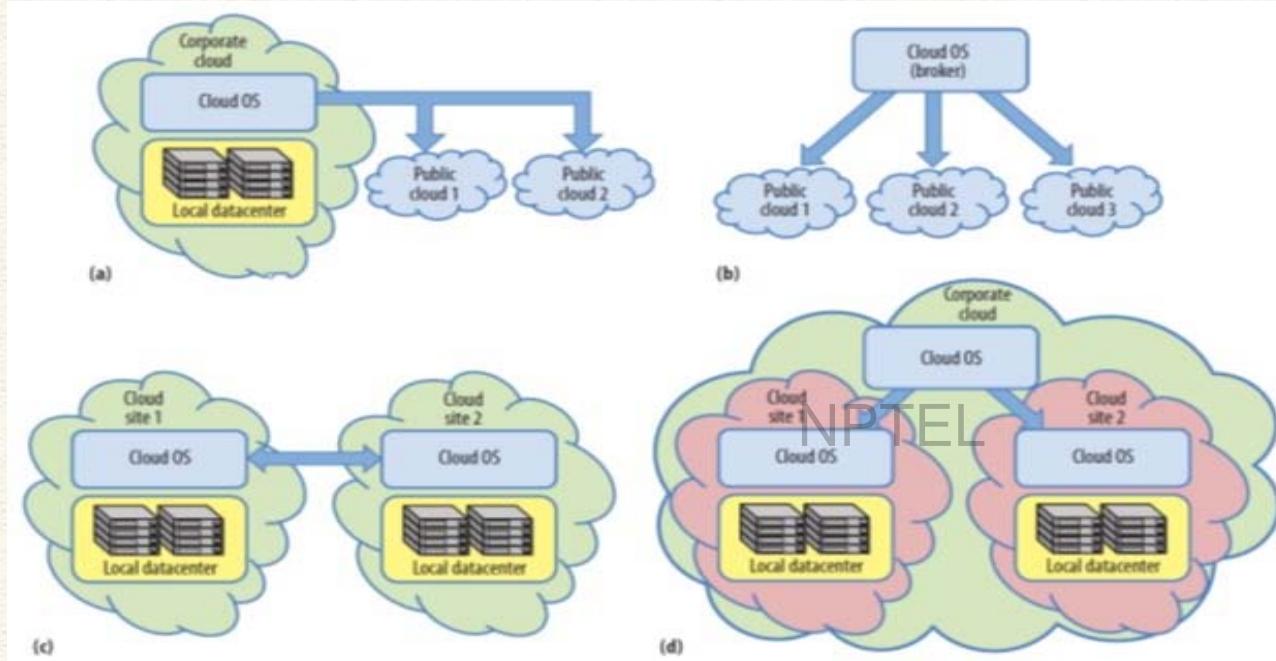
Tightly Coupled Federation

- In this case the clouds are normally governed by the same cloud administration.
- A cloud instance can have advanced control over remote resources—for example, allowing decisions about the exact placement of a remote VM—and can access all the monitoring information available about remote resources.
- May allow other advanced features, including the creation of cross-site networks, cross-site migration of VMs, implementation of high availability techniques among remote cloud instances, and creation of virtual storage systems across site boundaries.

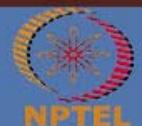
NPTEL



Cloud Federation Architectures



(a) Hybrid / Bursting, (b) Broker, (c) Aggregated, (d) Multiplier



Hybrid / Bursting Architecture

- Cloud bursting or hybrid architecture combines the existing on-premise infrastructure (usually a private cloud) with remote resources from one or more public clouds to provide extra capacity to satisfy peak demand periods.
- As the local cloud OS has no advanced control over the virtual resources deployed in external clouds beyond the basic operations the providers allow, this architecture is loosely coupled.
- Most existing open cloud managers support the hybrid cloud architecture

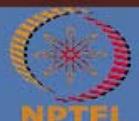
NPTEL



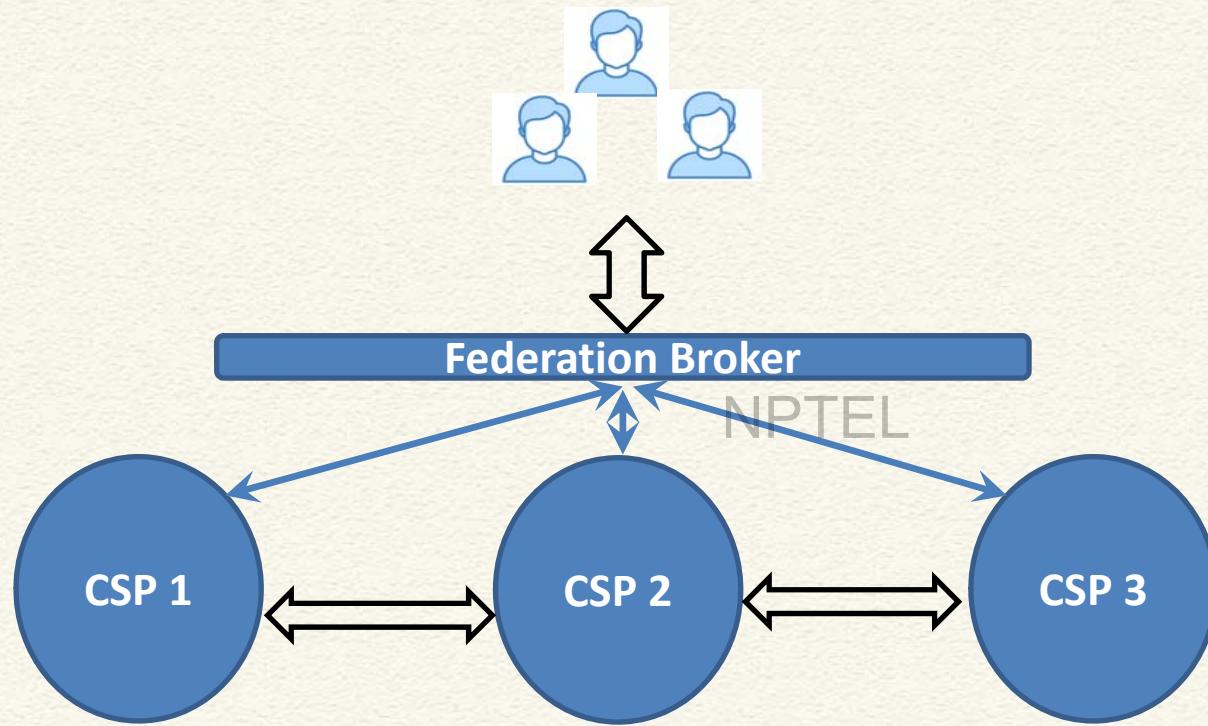
Broker Architecture

- A broker that serves various users and has access to several public cloud infrastructures. A simple broker should be able to deploy virtual resources in the cloud as selected by the user.
- Brokering is the most common federation scenario.
- An advanced broker offering service management capabilities could make scheduling decisions based on optimization criteria such as cost, performance, or energy consumption to automatically deploy virtual user service in the most suitable cloud
- It may even distribute the service components across multiple clouds. This architecture is also loosely coupled since public clouds typically do not allow advanced control over the deployed virtual resources.

NPTEL



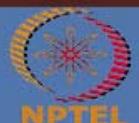
Broker Architecture



Aggregated Architecture

- Involves two or more partner clouds that interoperate to aggregate their resources and provide users with a larger virtual infrastructure.
- This architecture is usually partially coupled, since partners could be provided with some kind of advanced control over remote resources, depending on the terms and conditions of contracts with other partners.
- The partner clouds usually have a higher coupling level when they belong to the same corporation than when they are owned by different companies that agree to cooperate and aggregate their resources.

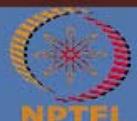
NPTEL



Multitier Architecture

- Involves two or more cloud sites, each running its own cloud OS and usually belonging to the same corporation, that are managed by a third cloud OS instance following a hierarchical arrangement.
- This root/top cloud OS instance has full control over resources in different cloud sites—a tightly coupled scenario—and it exposes the resources available in the different cloud sites as if they were located in a single cloud.
- This architecture is beneficial for corporations with geographically distributed cloud infrastructures because it provides uniform access.
- It may be useful for implementing advanced management features such as high availability, load balancing, and fault tolerance.

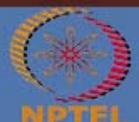
NPTEL



REFERENCES

- Moreno-Vozmediano, R., Montero, R., and Llorente, I., "IaaS Cloud Architecture: From Virtualized Data Centers to Federated Cloud Infrastructures", IEEE Computer Vol. 45 (12), Dec. 2012, pp. 65-72.
- Sanjay P. Ahuja, IaaS Cloud Architectures: Virtualized Data Centers to Federated Cloud Infrastructures, School of Computing, UNF

NPTEL



*Thank
you*



NPTEL





NPTEL ONLINE CERTIFICATION COURSES

Cloud Computing

Prof. Soumya K Ghosh

**Department of Computer Science
and Engineering**

Module 10: Cloud Computing Paradigm

Lecture 45: Cloud Migration - I

CONCEPTS COVERED

- VM Migration - Basics
- Migration strategies

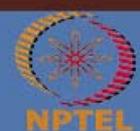
NPTEL



KEYWORDS

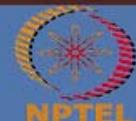
- Virtual Machine (VM)
- VM Migration

NPTEL



VM Migration

NPTEL



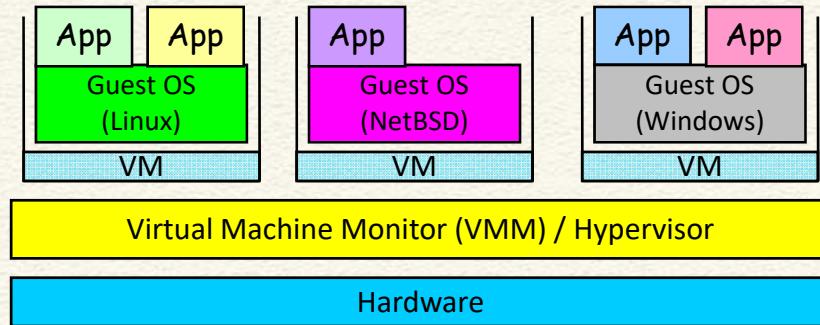
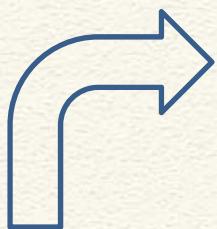
VM Migration

- VM Migration – It is process to move running applications or VMs from one physical server/ host to another host.
- Processor state, storage, memory and network connection are moved from one host to another host
- Why to migrate VMs?
 - Distribute VM load efficiently across servers in a cloud
 - System maintenance

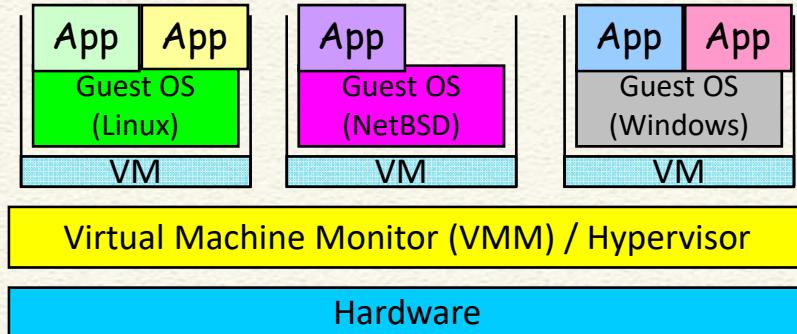
NPTEL



Virtualization



NPTEL



VM Migration – Needs

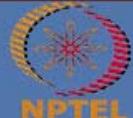
- **Load Balancing:** For fair distribution of workload among computing resources.
- **Maintenance:** For server maintenance VMs can be migrated transparently from one server to another.
- **Manage Operational Parameters:** To reduce operational parameters like power consumption, VMs can be consolidated on minimal number of servers. Under-utilized servers can be put on a low power mode to reduce power consumption.
- **Quality-of-Service violation:** When the service provider fails to meet the desired quality-of-services (QoS) a user can migrate his VM to another service provider.
- **Fault Tolerance:** In case of failure, VMs can be migrated from one data center to another where they can be executed

NPTEL



VM Migration – Types

- **Cold or Non-Live Migration:** In case of cold migration the VM executing on the source machine is turned off or suspended during the migration process.
- **Hot or Live Migration:** In case of a hot or live migration the VM executing on the source machine continues to provide service during the migration process. In fact the target VM is not suspended during the migration process.



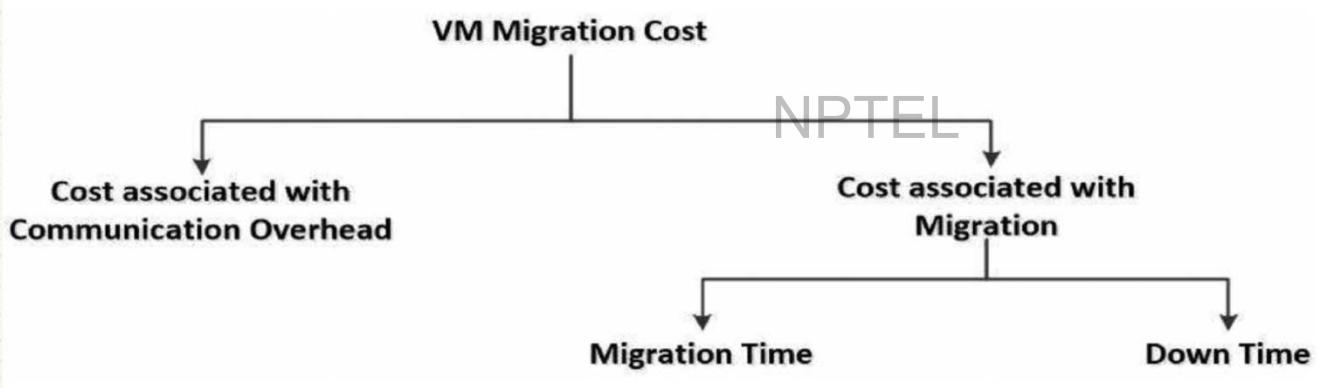
Live VM Migration

- Migrate an entire VM from one physical host to another
 - All user processes and kernel state
 - Without having to shut down the machine
- In case of non live migration the VM providing the services remains suspended during the entire migration process. Hence for large sized VMs the service downtime might be very high.
- For real time applications non live migration can cause severe degradation in service quality which is not tolerable.
- Two main approaches: Pre-copy and Post-copy



When to Migrate?

- To remove a physical machine from service.
- To relieve load on congested hosts.



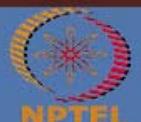
Migration - Concerns

- Minimize the downtime
 - Downtime refers to the total amount of time services remain unavailable to the users.
- Minimize total migration time
 - Migration time refers to the total time taken to move a VM from the source host to the destination host. It can be considered as the total time taken for the entire migration process.
- Migration does not unnecessarily disrupt active services through resource contention (e.g., CPU, network bandwidth) with the migrating OS.



What to Migrate?

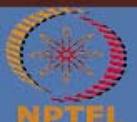
- CPU context of VM, contents of Main Memory
- Disk
 - If NAS (network attached storage) that is accessible from both hosts, or local disk is mirrored – migrating disk data may not be critical
- Network: *assume both hosts on same LAN*
 - Migrate IP address, advertise new MAC address to IP mapping via ARP reply
 - Migrate MAC address, let switches learn new MAC location
 - Network packets redirected to new location (with transient losses)
- I/O devices
 - Virtual I/O devices easier to migrate, direct device assignment of physical devices to VMs may be difficult to migrate



Memory Migration - Steps

- **Push** - Source VM continues running while certain pages are pushed across the network to the new destination. To ensure consistency, pages modified during this process must be re-sent.
 - **Stop-and-copy** - Source VM is stopped, pages are copied across to the destination VM, then the new VM is started.
 - **Pull** - The new VM executes and, if it accesses a page that has not yet been copied, this page is faulted in ("pulled") across the network from the source VM.
- Pure Stop-and-Copy
- Simple but both downtime and total migration time are proportional to the amount of physical memory allocated to the VM.
 - May lead to an unacceptable outage if the VM is running a live service.

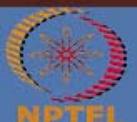
NPTEL



Live Migration - Phases

- **Pre-Copy Phase:** It is carried out over several rounds. The VM continues to execute at the source, while its memory is copied to the destination.
- **Pre-copy Termination Phase:** Stopping criteria of Pre-Copy phase takes one of the following thresholds into account: (i) The number of rounds exceeds a threshold. (ii) The total memory transmitted exceeds a threshold. (iii) The number of dirtied pages in the previous round drops below a threshold.
- **Stop-and-Copy Phase:** In this phase, execution of the VM to be migrated is suspended at the source. Then, the remaining dirty pages and, state of the CPU is copied to the destination host, where the execution of VM is resumed.

NPTEL



Iterative Pre Copy Live Memory Migration

- **Pre-copy Phase:**
 - This phase may be carried out over several rounds.
 - The VM continues to execute at the source host, while its memory is copied to the destination host.
 - Active pages of the VM to be migrated are copied iteratively in each round.
 - During the copying process some active page might get dirtied at the source host, which are again resent in the subsequent rounds to ensure memory consistency.
- **Pre copy-termination phase:** Stopping criteria - options
 - Number of rounds exceeds a threshold.
 - Total memory transmitted exceeds a threshold.
 - Number of dirtied pages in the previous round drops below a threshold.
- **Stop-and-Copy Phase:**
 - In this phase, execution of the VM to be migrated is suspended at the source.
 - Then, the remaining dirty pages and, state of the CPU is copied to the destination host, where the execution of VM is resumed.
- **Restarting Phase:** Restart the VM on destination server.

NPTEL



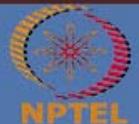
Post-copy Live Memory Migration

- **Stop Phase:** Stop the source VM and copy the CPU state to the destination VM.
 - **Restart Phase:** Restart the destination VM.
 - **On-demand Copy:** Copy the VM memory according to the demand.
- *In the post-copy strategy, when the VM is restarted, the VM memory is empty. If the VM tries to access a memory page that has not yet been copied, this memory page needs to be brought from the source VM. However, most of the time, some memory pages will not be used, so we only need to copy the VM memory according to the demand.*



REFERENCES

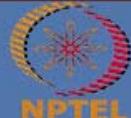
- Kai Hwang, Geoffrey C. Fox, Jack J. Dongarra, Distributed and Cloud Computing - From Parallel Processing to the Internet of Things, Morgan Kaufmann, Elsevier, 2012
- Christian Limpach, Ian Pratt, Christopher Clark, Keir Fraser, Steven Hand, Jacob Gorm Hansen, Eric Jul, Andrew Warfield, Live Migration of Virtual Machines
- Michael R. Hines and Kartik Gopalan, Post-Copy Based Live Virtual Machine Migration Using Adaptive Pre-Paging and Dynamic Self-Ballooning



*Thank
you*



NPTEL





NPTEL ONLINE CERTIFICATION COURSES

Cloud Computing

Prof. Soumya K Ghosh

**Department of Computer Science
and Engineering**

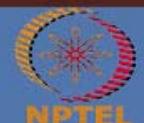
Module 10: Cloud Computing Paradigm

Lecture 46: Cloud Migration - II

CONCEPTS COVERED

- VM Migration - Basics
- Migration strategies

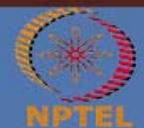
NPTEL



KEYWORDS

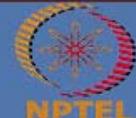
- Virtual Machine (VM)
- VM Migration

NPTEL



VM Migration (contd.)

NPTEL



VM Migration

- VM Migration – It is process to move running applications or VMs from one physical server/ host to another host.
- Processor state, storage, memory and network connection are moved from one host to another host
- Why to migrate VMs?
 - Distribute VM load efficiently across servers in a cloud
 - System maintenance

NPTEL



VM Live Migration – Requirements

- **Load Balancing:** When the load is considerably unbalanced and impending downtime often require simultaneous VM (s) migration.
- **Fault tolerance:** Fault is another challenge to guarantee the critical service availability and reliability. Failures should be anticipated and proactively handled.
- **Power management:** Switching the idle mode server to either sleep mode or off mode based on resource demands, that leads to energy saving. VM live migration is a good technique for cloud power efficiency.
- **Resource sharing:** Challenge of limited hardware resources like memory, cache, and CPU cycles can be solved by relocating VM's from over-loaded server to under-loaded server.
- **System maintenance:** Physical system required to be upgraded and serviced, so VMs of that physical server must be migrated to an alternate server so that services are available to users without interruption



Live Migration – Pre-copy Approach

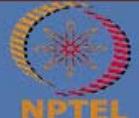
- Uses iterative push phase that is followed by stop-and-copy phase.
- Because of iterative procedure, some memory pages have been updated/modified, called dirty pages are regenerated on the source server during migration iterations.
- Dirty pages resend to the destination host in a future iteration, hence some of the or frequently access memory pages are sent several times. It causes long migration time.
- In the **first phase**, all pages are transferred while VM running continuously on the source host. Further round(s), dirty pages are re-sent.

NPTEL

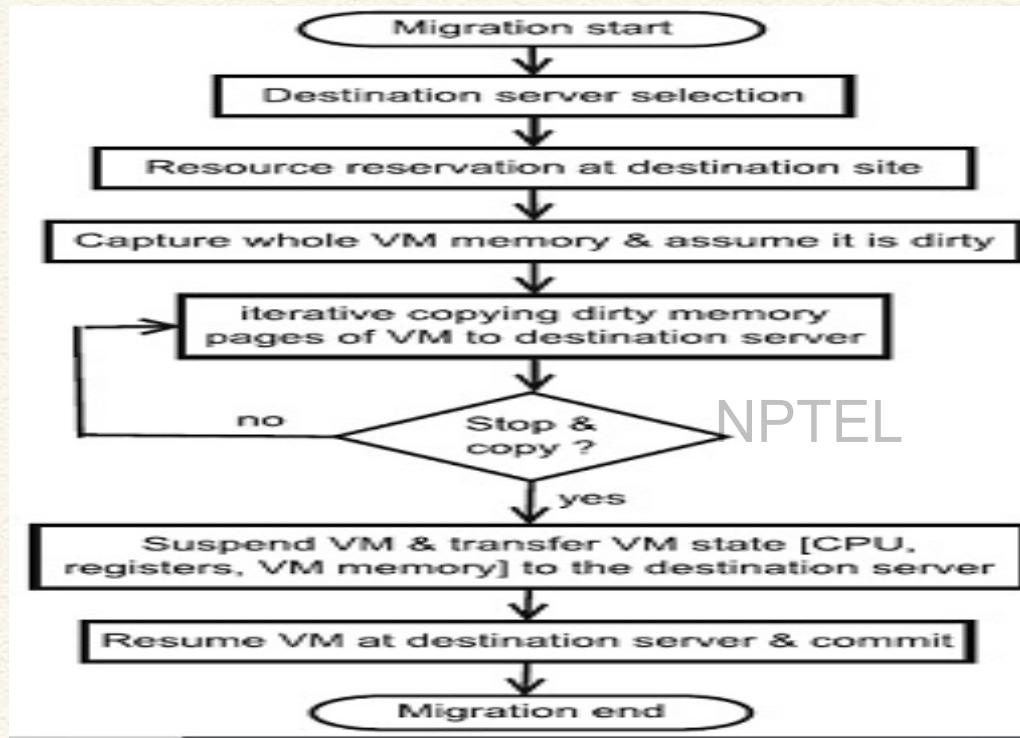


Live Migration – Pre-copy Approach (contd.)

- **Second phase** is termination phase which depends on the defined threshold. The termination is executed if any one out of three conditions: (i) the number of iterations exceeds pre-defined iterations, or (ii) the total amount of memory that has been sent or (iii) the number of dirty pages in just previous round fall below the defined threshold.
- In the last, **stops-and-copy phase**, migrating VM is suspended at source server, after that move processors state and remaining dirty pages.
- When VM migration process is completed in the correct way then hypervisor resumes migrant VM on the destination server.
- KVM, Xen, and VMware hypervisor use the pre-copy technique for live VM migration.



Live Migration – Pre-copy Approach



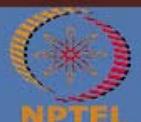
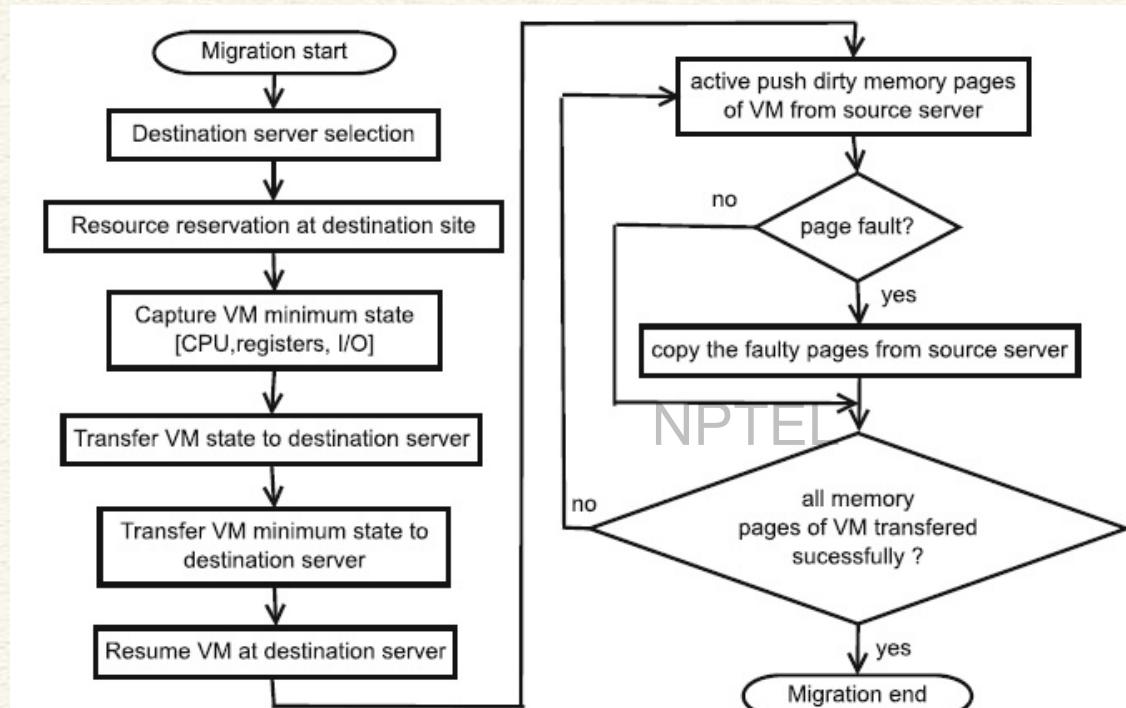
Live Migration – Post-copy Approach

- In post-copy migration technique, processor state transfer before memory content and then
 - VM could be started at the destination server.
 - Post-copy VM migration technique investigates demand paging, active push, pre-paging etc. approaches for prefetching of memory pages at the destination server.
- ✓ **Stop Phase:** Stop the source VM and copy the CPU state to the destination VM.
- ✓ **Restart Phase:** Restart the destination VM.
- ✓ **On-demand Copy:** Copy the VM memory according to the demand.

NPTEL



Live Migration – Post-copy Approach



VM Migration – Analysis

- Let T_{mig} be the total migration time.
- Let T_{down} be the total down time.
- For non-live migration of a single VM the migration time T_{mig} can be calculated as follows:

$$T_{\text{mig}} = V_m / R .$$

where, V_m is the size i.e. memory of the VM and R is the transmission rate.

- In non-live migration, down time is same as the migration time because the services of the VM is suspended during the entire migration process.

$$T_{\text{down}} = T_{\text{mig}}$$

Note: Transmission rate remains fixed for the entire duration of migration.

NPTEL



VM Migration – Analysis (contd.)

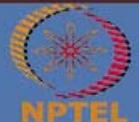
- Let n represent the total number of iterations in the pre copy cycle.
- Let $T_{i,j}$ represents the total time that the j^{th} iterative transmits the i^{th} virtual machine's memory.
- V_m : the memory of a VM.
- V_{th} : threshold for stopping the iterations.
- n_{\max} : maximum number of iterations.
- $r = (P*D)/R$.
where P is page size and D is the dirtying rate, R is the transmission rate.
- T_{res} denotes the time taken to restart the VM on the destination server.



VM Migration – Analysis (contd.)

- Pre-copy migration mechanism: the VMs memory can be migrated iteratively.
- We can compute the total migration time $T_{i,\text{mig}}$ of the i^{th} VM as follows.
- $T_{i,\text{mig}} = \sum_{j=0}^n (T_{i,j}) = \frac{V_m}{R} \left(\frac{1-r^{n+1}}{1-r} \right) + T_{\text{res}}$
- $T_{i,\text{down}} = r^n \left(\frac{V_m}{R} \right) + T_{\text{res}}$

NPTEL



VM Migration – Analysis (contd.)

- Round 0 : $t_0 = \frac{V_m}{R}$
- Round 1: $t_1 = \frac{(P*D)}{R} * t_0 = \frac{(P*D)}{R} * \frac{V_m}{R} = r * \left(\frac{V_m}{R}\right)$
- Round 2: $t_2 = \frac{(P*D)}{R} * t_1 = \frac{(P*D)}{R} * \left(r * \frac{V_m}{R}\right) = r^2 \left(\frac{V_m}{R}\right)$
- Round 3: $t_3 = \frac{(P*D)}{R} * t_2 = \frac{(P*D)}{R} * \left(r^2 * \frac{V_m}{R}\right) = r^3 \left(\frac{V_m}{R}\right)$
-
- Round $n-1$: $t_{n-1} = \frac{(P*D)}{R} * t_{n-2} = \frac{(P*D)}{R} * \left(r^{n-2} * \frac{V_m}{R}\right) = r^{n-1} * \left(\frac{V_m}{R}\right)$
- Round n (Stop and Copy): $t_n = \frac{(P*D)}{R} * t_{n-1} = \frac{(P*D)}{R} \left(r^{n-1} * \frac{V_m}{R}\right) = r^n \left(\frac{V_m}{R}\right)$
- $T = t_0 + t_1 + \dots + t_{n-1} + t_n = \frac{V_m}{R} (1 + r + r^2 + r^3 + \dots + r^{n-1} + r^n) = \frac{V_m}{R} \left(\frac{1 - r^{n+1}}{1 - r}\right)$.

NPTEL



VM Migration – Analysis (contd.)

Estimation of Number of Rounds (n)

- Volume of dirty data to be transferred in round j : $r^j \cdot V_m$
- $r^j \cdot V_m < V_{th}$
- $\Rightarrow j = \lceil \log_r \frac{V_{th}}{V_m} \rceil$
- $n = \min(\lceil \log_r \frac{V_{th}}{V_m} \rceil, n_{max})$

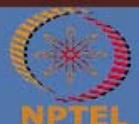
NPTEL



Multiple VMs Migration

- Generally multiple VMs are migrated from a source host to the destination host.
- Typical strategies for migration multiple VMs:
 - Serial Migration.
 - Parallel migration.

NPTEL

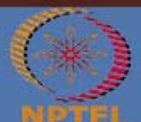


Serial Migration

In case of serial migration of ' m ' correlated VMs of same type the procedure is as follows:

- The first VM that is selected to be migrated executes its pre-copy cycle and the other $(m-1)$ VMs continue to provide services.
- As soon as the first VM enters into the stop and copy phase the remaining $(m-1)$ VMs are suspended and are copied after the first VM completes its stop and copy phase.
- Reason for stopping the remaining $(m-1)$ VMs is to stop those VMs from dirtying memory.
- Assumption: each VM that is copied at full transmission rate (R).
- Downtime for the serial migration includes the stop and copy phase of the first VM, the migration time for the $(m-1)$ VMs and the time to resume the VMs at the destination host.

NPTEL



Serial Migration

- Consider there are 'm' VMs that are to be migrated serially.
- Migration time and downtime for serial migration strategy can be calculated as follows:

$$T_{\text{mig}}^s = \sum_{i=1}^m (T_{i,\text{mig}}) = \frac{m \cdot V_m}{R} \left(\frac{1-r^{n+1}}{1-r} \right) + T_{\text{res}}$$

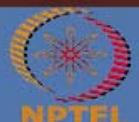
$$T_{\text{down}}^s = \frac{V_m}{R} \cdot r^n + (m - 1) \frac{V_m}{R} \left(\frac{1-r^{n+1}}{1-r} \right) + T_{\text{res}}$$



Parallel Migration

- Major difference between parallel and serial migrations is that all 'm' VMs start their pre-copy cycles simultaneously.
- In fact each VM shares (R/m) of the transmission capacity.
- As the VM sizes are same and transmission rates are same the VMs begin the stop and copy phase at the same time and they end the stop and copy phase also at the same time.
- Since the stop and copy phase is executed in parallel and they consume the same amount of time the downtime is in fact equivalent to the time taken by the stop and copy phase for any VM added to the time taken to resume the VMs at the destination host.

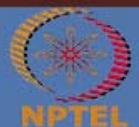
NPTEL



Parallel Migration

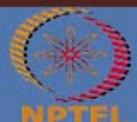
- $T_{mig}^p = \sum_{i=1}^m (T_{i,mig}) = \frac{m.Vm}{R} \left(\frac{1-mr}{1-mr}^{n(p)+1} \right) + T_{res}$
- $T_{down}^p = \frac{m.Vm}{R} \cdot (m.r)^{n(p)} + T_{res}$

NPTEL



REFERENCES

- Kai Hwang, Geoffrey C. Fox, Jack J. Dongarra, Distributed and Cloud Computing - From Parallel Processing to the Internet of Things, Morgan Kaufmann, Elsevier, 2012
- Christian Limpach, Ian Pratt, Christopher Clark, Keir Fraser, Steven Hand, Jacob Gorm Hansen, Eric Jul, Andrew Warfield, Live Migration of Virtual Machines, NSDI, 2005
- Michael R. Hines and Kartik Gopalan, Post-Copy Based Live Virtual Machine Migration Using Adaptive Pre-Paging and Dynamic Self-Ballooning, 2009
- Anita Choudhary, Mahesh Chandra Govil, Girdhari Singh, Lalit K. Awasthi, Emmanuel S. Pilli, Divya Kapil, A critical survey of live virtual machine migration techniques, Journal of Cloud Computing, Springer, 6(23), 2017



*Thank
you*



NPTEL





NPTEL ONLINE CERTIFICATION COURSES

Cloud Computing

Prof. Soumya K Ghosh

**Department of Computer Science
and Engineering**

Module 10: Cloud Computing Paradigm

Lecture 47: Container based Virtualization - I

CONCEPTS COVERED

- Containers
- Container based Virtualization
- Kubernetes
- Docker Container

NPTEL



KEYWORDS

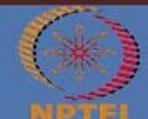
- Container
- Virtualization

NPTEL



Containers

NPTEL



Containers - Introduction

- Virtualization helps to share resources among many customers in cloud computing.
- Container is a lightweight virtualization technique.
- Container packages the code and all its dependencies so the application runs quickly and reliably from one computing environment to another.
- *Docker* is an open platform for developing, shipping, and running applications.
- *Kubernetes* is an open-source system for automating deployment, scaling, and management of containerized applications.

NPTEL



Containers - Introduction

- Containers are packages of software that contain all of the necessary elements to run in any environment.
- Containers virtualize the operating system and run anywhere, from a private data center to the public cloud or even on a developer's personal laptop.
- *Containers are lightweight packages of the application code together with dependencies such as specific versions of programming language runtimes and libraries required to run the software services.*
- Containers make it easy to share CPU, memory, storage, and network resources at the operating systems level and offer a logical packaging mechanism in which applications can be abstracted from the environment in which they actually run.

Ref: <https://cloud.google.com/learn/what-are-containers>



Containers - Needs

- Containers offer a logical packaging mechanism in which applications can be abstracted from the environment in which they actually run.
- This decoupling allows container-based applications to be deployed easily and consistently, regardless of whether the target environment is a private data center, the public cloud, or even a developer's personal laptop.
- *Agile development*: Containers allow the developers to move much more quickly by avoiding concerns about dependencies and environments.
- *Efficient operations*: Containers are lightweight and allow to use just the computing resources one need – thus running the applications efficiently.
- *Run anywhere*: Containers are able to run virtually anywhere. .

Ref: <https://cloud.google.com/learn/what-are-containers>



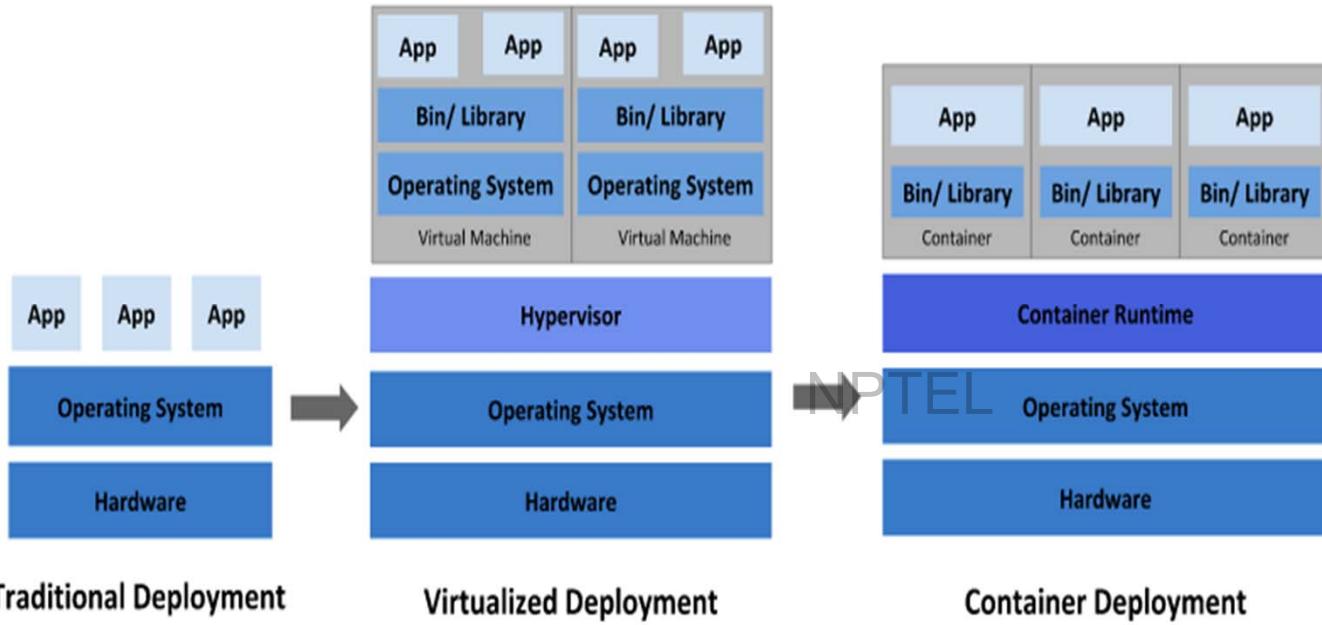
Containers – Major Benefits

- **Separation of responsibility:** Containerization provides a clear separation of responsibility, as developers focus on application logic and dependencies, while IT operations teams can focus on deployment and management instead of application details such as specific software versions and configurations.
- **Workload portability:** Containers can run virtually anywhere, greatly easing development and deployment: on Linux, Windows, and Mac operating systems; on virtual machines or on physical servers; on a developer's machine or in data centers on-premises; and of course, in the public cloud.
- **Application isolation:** Containers virtualize CPU, memory, storage, and network resources at the operating system level, providing developers with a view of the OS logically isolated from other applications.

Ref: <https://cloud.google.com/learn/what-are-containers>



Application Deployment



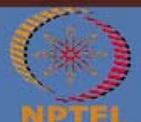
Ref: <https://kubernetes.io/docs/concepts/overview/what-is-kubernetes/>



Traditional – Virtualized – Container Deployments

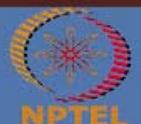
- **Traditional deployment :** Applications run on physical servers. There was no way to define resource boundaries for applications in a physical server, and this caused resource allocation issues.
- **Virtualized deployment :** Allows to run multiple Virtual Machines (VMs) on a single physical server's CPU. Virtualization allows applications to be isolated between VMs. It allows better utilization of resources in a physical server and allows better scalability. Each VM is a full machine running all the components, including its own operating system, on top of the virtualized hardware.
- **Container deployment:** Containers are similar to VMs, but they have relaxed isolation properties to share the Operating System (OS) among the applications. Therefore, containers are considered lightweight. A container has its own filesystem, share of CPU, memory, process space, and more. As containers are decoupled from the underlying infrastructure, they are portable across clouds and different OS distributions.

NPTEL



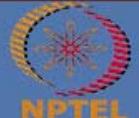
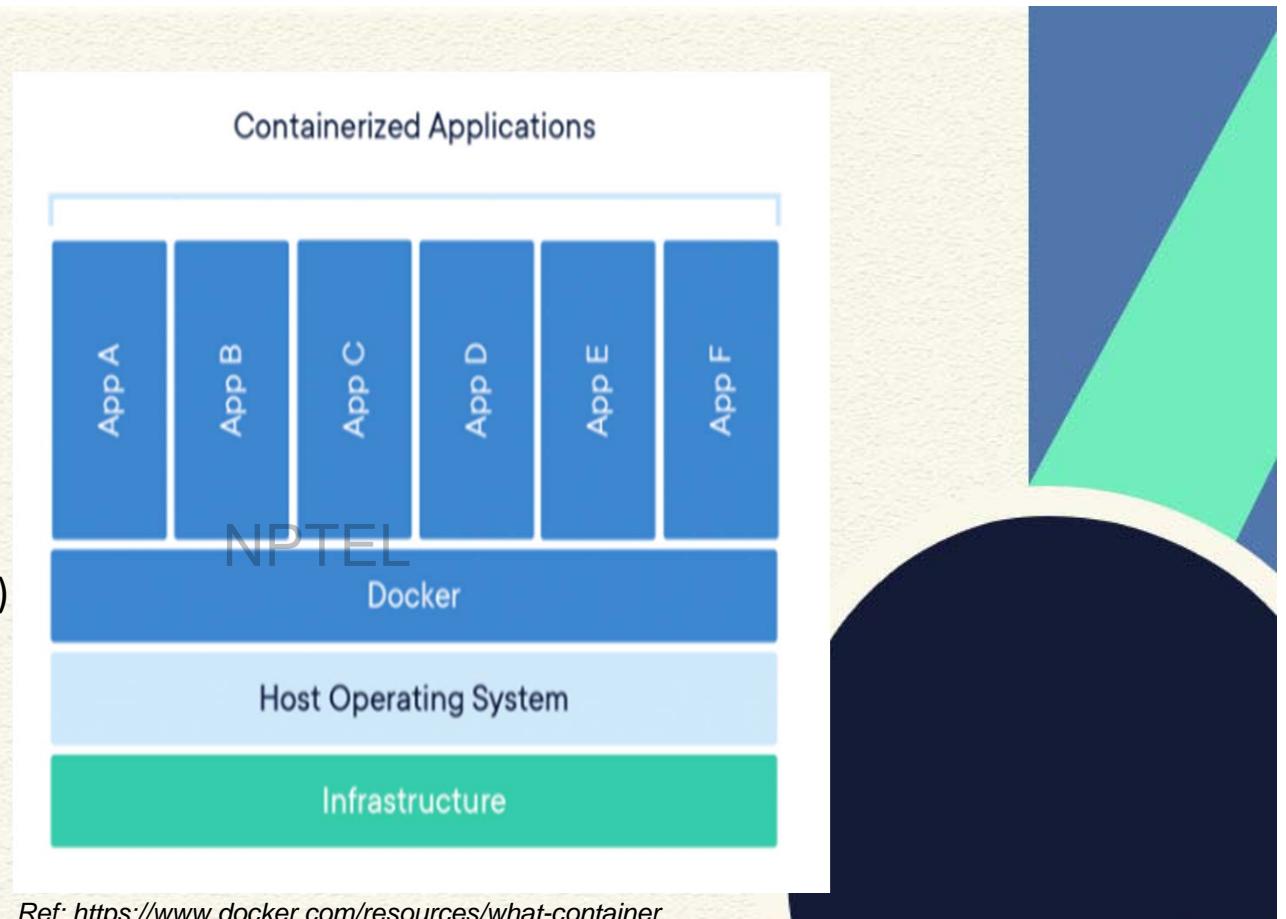
Containers and VMs

- VMs: a guest operating system such as Linux or Windows runs on top of a host operating system with access to the underlying hardware.
- Containers are often compared to virtual machines (VMs). Like virtual machines, containers allow one to package the application together with libraries and other dependencies, providing isolated environments for running your software services.
- However, the containers offer a far more lightweight unit for developers and IT Ops teams to work with, carrying a myriad of benefits.
 - Containers are much more lightweight than VMs
 - Containers virtualize at the OS level while VMs virtualize at the hardware level
 - Containers share the OS kernel and use a fraction of the memory VMs require



Container

- A container is a sandboxed process that is isolated from all other processes on the host machine.
- A container is a runnable instance of an image.
- One can create, start, stop, move, or delete a container using the API (e.g. DockerAPI) or CLI.
- A container can be run on local machines, virtual machines or deployed to the cloud.



Kubernetes

- Kubernetes is a portable, extensible, open-source platform for managing containerized workloads and services, that facilitates both declarative configuration and automation. It has a large, rapidly growing ecosystem. Kubernetes services, support, and tools are widely available.
- The name Kubernetes originates from Greek, meaning helmsman or pilot.
- Kubernetes operates at the container level rather than at the hardware level, it provides some generally applicable features common to PaaS offerings, such as deployment, scaling, load balancing, and lets users integrate their logging, monitoring, and alerting solutions.
- However, Kubernetes is not monolithic, and these default solutions are optional and pluggable.
- Kubernetes provides the building blocks for building developer platforms, but preserves user choice and flexibility where it is important.

Ref: <https://kubernetes.io/docs/concepts/overview/what-is-kubernetes/>



Kubernetes Components

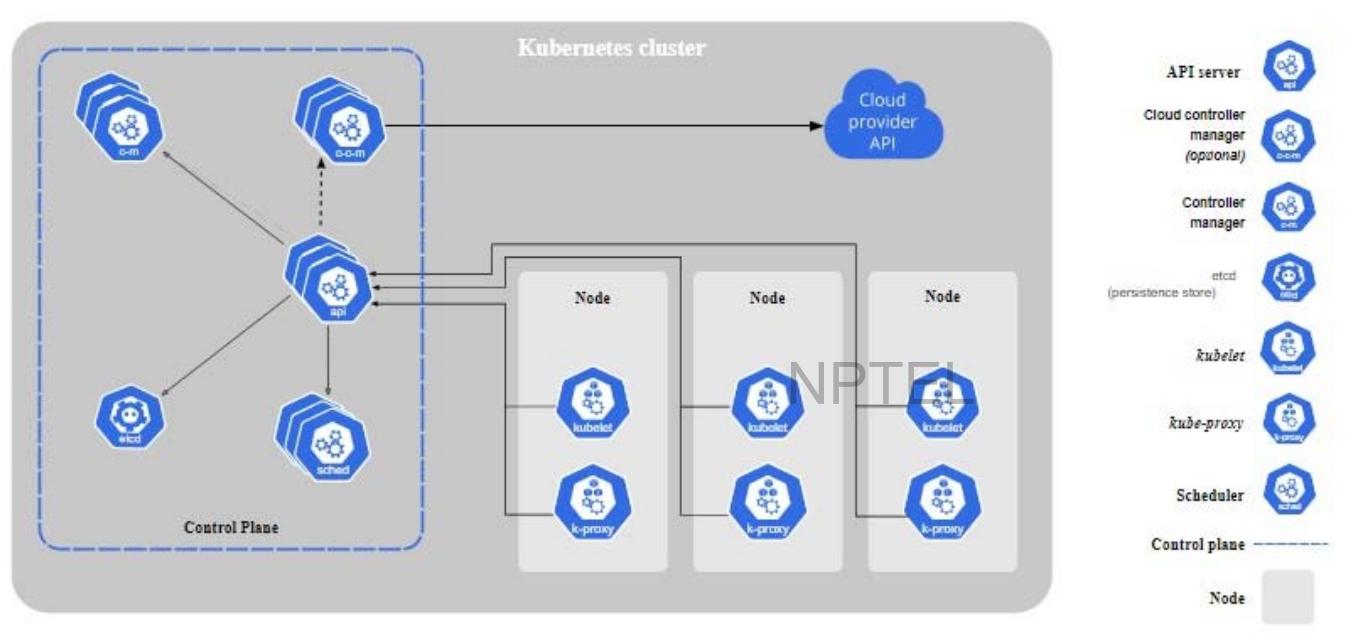
- A Kubernetes cluster consists of a set of worker machines, called **nodes**, that run containerized applications. Every cluster has at least one worker node.
- The worker node(s) host the **Pods** that are the components of the application workload.
- The **control plane** manages the worker nodes and the Pods in the cluster.
- In production environments, the control plane usually runs across multiple computers and a cluster usually runs multiple nodes, providing fault-tolerance and high availability.

NPTEL

Ref: <https://kubernetes.io/docs/concepts/overview/what-is-kubernetes/>



Kubernetes Cluster Components



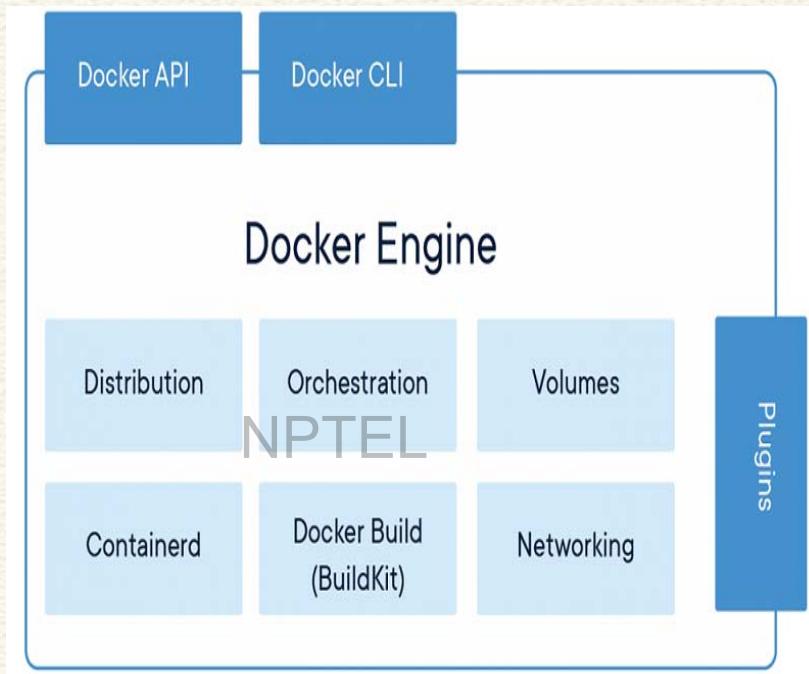
Ref: <https://kubernetes.io/docs/concepts/overview/what-is-kubernetes/>



Docker Engine

- Docker containers that run on Docker Engine:
- **Standard:** Docker created the industry standard for containers, so they could be portable anywhere
- **Lightweight:** Containers share the machine's OS system kernel and therefore do not require an OS per application, driving higher server efficiencies and reducing server and licensing costs
- **Secure:** Applications are safer in containers and Docker provides the strongest default isolation capabilities in the industry

Ref: <https://www.docker.com/resources/what-container>



Dockers

- A Docker container image is a lightweight, standalone, executable package of software that includes everything needed to run an application: code, runtime, system tools, system libraries and settings.
- Container images become containers at runtime and in the case of Docker containers - images become containers when they run on [Docker Engine](#).
- Available for both Linux and Windows-based applications, containerized software will always run the same, regardless of the infrastructure. Containers isolate software from its environment and ensure that it works uniformly despite differences for instance between development and staging.

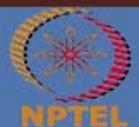
Ref: <https://www.docker.com/resources/what-container>



REFERENCES

- <https://kubernetes.io/docs/concepts/overview/what-is-kubernetes/>
- <https://www.docker.com/resources/what-container>
- <https://cloud.google.com/kubernetes-engine/docs/concepts/verticalpodautoscaler>

NPTEL



*Thank
you*



NPTEL





NPTEL ONLINE CERTIFICATION COURSES

Cloud Computing

Prof. Soumya K Ghosh

**Department of Computer Science
and Engineering**

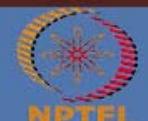
Module 10: Cloud Computing Paradigm

Lecture 48: Container - II (Docker)

CONCEPTS COVERED

- Docker Container – Overview
- Docker – Components
- Docker – Architecture

NPTEL



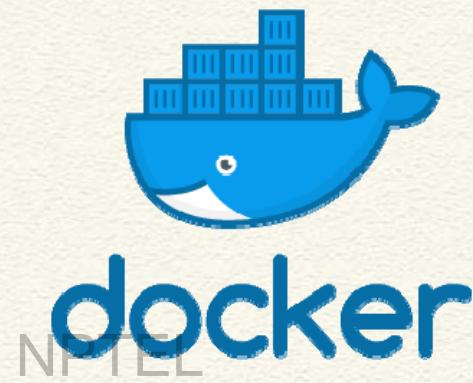
KEYWORDS

- Container
- Docker Container

NPTEL



Docker



<https://www.docker.com/>



Docker - Overview

- Docker is a platform that allows you to “build, ship, and run any app, anywhere.”
- Considered to be a standard way of solving one of the challenging aspects of software: deployment.
- Traditionally, the development pipeline typically involved combinations of various technologies for managing the movement of software, such as virtual machines, configuration management tools, package management systems, and complex webs of library dependencies.
 - All these tools needed to be managed and maintained by specialist engineers, and most had their own unique ways of being configured.

Ref: *Docker in Practice, Second Edition, Ian Miell and Aidan Hobson Sayers, February 2019, ISBN 9781617294808*



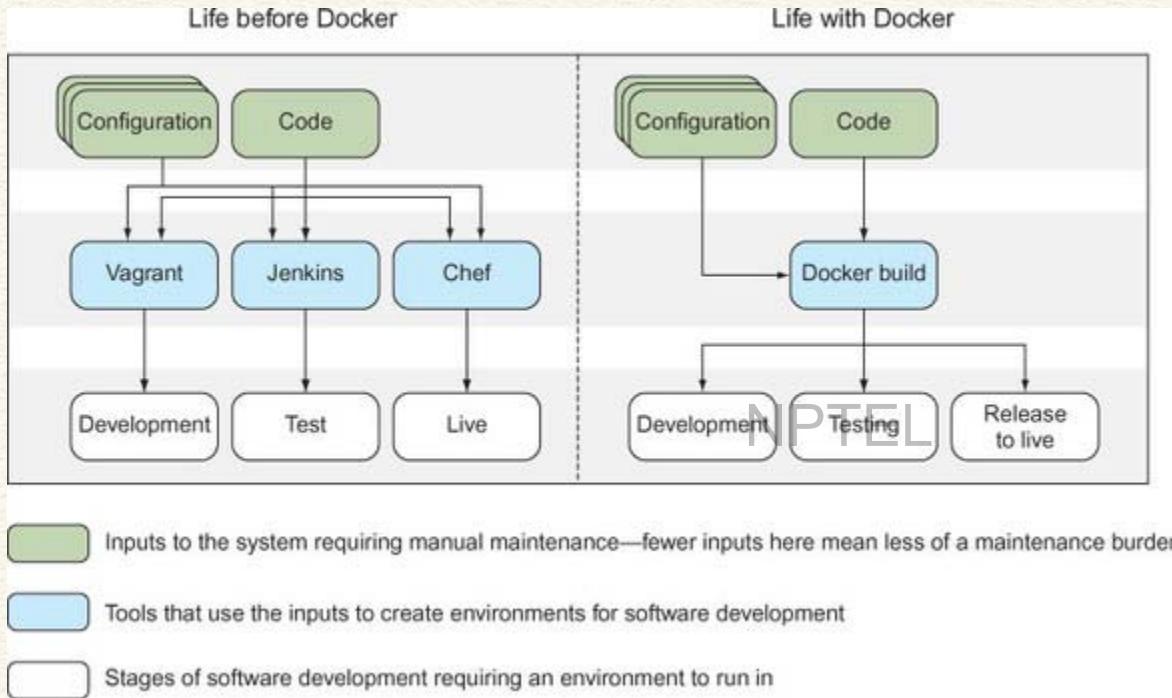
Docker - Overview

- Docker has changed the traditional approach - Everything goes through a common pipeline to a single output that can be used on any target—there's no need to continue maintaining a bewildering array of tool configurations
- At the same time, there's no need to throw away the existing software stack if it works for you—you can package it up in a Docker container as-is, for others to consume.
- Additionally, you can see how these containers were built, so if you need to dig into the details, you can.

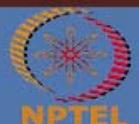
Ref: *Docker in Practice, Second Edition, Ian Miell and Aidan Hobson Sayers, February 2019, ISBN 9781617294808*



Docker – Big Picture



Ref: *Docker in Practice, Second Edition, Ian Miell and Aidan Hobson Sayers, February 2019, ISBN 9781617294808*



Docker - Analogy

- *Analogy:* Tradionally, a docker was a laborer who moved commercial goods into and out of ships when they docked at ports. There were boxes and items of differing sizes and shapes, and experienced dockers were prized for their ability to fit goods into ships by hand in cost-effective ways. Hiring people to move stuff around wasn't cheap, but there was no alternative!
- This may sound familiar to anyone working in software. Much time and intellectual energy is spent getting metaphorically odd-shaped software into differently-sized metaphorical ships full of other odd-shaped software, so they can be sold to users or businesses elsewhere.

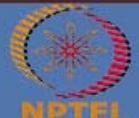
Ref: Docker in Practice, Second Edition, Ian Miell and Aidan Hobson Sayers, February 2019, ISBN 9781617294808



Docker - Benefit

- Before Docker, deploying software to different environments required significant effort. Even if you weren't hand-running scripts to provision software on different machines (and plenty of people do exactly that), you'd still have to handle the configuration management tools that manage state on what are increasingly fast-moving environments starved of resources.
- Even when these efforts were encapsulated in VMs, a lot of time was spent managing the deployment of these VMs, waiting for them to boot, and managing the overhead of resource use they created.

Ref: *Docker in Practice, Second Edition, Ian Miell and Aidan Hobson Sayers, February 2019, ISBN 9781617294808*



Docker - Benefit

Three times the effort to manage deployment

Life before Docker

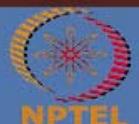
Install, configure, and maintain complex application
↓
Dev laptop Test server Live server

A single effort to manage deployment

Life with Docker

Install, configure, and maintain complex application
↓
Docker image
docker run → Dev laptop
docker run → Test server
docker run → Live server

Ref: Docker in Practice, Second Edition, Ian Miell and Aidan Hobson Sayers, February 2019, ISBN 9781617294808



Docker - Benefit

- With Docker, the configuration effort is separated from the resource management, and the deployment effort is trivial:
 - run docker, and the environment's image is pulled down and ready to run, consuming fewer resources and contained so that it doesn't interfere with other environments.
- You don't need to worry about whether your container is going to be shipped to a Red Hat machine, an Ubuntu machine, or a CentOS VM image; as long as it has Docker on it, it will run

Ref: Docker in Practice, Second Edition, Ian Miell and Aidan Hobson Sayers, February 2019, ISBN 9781617294808



Docker - Advantage

- **Replacing virtual machines (VMs):** Docker can be used to replace VMs in many situations. If you only care about the application, not the operating system, Docker can replace the VM.
 - Not only is Docker quicker than a VM to spin up, it's more lightweight to move around, and due to its layered filesystem, you can more easily and quickly share changes with others. It's also rooted in the command line and is scriptable.
- **Prototyping software:** If you want to quickly experiment with software without either disrupting your existing setup or going through the hassle of provisioning a VM, Docker can give you a sandbox environment almost instantly. .

Ref: *Docker in Practice, Second Edition, Ian Miell and Aidan Hobson Sayers, February 2019, ISBN 9781617294808*

NPTEL



Docker - Advantage

- **Packaging software:** Because a Docker image has effectively no dependencies, it's a great way to package software. You can build your image and be sure that it can run on any modern Linux machine—think Java, without the need for a JVM.
- **Enabling a Microservices architecture:** Docker facilitates the decomposition of a complex system to a series of composable parts, which allows you to reason about your services in a more discrete way. This can allow you to restructure your software to make its parts more manageable and pluggable without affecting the whole.

Ref: *Docker in Practice, Second Edition, Ian Miell and Aidan Hobson Sayers, February 2019, ISBN 9781617294808*



Docker - Advantage

- **Modeling networks:** Several hundreds (even thousands) of isolated containers can be initiated on one machine, modeling a network can be done efficiently. .
- **Enabling full-stack productivity when offline** - All the parts of the system can be bundled into Docker containers, you can orchestrate these to run on your laptop and work on the move, even when offline.

Ref: *Docker in Practice, Second Edition, Ian Miell and Aidan Hobson Sayers, February 2019, ISBN 9781617294808*



Docker - Advantage

- **Reducing debugging overhead:** Complex negotiations between different teams about software delivered is a commonplace within the industry.
 - Docker allows you to state clearly (even in script form) the steps for debugging a problem on a system with known properties, making bug and environment reproduction a much simpler affair, and one normally separated from the host environment provided.
- **Documenting software dependencies:** By building your images in a structured way, ready to be moved to different environments, Docker forces you to document your software dependencies explicitly from a base starting point..

Ref: *Docker in Practice, Second Edition, Ian Miell and Aidan Hobson Sayers, February 2019, ISBN 9781617294808*



Docker - Advantage

- **Enabling continuous delivery:** Continuous delivery (CD) is a paradigm for software delivery based on a pipeline that rebuilds the system on every change and then delivers to production (or “live”) through an automated (or partly automated) process.
- Docker builds are more reproducible and replicable than traditional software building methods. This makes implementing Continuous delivery (CD) much easier.

Ref: *Docker in Practice, Second Edition, Ian Miell and Aidan Hobson Sayers, February 2019, ISBN 9781617294808*

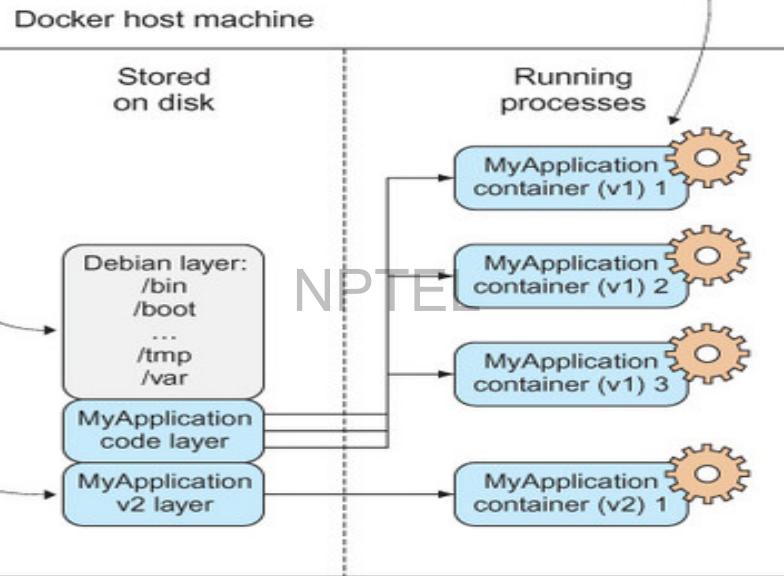


Docker – Key Concepts

Images: An image is a collection of filesystem layers and some metadata. Taken together, they can be spun up as Docker containers.

Layers: A layer is a collection of changes to files. The differences between v1 and v2 of MyApplication are stored in this layer.

Containers: A container is a running instance of an image. You can have multiple containers running from the same image.



Ref: Docker in Practice, Second Edition, Ian Miell and Aidan Hobson Sayers, February 2019, ISBN 9781617294808



Docker – Key Commands

Command	Purpose
docker build	Build a Docker image
docker run	Run a Docker image as a container
docker commit	Commit a Docker container as an image
docker tag	Tag a Docker image

Ref: *Docker in Practice, Second Edition, Ian Miell and Aidan Hobson Sayers, February 2019, ISBN 9781617294808*



Docker – Architecture

- Docker on your host machine is split into two parts—a daemon with a RESTful API and a client that talks to the daemon.
- The private Docker registry is a service that stores Docker images. These can be requested from any Docker daemon that has the relevant access. This registry is on an internal network and isn't publicly accessible, so it's considered private.

Ref: *Docker in Practice, Second Edition, Ian Miell and Aidan Hobson Sayers, February 2019, ISBN 9781617294808*



Docker – Architecture

- One invokes the Docker client to get information from or give instructions to the daemon; the daemon is a server that receives requests and returns responses from the client using the HTTP protocol.
- In turn, it will make requests to other services to send and receive images, also using the HTTP protocol.
- The server will accept requests from the command-line client or anyone else authorized to connect.
- 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.

NPTEL

Ref: Docker in Practice, Second Edition, Ian Miell and Aidan Hobson Sayers, February 2019, ISBN 9781617294808



Docker – Architecture

Your host machine, on which you've installed Docker. The host machine will typically sit on a private network.

You invoke the Docker client program to get information from or give instructions to the Docker daemon.

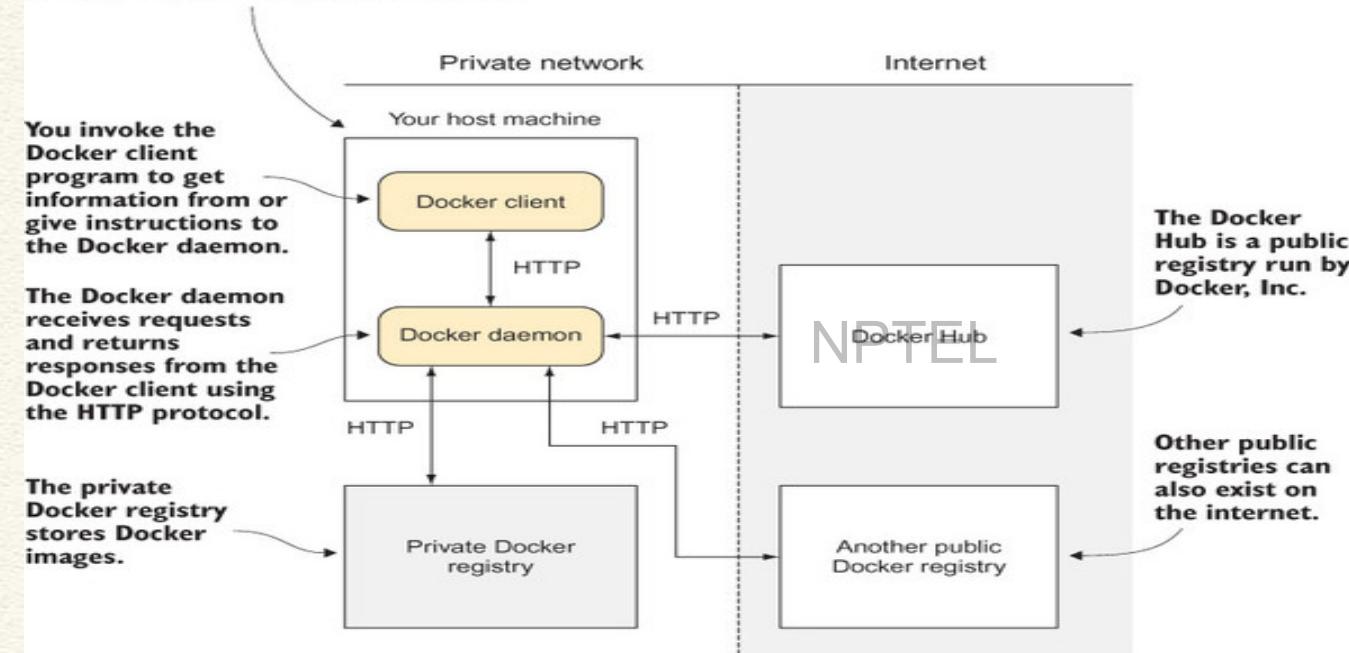
The Docker daemon receives requests and returns responses from the Docker client using the HTTP protocol.

The private Docker registry stores Docker images.

Internet

The Docker Hub is a public registry run by Docker, Inc.

Other public registries can also exist on the internet.



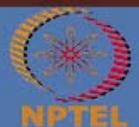
Ref: Docker in Practice, Second Edition, Ian Miell and Aidan Hobson Sayers, February 2019, ISBN 9781617294808



REFERENCES

- <https://www.docker.com/>
- Docker in Practice, Second Edition, Ian Miell and Aidan Hobson Sayers, February 2019, ISBN 9781617294808

NPTEL



*Thank
you*



NPTEL





NPTEL ONLINE CERTIFICATION COURSES

Cloud Computing

Prof. Soumya K Ghosh

**Department of Computer Science
and Engineering**

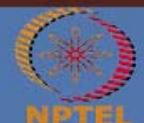
Module 10: Cloud Computing Paradigm

Lecture 49: Docker Container - Demo (Part-I)

CONCEPTS COVERED

- Docker Container - Demo

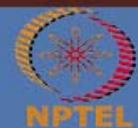
NPTEL



KEYWORDS

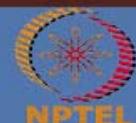
- Container
- Docker

NPTEL



Docker Demo - I

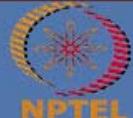
NPTEL



Introduction

- **Containers**
 - Standard unit of software
 - Packages up code and all its dependencies
 - Application runs quickly and reliably
- **Docker container image**
 - Lightweight
 - Standalone
 - Executable package of software
 - Includes everything needed to run an application

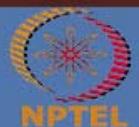
NPTEL



Demo - Objective

- MySQL and [PHPMyAdmin](#) on Docker platform
- MySQL
 - Widely used relational database package
 - Open-source
- PHPMyAdmin
 - A graphical user interface
 - Web-based
 - Connects to MySQL database
 - Widely used for managing MySQL databases

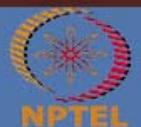
NPTEL



Standalone System (No Container)

- **Separate installation for**
 - MySQL
 - Web Server (Apache)
 - PHP
 - PHPMyAdmin
- **Transferring to other machine/ system**
 - Separate installation
 - Backup of data from old MySQL server
 - Restore the backup to new MySQL server

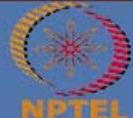
NPTEL



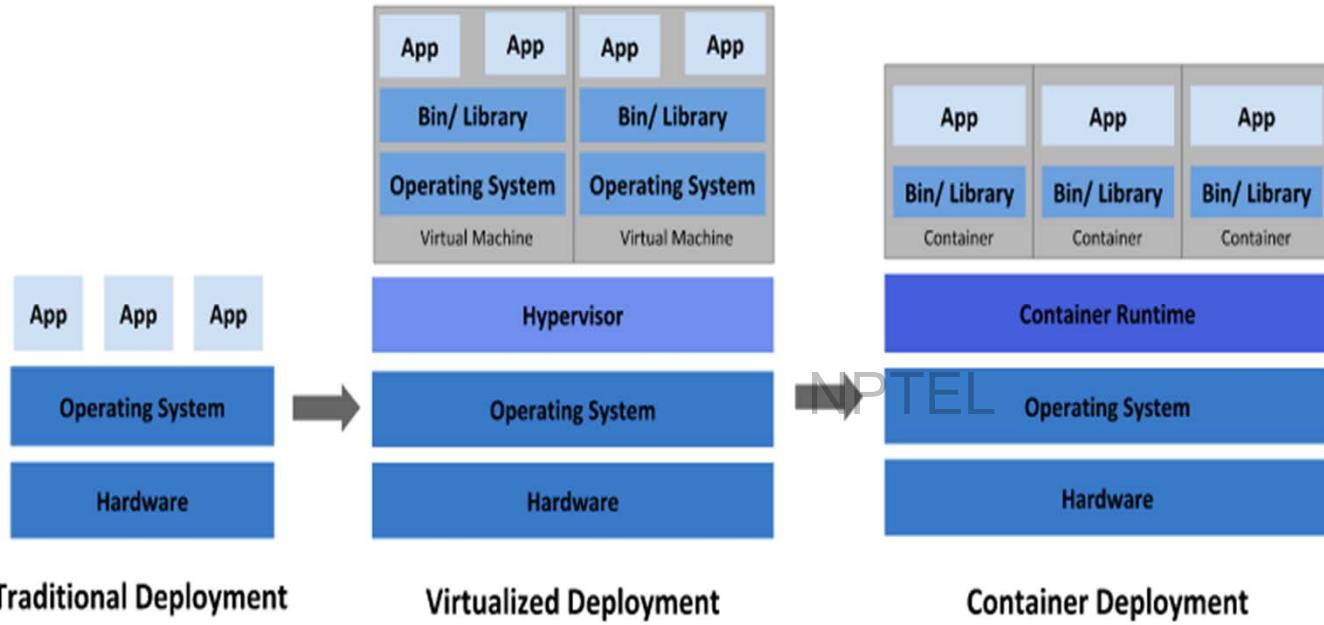
Containers – Major Benefits

- **Separation of responsibility:** Containerization provides a clear separation of responsibility, as developers focus on application logic and dependencies, while IT operations teams can focus on deployment and management instead of application details such as specific software versions and configurations.
- **Workload portability:** Containers can run virtually anywhere, greatly easing development and deployment: on Linux, Windows, and Mac operating systems; on virtual machines or on physical servers; on a developer's machine or in data centers on-premises; and of course, in the public cloud.
- **Application isolation:** Containers virtualize CPU, memory, storage, and network resources at the operating system level, providing developers with a view of the OS logically isolated from other applications.

Ref: <https://cloud.google.com/learn/what-are-containers>



Application Deployment



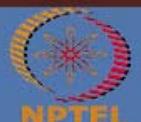
Ref: <https://kubernetes.io/docs/concepts/overview/what-is-kubernetes/>



Traditional – Virtualized – Container Deployments

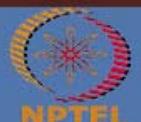
- **Traditional deployment :** Applications run on physical servers. There was no way to define resource boundaries for applications in a physical server, and this caused resource allocation issues.
- **Virtualized deployment :** Allows to run multiple Virtual Machines (VMs) on a single physical server's CPU. Virtualization allows applications to be isolated between VMs. It allows better utilization of resources in a physical server and allows better scalability. Each VM is a full machine running all the components, including its own operating system, on top of the virtualized hardware.
- **Container deployment:** Containers are similar to VMs, but they have relaxed isolation properties to share the Operating System (OS) among the applications. Therefore, containers are considered lightweight. A container has its own filesystem, share of CPU, memory, process space, and more. As containers are decoupled from the underlying infrastructure, they are portable across clouds and different OS distributions.

NPTEL



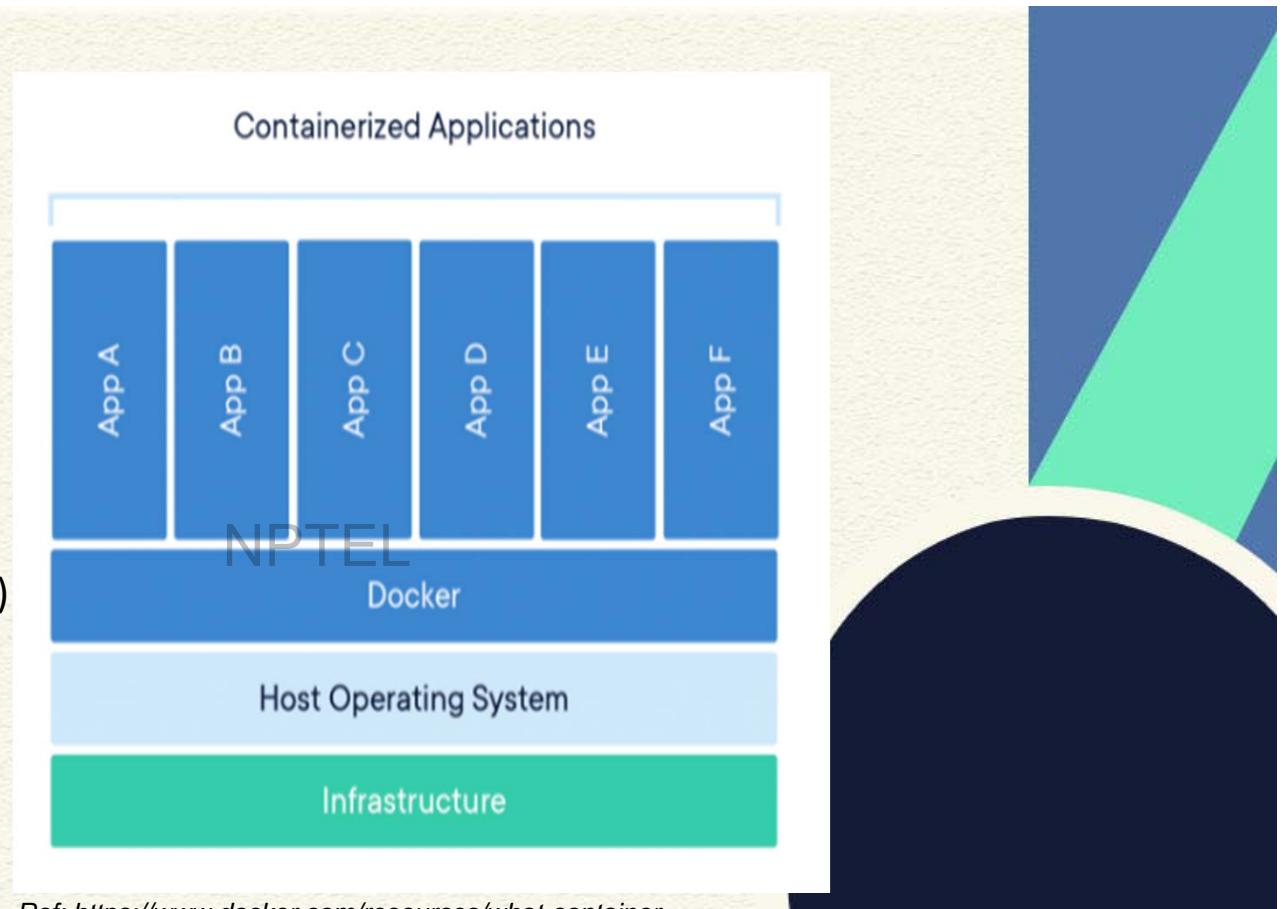
Containers and VMs

- VMs: a guest operating system such as Linux or Windows runs on top of a host operating system with access to the underlying hardware.
- Containers are often compared to virtual machines (VMs). Like virtual machines, containers allow one to package the application together with libraries and other dependencies, providing isolated environments for running your software services.
- However, the containers offer a far more lightweight unit for developers and IT Ops teams to work with, carrying a myriad of benefits.
 - Containers are much more lightweight than VMs
 - Containers virtualize at the OS level while VMs virtualize at the hardware level
 - Containers share the OS kernel and use a fraction of the memory VMs require



Container

- A container is a sandboxed process that is isolated from all other processes on the host machine.
- A container is a runnable instance of an image.
- One can create, start, stop, move, or delete a container using the API (e.g. DockerAPI) or CLI.
- A container can be run on local machines, virtual machines or deployed to the cloud.



Ref: <https://www.docker.com/resources/what-container>



Kubernetes

- Kubernetes is a portable, extensible, open-source platform for managing containerized workloads and services, that facilitates both declarative configuration and automation. It has a large, rapidly growing ecosystem. Kubernetes services, support, and tools are widely available.
- The name Kubernetes originates from Greek, meaning helmsman or pilot.
- Kubernetes operates at the container level rather than at the hardware level, it provides some generally applicable features common to PaaS offerings, such as deployment, scaling, load balancing, and lets users integrate their logging, monitoring, and alerting solutions.
- However, Kubernetes is not monolithic, and these default solutions are optional and pluggable.
- Kubernetes provides the building blocks for building developer platforms, but preserves user choice and flexibility where it is important.

Ref: <https://kubernetes.io/docs/concepts/overview/what-is-kubernetes/>



Kubernetes Components

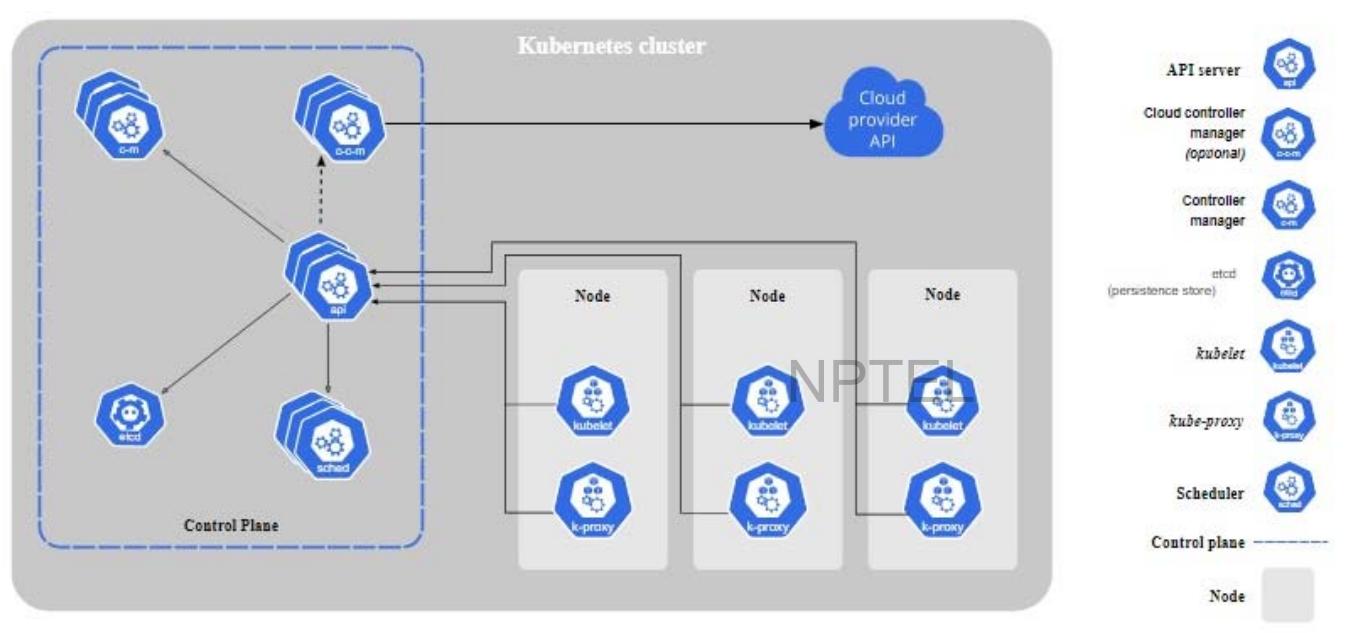
- A Kubernetes cluster consists of a set of worker machines, called **nodes**, that run containerized applications. Every cluster has at least one worker node.
- The worker node(s) host the **Pods** that are the components of the application workload.
- The **control plane** manages the worker nodes and the Pods in the cluster.
- In production environments, the control plane usually runs across multiple computers and a cluster usually runs multiple nodes, providing fault-tolerance and high availability.

NPTEL

Ref: <https://kubernetes.io/docs/concepts/overview/what-is-kubernetes/>



Kubernetes Cluster Components



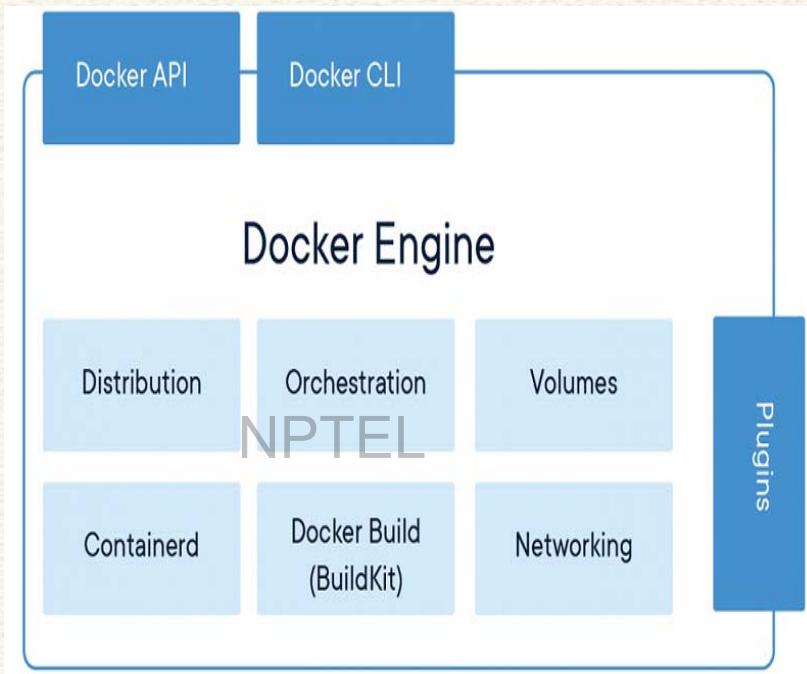
Ref: <https://kubernetes.io/docs/concepts/overview/what-is-kubernetes/>



Docker Engine

- Docker containers that run on Docker Engine:
- **Standard:** Docker created the industry standard for containers, so they could be portable anywhere
- **Lightweight:** Containers share the machine's OS system kernel and therefore do not require an OS per application, driving higher server efficiencies and reducing server and licensing costs
- **Secure:** Applications are safer in containers and Docker provides the strongest default isolation capabilities in the industry

Ref: <https://www.docker.com/resources/what-container>



Dockers

- A Docker container image is a lightweight, standalone, executable package of software that includes everything needed to run an application: code, runtime, system tools, system libraries and settings.
- Container images become containers at runtime and in the case of Docker containers - images become containers when they run on [Docker Engine](#).
- Available for both Linux and Windows-based applications, containerized software will always run the same, regardless of the infrastructure. Containers isolate software from its environment and ensure that it works uniformly despite differences for instance between development and staging.

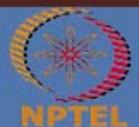
Ref: <https://www.docker.com/resources/what-container>



REFERENCES

- <https://kubernetes.io/docs/concepts/overview/what-is-kubernetes/>
- <https://www.docker.com/resources/what-container>
- <https://cloud.google.com/kubernetes-engine/docs/concepts/verticalpodautoscaler>

NPTEL



*Thank
you*



NPTEL





NPTEL ONLINE CERTIFICATION COURSES

Cloud Computing

Prof. Soumya K Ghosh

**Department of Computer Science
and Engineering**

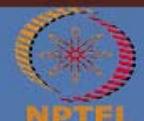
Module 10: Container

Lecture 50: Docker Container - Demo (Part-II)

CONCEPTS COVERED

- Docker Container - Demo

NPTEL



KEYWORDS

- Container
- Docker

NPTEL



Docker Demo - II

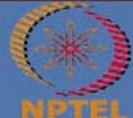
NPTEL



Introduction

- **Containers**
 - Standard unit of software
 - Packages up code and all its dependencies
 - Application runs quickly and reliably
- **Docker container image**
 - Lightweight
 - Standalone
 - Executable package of software
 - Includes everything needed to run an application

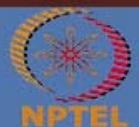
NPTEL



Demo - Objective

- MySQL and [PHPMyAdmin](#) on Docker platform
- MySQL
 - Widely used relational database package
 - Open-source
- PHPMyAdmin
 - A graphical user interface
 - Web-based
 - Connects to MySQL database
 - Widely used for managing MySQL databases

NPTEL



Standalone System (No Container)

- **Separate installation for**
 - MySQL
 - Web Server (Apache)
 - PHP
 - PHPMyAdmin
- **Transferring to other machine/ system**
 - Separate installation
 - Backup of data from old MySQL server
 - Restore the backup to new MySQL server

NPTEL



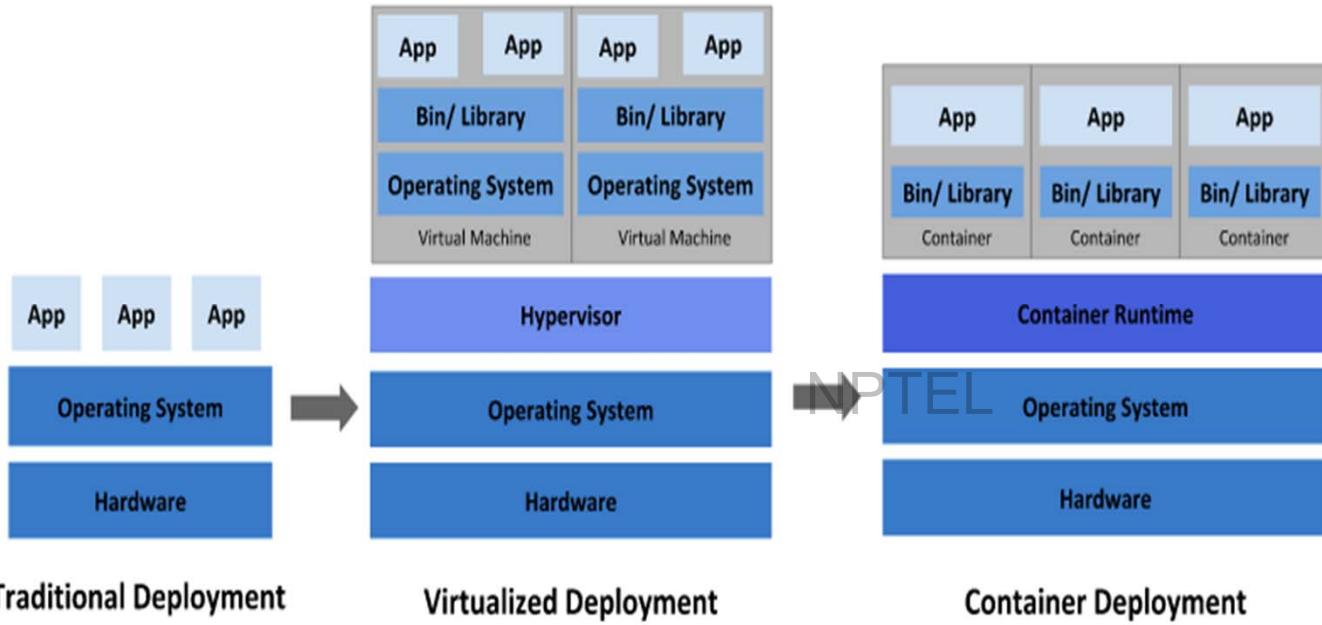
Containers – Major Benefits

- **Separation of responsibility:** Containerization provides a clear separation of responsibility, as developers focus on application logic and dependencies, while IT operations teams can focus on deployment and management instead of application details such as specific software versions and configurations.
- **Workload portability:** Containers can run virtually anywhere, greatly easing development and deployment: on Linux, Windows, and Mac operating systems; on virtual machines or on physical servers; on a developer's machine or in data centers on-premises; and of course, in the public cloud.
- **Application isolation:** Containers virtualize CPU, memory, storage, and network resources at the operating system level, providing developers with a view of the OS logically isolated from other applications.

Ref: <https://cloud.google.com/learn/what-are-containers>



Application Deployment



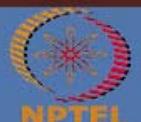
Ref: <https://kubernetes.io/docs/concepts/overview/what-is-kubernetes/>



Traditional – Virtualized – Container Deployments

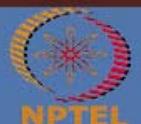
- **Traditional deployment :** Applications run on physical servers. There was no way to define resource boundaries for applications in a physical server, and this caused resource allocation issues.
- **Virtualized deployment :** Allows to run multiple Virtual Machines (VMs) on a single physical server's CPU. Virtualization allows applications to be isolated between VMs. It allows better utilization of resources in a physical server and allows better scalability. Each VM is a full machine running all the components, including its own operating system, on top of the virtualized hardware.
- **Container deployment:** Containers are similar to VMs, but they have relaxed isolation properties to share the Operating System (OS) among the applications. Therefore, containers are considered lightweight. A container has its own filesystem, share of CPU, memory, process space, and more. As containers are decoupled from the underlying infrastructure, they are portable across clouds and different OS distributions.

NPTEL



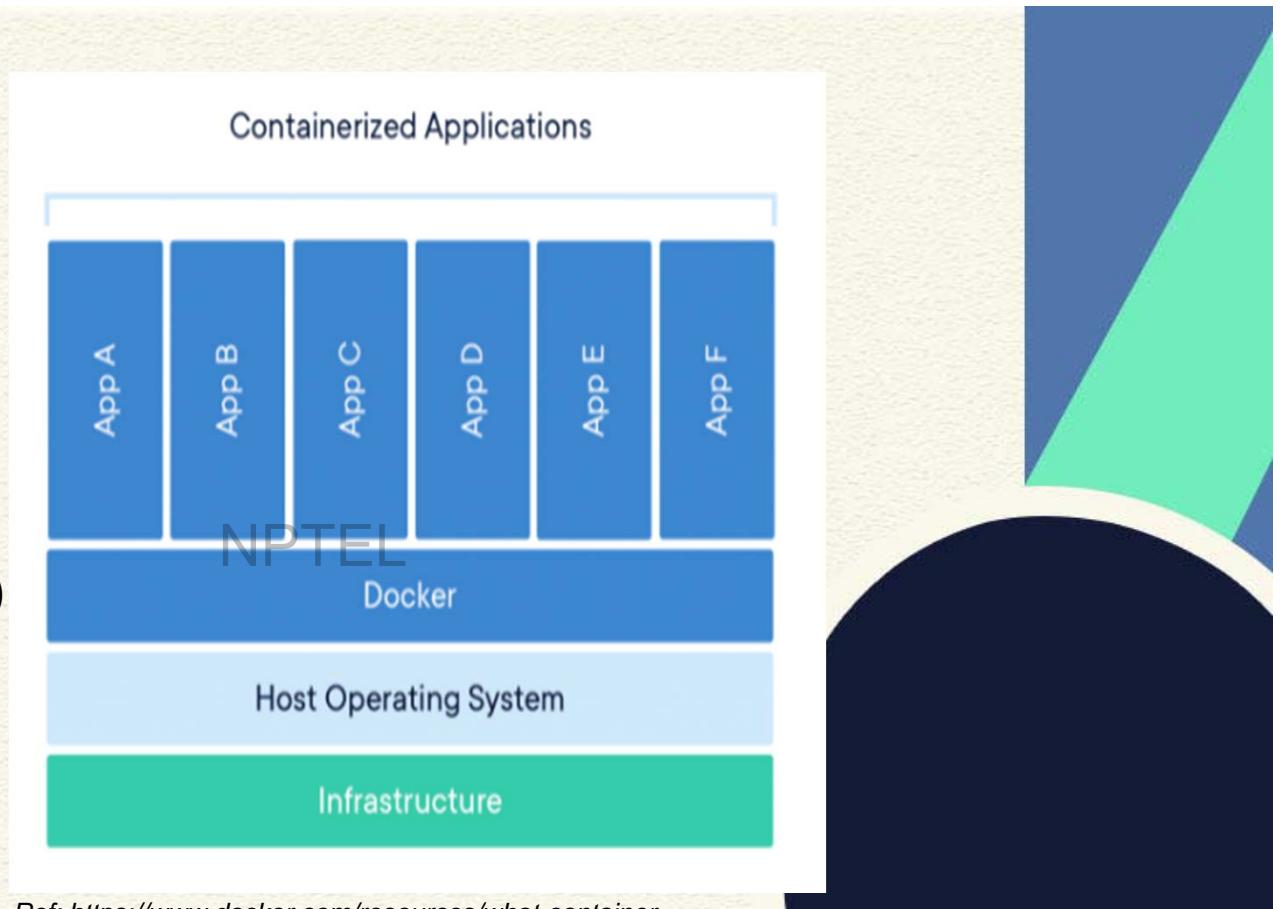
Containers and VMs

- VMs: a guest operating system such as Linux or Windows runs on top of a host operating system with access to the underlying hardware.
- Containers are often compared to virtual machines (VMs). Like virtual machines, containers allow one to package the application together with libraries and other dependencies, providing isolated environments for running your software services.
- However, the containers offer a far more lightweight unit for developers and IT Ops teams to work with, carrying a myriad of benefits.
 - Containers are much more lightweight than VMs
 - Containers virtualize at the OS level while VMs virtualize at the hardware level
 - Containers share the OS kernel and use a fraction of the memory VMs require



Container

- A container is a sandboxed process that is isolated from all other processes on the host machine.
- A container is a runnable instance of an image.
- One can create, start, stop, move, or delete a container using the API (e.g. DockerAPI) or CLI.
- A container can be run on local machines, virtual machines or deployed to the cloud.



Ref: <https://www.docker.com/resources/what-container>



Kubernetes

- Kubernetes is a portable, extensible, open-source platform for managing containerized workloads and services, that facilitates both declarative configuration and automation. It has a large, rapidly growing ecosystem. Kubernetes services, support, and tools are widely available.
- The name Kubernetes originates from Greek, meaning helmsman or pilot.
- Kubernetes operates at the container level rather than at the hardware level, it provides some generally applicable features common to PaaS offerings, such as deployment, scaling, load balancing, and lets users integrate their logging, monitoring, and alerting solutions.
- However, Kubernetes is not monolithic, and these default solutions are optional and pluggable.
- Kubernetes provides the building blocks for building developer platforms, but preserves user choice and flexibility where it is important.

Ref: <https://kubernetes.io/docs/concepts/overview/what-is-kubernetes/>



Kubernetes Components

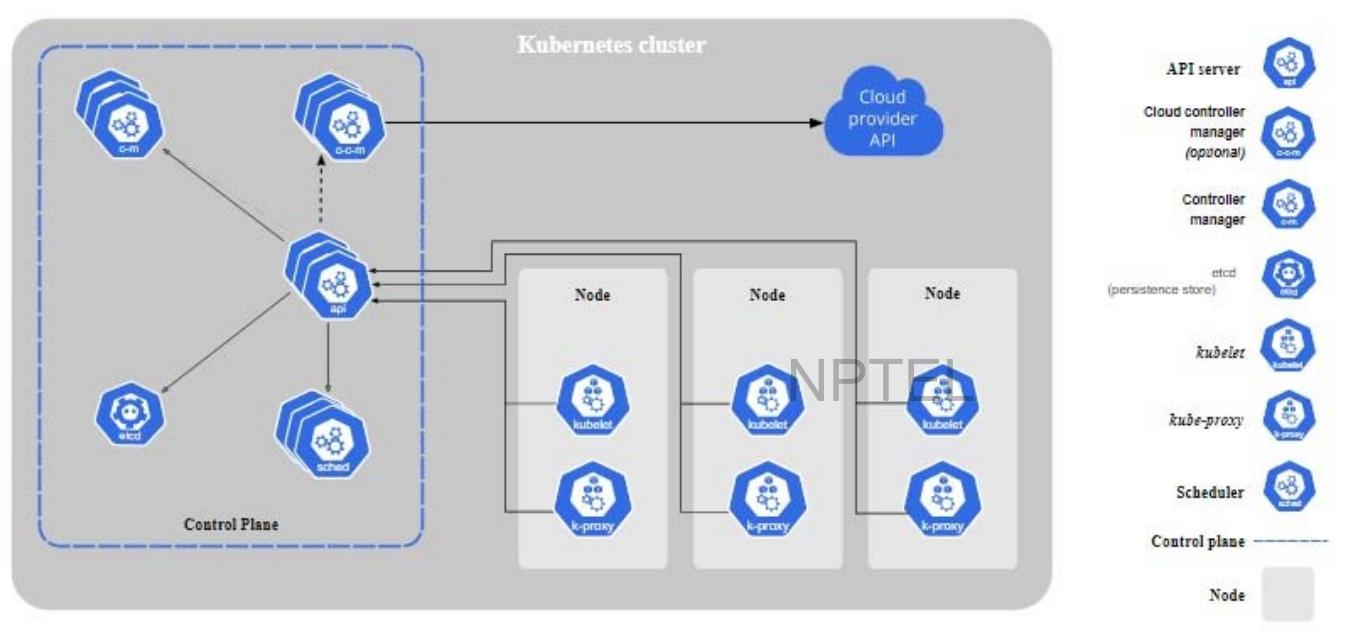
- A Kubernetes cluster consists of a set of worker machines, called **nodes**, that run containerized applications. Every cluster has at least one worker node.
- The worker node(s) host the **Pods** that are the components of the application workload.
- The **control plane** manages the worker nodes and the Pods in the cluster.
- In production environments, the control plane usually runs across multiple computers and a cluster usually runs multiple nodes, providing fault-tolerance and high availability.

NPTEL

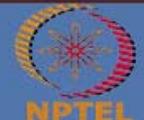
Ref: <https://kubernetes.io/docs/concepts/overview/what-is-kubernetes/>



Kubernetes Cluster Components



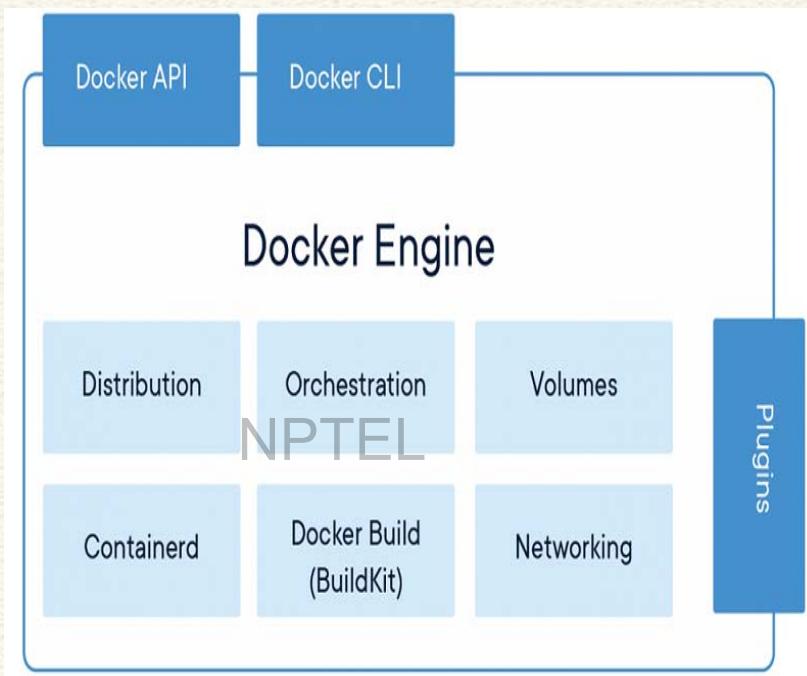
Ref: <https://kubernetes.io/docs/concepts/overview/what-is-kubernetes/>



Docker Engine

- Docker containers that run on Docker Engine:
- **Standard:** Docker created the industry standard for containers, so they could be portable anywhere
- **Lightweight:** Containers share the machine's OS system kernel and therefore do not require an OS per application, driving higher server efficiencies and reducing server and licensing costs
- **Secure:** Applications are safer in containers and Docker provides the strongest default isolation capabilities in the industry

Ref: <https://www.docker.com/resources/what-container>



Dockers

- A Docker container image is a lightweight, standalone, executable package of software that includes everything needed to run an application: code, runtime, system tools, system libraries and settings.
- Container images become containers at runtime and in the case of Docker containers - images become containers when they run on [Docker Engine](#).
- Available for both Linux and Windows-based applications, containerized software will always run the same, regardless of the infrastructure. Containers isolate software from its environment and ensure that it works uniformly despite differences for instance between development and staging.

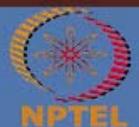
Ref: <https://www.docker.com/resources/what-container>



REFERENCES

- <https://kubernetes.io/docs/concepts/overview/what-is-kubernetes/>
- <https://www.docker.com/resources/what-container>
- <https://cloud.google.com/kubernetes-engine/docs/concepts/verticalpodautoscaler>

NPTEL



*Thank
you*



NPTEL





NPTEL ONLINE CERTIFICATION COURSES

Cloud Computing

Prof. Soumya K Ghosh

**Department of Computer Science
and Engineering**

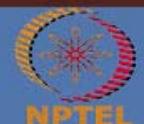
Module 11: Cloud Computing Paradigms

Lecture 51: Dew Computing

CONCEPTS COVERED

- Dew Computing – Overview
- Dew Computing – Features
- Dew Computing – Architecture
- Dew Computing – Applications

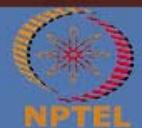
NPTEL



KEYWORDS

- Dew Computing

NPTEL

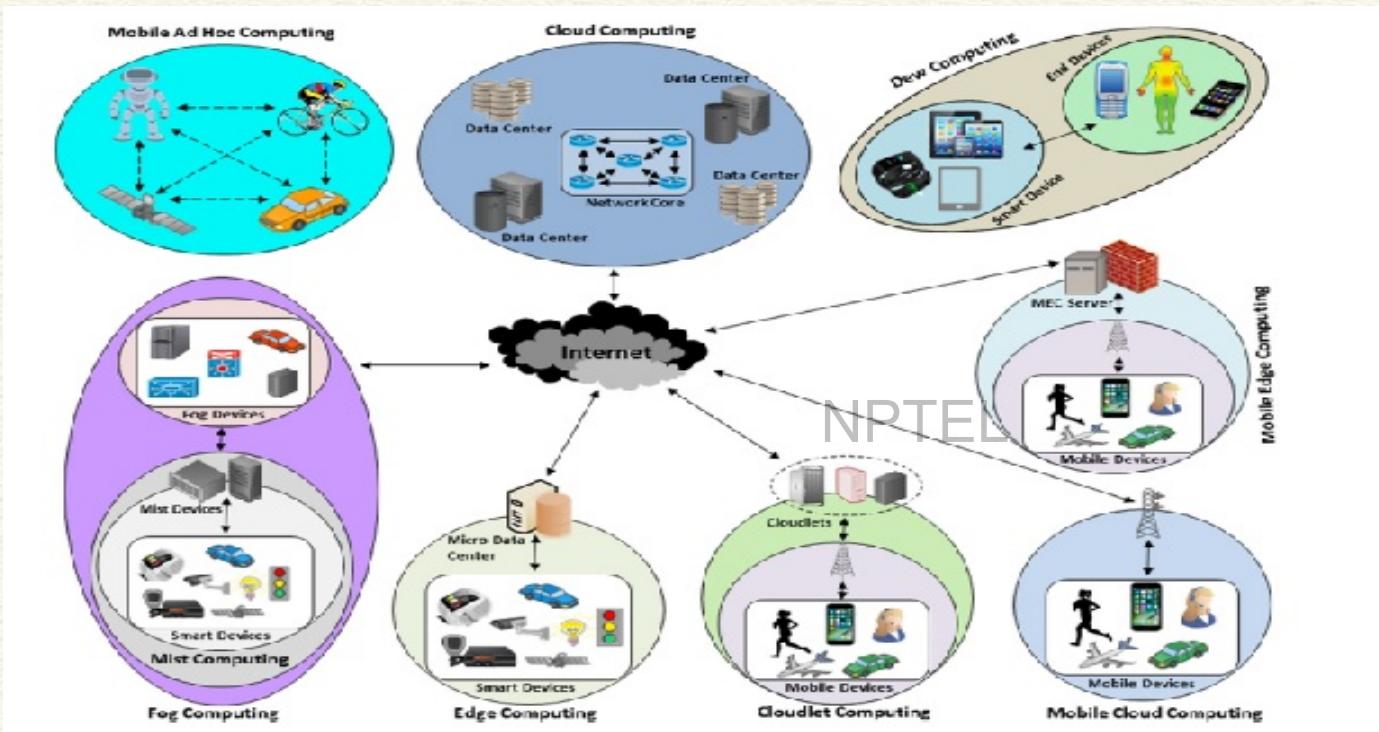


Dew Computing

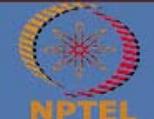
NPTEL



Cloud Computing “Family”



Ref: Sunday Oyinlola Ogundoyin, Ismaila Adeniyi Kamil, Optimization techniques and applications in fog computing: An exhaustive survey, Swarm and Evolutionary Computation, Elsevier



Dew Computing (DC)

- Dew computing is a computing paradigm that combines the core concept of cloud computing with the capabilities of end devices (personal computers, mobile phones, etc.).
- It is used to enhance the experience for the end user in comparison to only using cloud computing.
- Dew computing attempts to solve one of the major problems related to cloud computing, such as reliance on internet access.



Dew Computing (DC)

- Dew Computing is a computing model for enabling ubiquitous, pervasive, and convenient ready-to-go, plug-in facility empowered personal network that includes Single-Super-Hybrid-Peer P2P communication link.
- Primary goal: To access a pool of raw data equipped with meta-data that can be rapidly created, edited, stored, and deleted with minimal internetwork management effort (i.e. offline mode).
- To utilise all functionalities of cloud computing , Network users are heavily dependent on *Internet Connectivity* all the time.



Dew Computing (DC)

- Dew computing (DC) is a new paradigm where user-centric, flexible, and personalized-supported applications are prioritized. It is located very close to the end devices and it is the first in the IoT-fog-cloud continuum.
- DC is a micro-service-based computing paradigm with vertically distributed hierarchy.
- DC comprises smart devices, such as smart-phones, smart-watches, tablets, etc., located at the edge of the network to connect with the end devices, collect and process the IoT sensed data, and offer other services.
- The services in DC are relatively available and it is not mandatory to have a permanent Internet connection.
- DC is micro-service based which means that it does not depend on any centralized server or cloud data center.
- DC does not rely on the centralized computing devices nor a permanent Internet connection.



DC – Typical Example

Dropbox is an example of the dew computing paradigm, as it provides access to the files and folders in the cloud in addition to keeping copies on local devices.

Allows the user to access files during times without an internet connection; when a connection is established again, files and folders are synchronized back to the cloud server.

Ref: <https://www.dropbox.com>



Dew Computing - Features

- Key features of dew computing are *independence* and *collaboration*.
- Independence means that the local device must be able to provide service without a continuous connection to the Internet.
- Collaboration means that the application must be able to connect to the cloud service and synchronize data when appropriate.

NPTEL

- The word "dew" reflects natural phenomena: clouds are far from the ground, fog is closer to the ground, and dew is on the ground. Analogically, cloud computing is a remote service, fog computing is beside the user, and dew computing is at the user end.



Dew Service Models and Typical Applications

Infrastructure-As-Dew

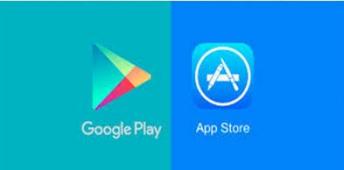
The local device is dynamically supported by cloud services.



iCloud

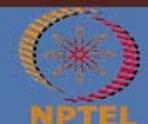
Software-In-Dew

The configuration and ownership of software are saved in the cloud.



Platform-In-Dew

A software development suite must be installed on the local device with the settings and application data synchronized to the cloud service.



Dew Service Models and Typical Applications

Storage-In-Dew

The storage of the local device is partially or fully copied into the cloud.

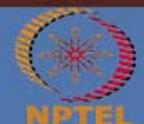


Google Drive

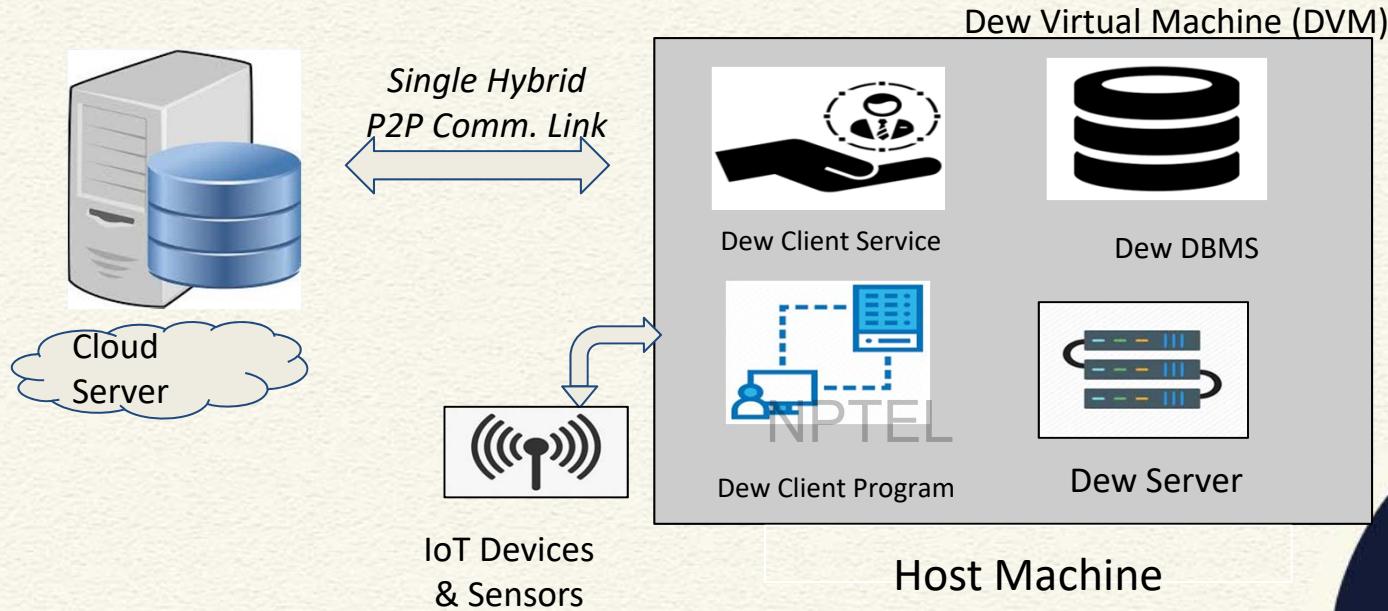
Web-in-Dew

The local device must possess a duplicated fraction of the World Wide Web (WWW).

NPTEL



Dew Computing Architecture



- To establish a cloud-dew architecture on a local machine, a dew virtual machine (DVM) is needed. The DVM is an isolated environment for executing the dew server on the local system



DC - Architecture

Dew Server functions:

- to serve user with requested services
- to perform synchronization and correlate between local data and remote data

Attempt to achieve three goals:

- Data Replication
- Data Distribution
- Synchronization

NPTEL



DC – Application Areas

- **Web in Dew (WiD)** - Possess a duplicated fraction of the World Wide Web (WWW) or a modified copy of that fraction to satisfy the independence feature. Because this fraction synchronizes with the web, it satisfies the collaboration feature of dew computing.
- **Storage in Dew (SiD)** The storage of the local device is partially or fully copied into the cloud. Since the user can access files at any time without the need for constant Internet access, this category meets the independence feature of dew computing. SiD also meets the collaboration feature because the folder and its contents automatically synchronize with the cloud service.
- **Database in Dew (DBiD)**: The local device and the cloud both store copies of the same database. One of these two databases is considered the main version and can be defined as such by the database administrator. This service increases the reliability of a database, as one of the databases can act as the backup for the other.

NPTEL



DC – Application Areas

- **Software in Dew (SiD):** The configuration and ownership of software are saved in the cloud. Examples include the Apple App Store and Google Play, where the applications the user installs are saved to their account and can then be installed on any device linked to their account.
- **Platform in Dew (PiD):** A 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 development data, system deployment data, and online backups. Example: GitHub.
- **Infrastructure as Dew (IaD):** The local device is dynamically supported by cloud services. IaD can come in different forms, but the following two forms can be used: (1) the local device can have an exact duplicate DVM instance in the cloud, which is always kept in the same state as the local instance, or (2) the local device can have all its settings/data saved in the cloud, including system settings/data and data for each application



DC – Challenges

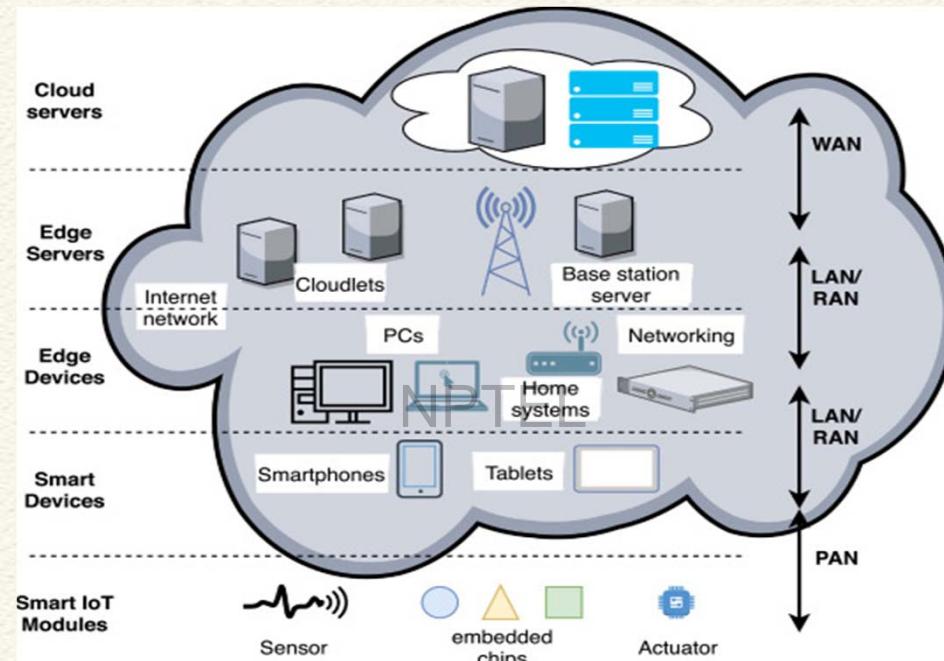
- Power Management
- Processor Utility
- Data Storage
- Viability of Operating System
- Programming Principles
- Database Security

NPTEL



Cloud Computing and Dew

Dew-enabled
Computing



REFERENCES

- https://en.wikipedia.org/wiki/Dew_computing
- Wang, Yingwei (2016). "Definition and Categorization of Dew Computing". Open Journal of Cloud Computing. 3 (1). ISSN 2199-1987.
- "Dew Computing and Transition of Internet Computing Paradigms" - ZTE Corporation
- Yingwei, Wang (2015). "The initial definition of dew computing". Dew Computing Research.
- Ray, Partha Pratim (2018). "An Introduction to Dew Computing: Definition, Concept and Implications - IEEE Journals & Magazine". IEEE Access. 6: 723–737. doi:10.1109/ACCESS.2017.2775042.
- Sunday Oyinlola Ogundoyin, Ismaila Adeniyi Kamil, Optimization techniques and applications in fog computing: An exhaustive survey, Swarm and Evolutionary Computation, Elsevier, Volume 66, 2021, <https://doi.org/10.1016/j.swevo.2021.100937>

*Thank
you*



NPTEL





NPTEL ONLINE CERTIFICATION COURSES

Cloud Computing

Prof. Soumya K Ghosh

**Department of Computer Science
and Engineering**

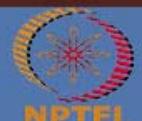
Module 11: Cloud Computing Paradigms

Lecture 52: Serverless Computing - I

CONCEPTS COVERED

- Serverless Computing
- Function-as-a-Service

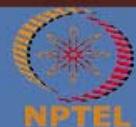
NPTEL



KEYWORDS

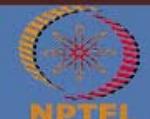
- Serverless Computing
- Function-as-a-Service

NPTEL



Serverless Computing - I

NPTEL



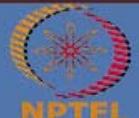
Serverless Computing

- Serverless computing is a method of providing backend services on an as-needed basis. A serverless provider allows users to write and deploy code without the hassle of worrying about the underlying infrastructure.
- Serverless architecture simplifies the code deployment and eliminates the need for system administration, allowing developers to focus on the core logic without creating additional overhead by instantiating resources, such as VMs or containers in the monitoring infrastructure.



Serverless Computing

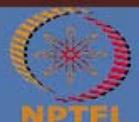
- In this model, developers execute their logic in the form of *functions* and submit to the cloud provider to run the task in a shared runtime environment; cloud providers manage the scalability needs of the *function*, by running multiple functions in parallel.
- Following the wide scale application of the containerization approach, the cloud services have adapted to offer better-fitting containers that require less time to load (boot) and to provide increased automation in handling (orchestration) containers on behalf of the client.
- Serverless computing promises to achieve full automation in managing fine-grained containers.



Serverless Computing

- *“Serverless computing is a form of cloud computing that allows users to run event-driven and granular applications, without having to address the operational logic”*
- Serverless as a computing abstraction: With serverless, developers focus on high-level abstractions (e.g., functions, queries, and events) and build applications that infrastructure operators map to concrete resources and supporting services.
- Developers focusing on the business logic and on ways to interconnect elements of business logic into complex workflows.
- Service providers ensure that the serverless applications are orchestrated—that is, containerized, deployed, provisioned, and available on demand—while billing the user for only the resources used.

NPTEL



Function-as-a-Service

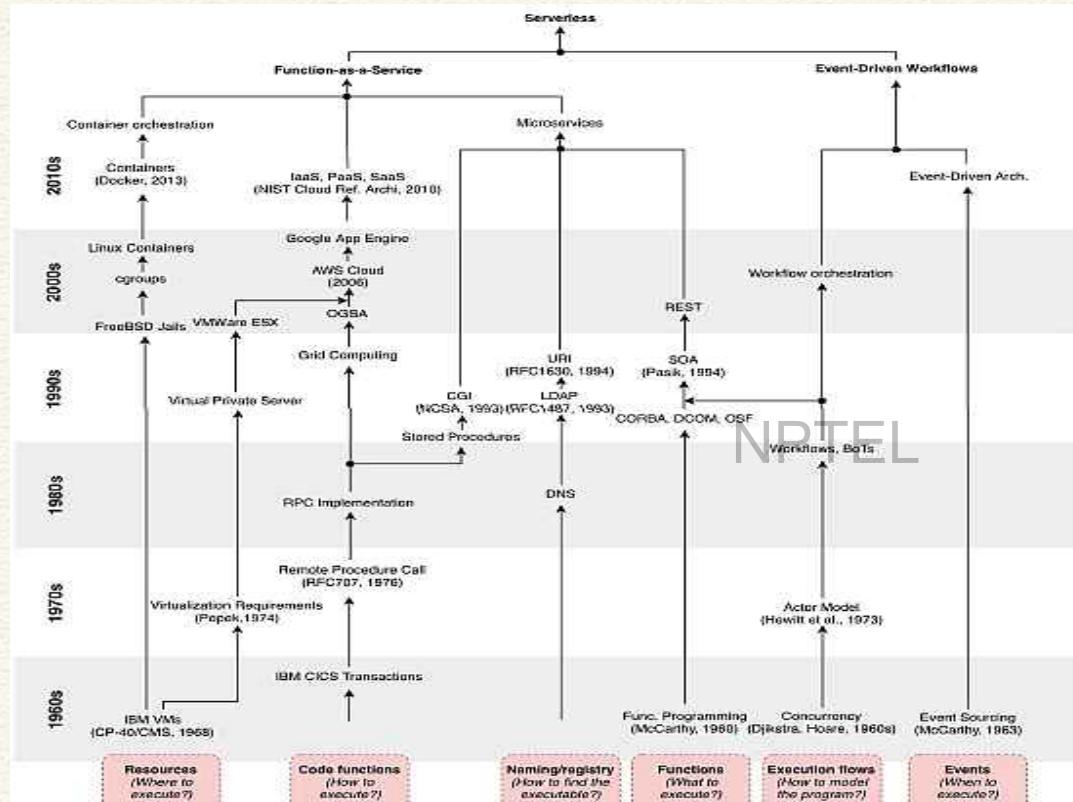
- Clients of serverless computing can use the *function-as-a-service* (FaaS) model
- *Function as a service (FaaS) is a form of serverless computing in which the cloud provider manages the resources, lifecycle, and event-driven execution of user-provided functions.*
- With FaaS, users provide small, stateless functions to the cloud provider, which manages all the operational aspects to run these functions.
- For example, consider the *ExCamera* application, which uses cloud functions and workflows to edit, transform, and encode videos with low latency and cost.

[Ref: S. Fouladi et al., “Encoding, fast and slow: Low-latency video processing using thousands of tiny threads,” Proceedings of the 14th USENIX Conference on Networked Systems Design and Implementation (NSDI 17), 2017, pp. 363–376]

- A majority of the tasks in these operations can be executed concurrently, allowing the application to improve its performance through parallelizing these tasks.



Evolution of Serverless



Serverless Computing

- In serverless, the cloud provider dynamically allocates and provisions servers.
- The code is executed in almost-stateless containers that are event-triggered, and ephemeral (may last for one invocation).
- Serverless covers a wide range of technologies, that can be grouped into two categories:
 - Backend-as-a-Service (BaaS)
 - Functions-as-a-Service (FaaS)

NPTEL



Backend-as-a-Service (BaaS)

- BaaS enables to replace server-side components with off-the-shelf services.
- BaaS enables developers to outsource all the aspects behind a scene of an application so that developers can choose to write and maintain all application logic in the frontend.
- Typical examples: remote authentication systems, database management, cloud storage, and hosting.
- Google Firebase, a fully managed database that can be directly used from an application.
- In this case, Firebase (the BaaS services) manage data components on the user's behalf.



Function-as-a-Service (FaaS)

- Serverless applications are event-driven cloud-based systems where application development relies solely on a combination of third-party services, client-side logic, and cloud-hosted remote procedure calls.
- FaaS allows developers to deploy code that, upon being triggered, is executed in an isolated environment.
- Each function typically describe a small part of an entire application. The execution time of functions is typically limited.
- Functions are not constantly active. Instead, the FaaS platforms listen for events that instantiate the functions.
- Thus, functions are triggered by events, such as client requests, events produced by any external systems, data streams, or others.
- FaaS provider is then responsible to horizontally scale function executions in response to the number of incoming events.



Serverless Computing - Challenges

- **Asynchronous calls:**
 - Asynchronous calls to and between Serverless Functions increase complexity of the system. Usually remote API calls follow request response model and are easier to implemented with synchronous calls.
- **Functions calling other functions**
 - Complex debugging, loose isolation of features. Extra costs if functions are called synchronously as we need to pay for two functions running at the same time.
- **Shared code between functions**
 - Might break existing Serverless Functions that depend on the shared code that is changed. Risk to hit the image size limit (50MB in AWS Lambda), warmup-time (the bigger the image, the longer it takes to start).

NPTEL



Serverless Computing - Challenges

- **Usage of too many libraries**
 - Increased space used by the libraries increase the risk to hit the image size limit and increase the warmup-time.
- **Adoption of too many technologies**
 - such as libraries, frameworks, languages.
 - Adds maintenance complexity and increases skill requirements for people working within the project.
- **Too many functions**
 - Creation of functions without reusing the existing one. Non-active Serverless Functions doesn't cost anything so there is temptation to create new functions instead of altering existing functionality to match changed requirements.
 - Decreased maintainability and lower system understandability.

NPTEL



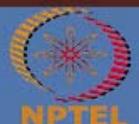
Serverless Computing – Major Providers

Service Provider	Virtual Servers	Function	Database	Storage
	Instances		 amazon DynamoDB	 S3
 Google Cloud	VMs	 Google Cloud Functions	 Google Cloud Datastore	 CLOUD STORAGE pictures, share, music, contacts, files, documents
	VM Instances	 Azure Functions	 Azure Cosmos DB	 S3 Azure Blob Storage



REFERENCES

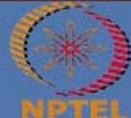
- Sanghyun Hong and Abhinav Srivastava and William Shambrook and Tudor Dumitras, Go Serverless: Securing Cloud via Serverless Design Patterns, 10th USENIX Workshop on Hot Topics in Cloud Computing (HotCloud 18), 2018, <https://www.usenix.org/conference/hotcloud18/presentation/hong>
- E. van Eyk, L. Toader, S. Talluri, L. Versluis, A. Ută and A. Iosup, "Serverless is More: From PaaS to Present Cloud Computing," in IEEE Internet Computing, vol. 22, no. 5, pp. 8-17, Sep./Oct. 2018, doi: 10.1109/MIC.2018.053681358.
- J. Nupponen and D. Taibi, "Serverless: What it Is, What to Do and What Not to Do," 2020 IEEE International Conference on Software Architecture Companion (ICSA-C), 2020, pp. 49-50, doi: 10.1109/ICSA-C50368.2020.00016.



*Thank
you*



NPTEL





NPTEL ONLINE CERTIFICATION COURSES

Cloud Computing

Prof. Soumya K Ghosh

**Department of Computer Science
and Engineering**

Module 11: Cloud Computing Paradigms

Lecture 53: Serverless Computing - II

CONCEPTS COVERED

- Serverless Computing
- AWS Lambda
- Google Cloud Functions
- Azure Functions

NPTEL



KEYWORDS

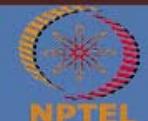
- Serverless Computing
- AWS Lambda
- Google Cloud Functions
- Azure Functions

NPTEL



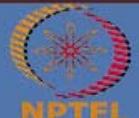
Serverless Computing - II

NPTEL



Serverless Computing

- Serverless computing hides the servers by providing programming abstractions for application builders that simplify cloud development, making cloud software easier to write.
- The focus/ target of Cloud Computing was system administrators and the Serverless is programmers. This change requires cloud providers to take over many of the operational responsibilities needed to run applications.



Serverless Computing

- To emphasize the change of focus from servers to applications, this new paradigm is known as serverless computing, although remote servers are still the invisible backend that powers it.
- This next phase of cloud computing will change the way programmers work as dramatically as the Cloud Computing has changed how operators work.
- Thus, Serverless applications are ones that don't need any server provision and do not require to manage servers.



Serverless Computing – Major Providers

Service Provider	Virtual Servers	Function	Database	Storage
	Instances		 amazon DynamoDB	 S3
 Google Cloud	VMs			
	VM Instances	 Azure Functions		 Azure Blob Storage



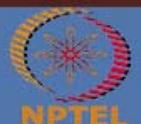
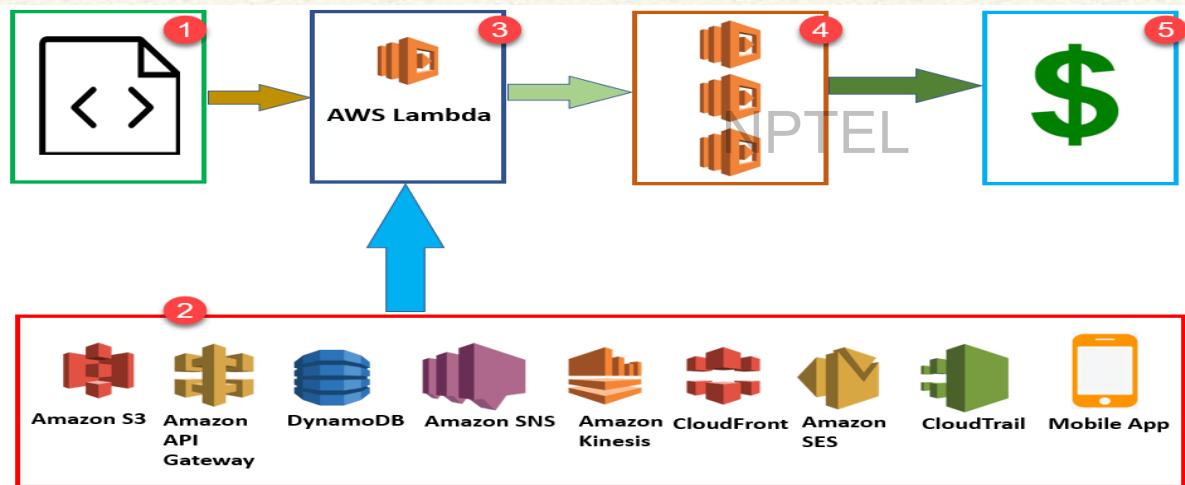
AWS Lambda

- AWS Lambda is an event-driven, serverless computing platform provided by Amazon as a part of Amazon Web Services.
- Thus one need to worry about which AWS resources to launch, or how to manage them. Instead, you need to put the code on Lambda, and it runs.
- In AWS Lambda the code is executed based on the response of events in AWS services such as add/delete files in S3 bucket, HTTP request from Amazon API gateway, etc.
- However, Amazon Lambda can only be used to execute background tasks.

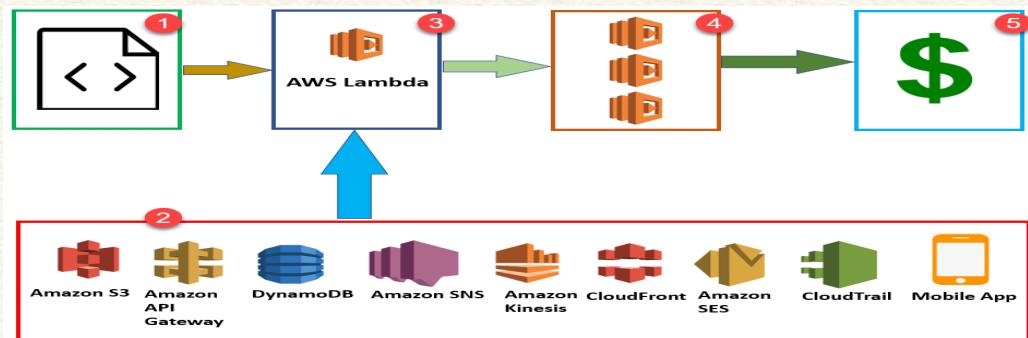


AWS Lambda

- AWS Lambda function helps you to focus on your core product and business logic instead of managing operating system (OS) access control, OS patching, right-sizing, provisioning, scaling, etc.
- AWS Lambda Block Diagram:



AWS Lambda



- 1) First upload your AWS Lambda code in any language supported by AWS Lambda. Java, Python, Go, and C# are some of the languages that are supported by AWS Lambda function.
- 2) These are some AWS services which allow you to trigger AWS Lambda.
- 3) AWS Lambda helps you to upload code and the event details on which it should be triggered.
- 4) Executes AWS Lambda Code when it is triggered by AWS services
- 5) AWS charges only when the AWS lambda code executes, and not otherwise.



AWS Lambda Concepts

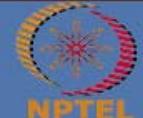
- **Function:** A function is a program or a script which runs in AWS Lambda. Lambda passes invocation events into your function, which processes an event and returns its response.
- **Runtimes:** Runtime allows functions in various languages which runs on the same base execution environment. This helps in configuring your function in runtime. It also matches your selected programming language.
- **Event source:** An event source is an AWS service, such as Amazon SNS (Simple Notification Service), or a custom service. This triggers function helps you to executes its logic.
- **Lambda Layers:** Lambda layers are an important distribution mechanism for libraries, custom runtimes, and other important function dependencies.
- **Log streams:** Log stream allows you to annotate your function code with custom logging statements which helps you to analyse the execution flow and performance of your AWS Lambda functions.



AWS Lambda – How to execute code

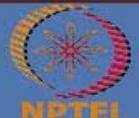
- 1) AWS Lambda URL: <https://aws.amazon.com/lambda/>
- 2) Create an Account or use Existing Account
Edit the code & Click Run...
 1. Edit the code
 2. Click Run
- 3) Check output

NPTEL

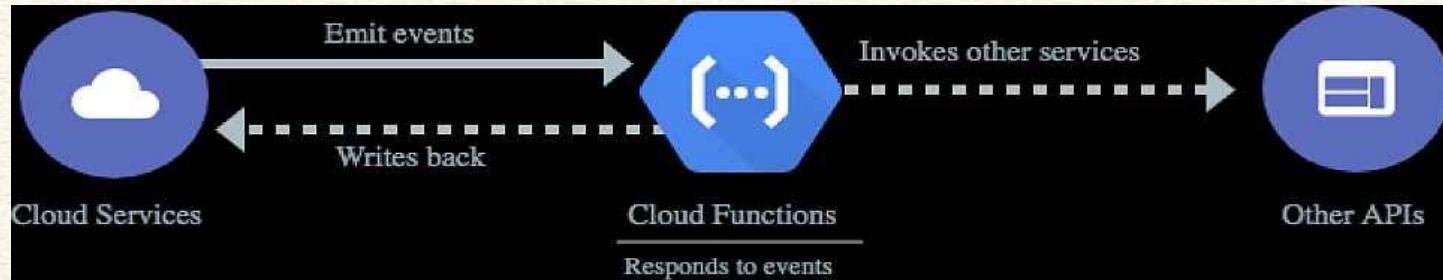


Google Cloud Functions

- Google Cloud Functions is a **serverless** execution environment for building and connecting cloud services.
- With Cloud Functions you write simple, **single-purpose functions** that are **attached to events emitted from your cloud infrastructure and services**.
- Cloud Function is **triggered when an event being watched is fired**.
- The **code executes in a fully managed environment**. There is **no need to provision any infrastructure** or worry about managing any servers.



Google Cloud Functions - Working



Cloud Services. This is the Google Cloud Platform and its various services. Services like: Google Cloud Storage, Google Cloud Pub/Sub, Stackdriver, Cloud Datastore, etc. They all have events that happen inside of them. For e.g. if a bucket has got a new object uploaded into it, deleted from it or metadata has been updated.



Google Cloud Functions - Working

Cloud Functions:

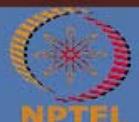
- Say an event (e.g. Object uploaded into a Bucket in Cloud Storage happens) is generated or fired or emitted. The Event data associated with that event has information on that event.
- If the Cloud Function is configured to be triggered by that event, then the Cloud Function is invoked or run or executed.
- As part of its execution, the event data is passed to it, so that it can decipher what has caused the event i.e. the event source, get meta information on the event and so on and do its processing.
- As part of the processing, it might also (maybe) invoke other APIs. (Google APIs or external APIs).
- It could even write back to the Cloud Services.



Google Cloud Functions - Working

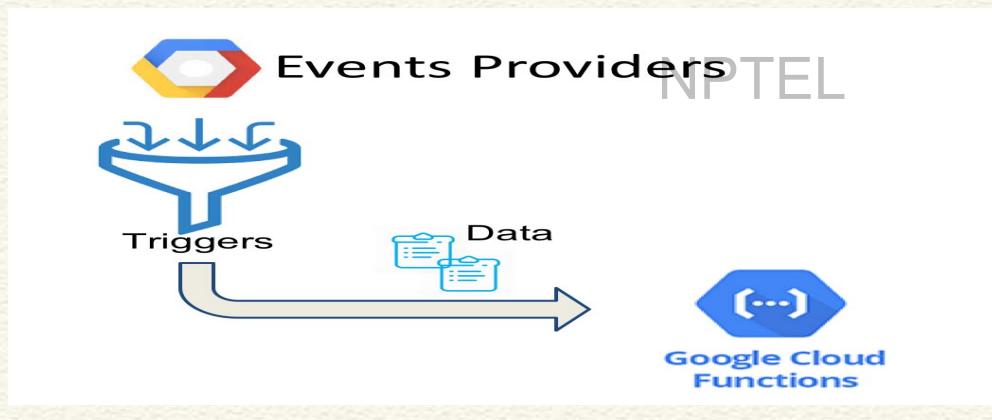
- When it has finished executing its logic, the Cloud Function mentions or specifies that it is done.
- Multiple Event occurrences will result in multiple invocations of your Cloud Functions. This is all handled for you by the Cloud Functions infrastructure. You focus on your logic inside the function and be a good citizen by keeping your function single purpose, use minimal execution time and indicate early enough that you are done and don't end up in a timeout.
- This should also indicate to you that this model works best in a stateless fashion and hence you cannot depend on any state that was created as part of an earlier invocation of your function. You could maintain state outside of this framework

NPTEL



Google Cloud Functions - Events, Triggers

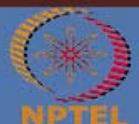
- Events : They occur in Google Cloud Platform Services E.g. File Uploaded to Storage, a Message published to a queue, Direct HTTP Invocation, etc.
- Triggers : You can chose to respond to events via a Trigger. A Trigger is the event + data associated the event.
- Event Data : This is the data that is passed to your Cloud Function when the event trigger results in your function execution.



Google Cloud Functions –Event Providers

- HTTP — invoke functions directly via HTTP requests
- Cloud Storage
- Cloud Pub/Sub
- Firebase (DB, Analytics, Auth)
- Stackdriver Logging
- Cloud Firestore
- Google Compute Engine
- BigQuery

NPTEL



Azure Functions

- Azure Functions is a serverless solution that allows you to write less code, maintain less infrastructure, and save on costs. Instead of worrying about deploying and maintaining servers, the cloud infrastructure provides all the up-to-date resources needed to keep your applications running.
- User focuses on the pieces of code, and Azure Functions handles the rest.
- A function is the primary concept in Azure Functions.
- A function contains two important pieces - your code, which can be written in a variety of languages, and some configurations, the function.json file.
- For compiled languages, this config file is generated automatically from annotations in your code. For scripting languages, you must provide the config file yourself.



Azure Functions – Build your Functions

Options and resources :

- **Use your preferred language:** Write functions in C#, Java, JavaScript, PowerShell, or Python, or use a custom handler to use virtually any other language.
- **Automate deployment:** From a tools-based approach to using external pipelines, there's a myriad of deployment options available.
- **Troubleshoot a function:** Use monitoring tools and testing strategies to gain insights into your apps.
- **Flexible pricing options:** With the Consumption plan, you only pay while your functions are running, while the Premium and App Service plans offer features for specialized needs.



Azure Functions

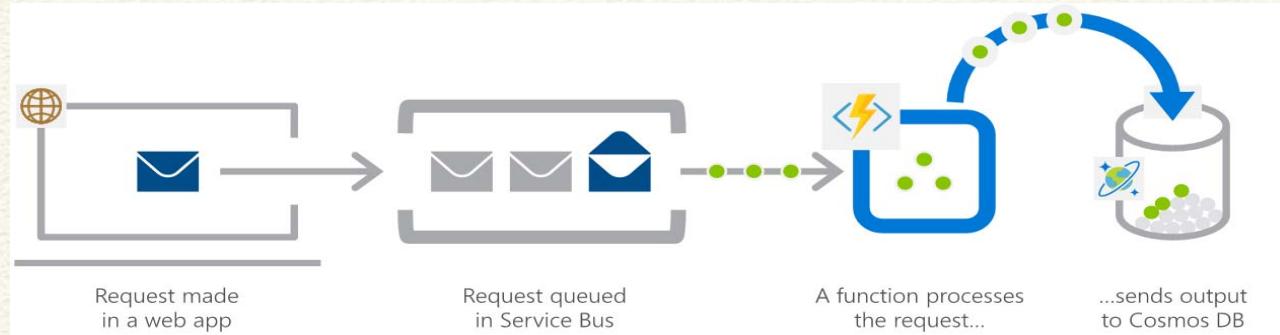
Common serverless architecture patterns include:

- Serverless APIs, mobile and web backends.
- Event and stream processing, Internet of Things (IoT) data processing, big data and machine learning pipelines.
- Integration and enterprise service bus to connect line-of-business systems, publish and subscribe (Pub/Sub) to business events.
- Automation and digital transformation and process automation.
- Middleware, software-as-a-Service (SaaS) like Dynamics, and big data projects.

NPTEL

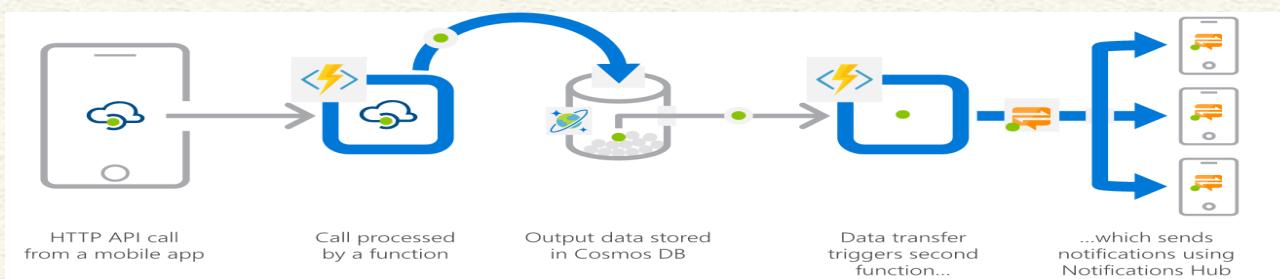


Azure Functions - Scenarios

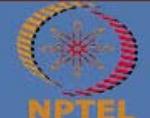


Web application backend: Retail scenario

NPTEL

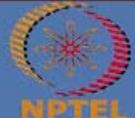


Mobile application backend: Financial services scenario



REFERENCES

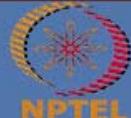
- Johann Schleier-Smith, Vikram Sreekanti, Anurag Khandelwal, Joao Carreira, Neeraja J. Yadwadkar, Raluca Ada Popa, Joseph E. Gonzalez, Ion Stoica, and David A. Patterson. 2021. What serverless computing is and should become: the next phase of cloud computing. *Commun. ACM* 64, 5 (May 2021), 76–84. DOI:<https://doi.org/10.1145/3406011>
- AWS Lambda: <https://aws.amazon.com/lambda/>
- AWS Lambda: <https://www.guru99.com/aws-lambda-function.html>
- Google Cloud Functions: <https://cloud.google.com/functions>
- Google Cloud Functions: <https://iromin.medium.com/google-cloud-functions-tutorial-what-is-google-cloud-functions-8796fa07fc7a>
- Azure Functions: <https://docs.microsoft.com/en-us/azure/azure-functions/>
- Azure Functions: <https://azure.microsoft.com/en-in/services/functions/>



*Thank
you*



NPTEL





NPTEL ONLINE CERTIFICATION COURSES

Cloud Computing

Prof. Soumya K Ghosh

**Department of Computer Science
and Engineering**

Module 11: Cloud Computing Paradigms

Lecture 54: Sustainable Cloud Computing - I

CONCEPTS COVERED

- Sustainable Computing
- Sustainable Cloud Computing

NPTEL



KEYWORDS

- Sustainable Cloud Computing
- Cloud Data Centre
- Energy Management
- Carbon Footprint

NPTEL



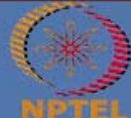
Sustainable Cloud Computing - I

NPTEL



Sustainable Cloud Computing

- Cloud Service Providers (CSPs) rely heavily on the Cloud Data Centers (CDCs) to support the ever-increasing demand for their computational and application services.
- The financial and carbon footprint related costs of running such large infrastructure negatively impacts the sustainability of cloud services. **NPTEL**
- Focus on minimizing the energy consumption and carbon footprints and ensuring reliability of the CDCs – goal of Sustainable Cloud Computing



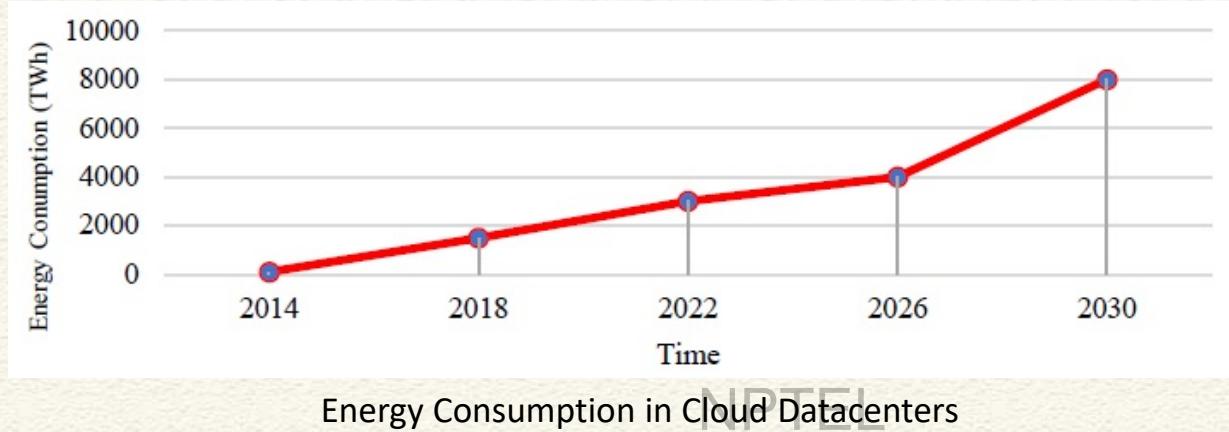
Sustainable Cloud Computing

- Cloud computing paradigm offers on-demand, subscription-oriented services over the Internet to host applications and process user workloads.
- To ensure the availability and reliability of the services, the components of Cloud Data Centers (CDCs), such as network devices, storage devices and servers are to be made available round-the-clock.
- However, creating, processing, and storing each bit of data adds to the energy cost, increases carbon footprints, and further impacts the environment.

NPTEL



Sustainable Cloud Computing



- Amount of energy consumed by the CDCs is increasing regularly and it is expected to be 8000 Tera Watt hours (TWh) by 2030

Ref: (1) Rajkumar Buyya and Sukhpal Singh Gill. "Sustainable Cloud Computing: Foundations and Future Directions." Business Technology & Digital Transformation Strategies, Cutter Consortium, Vol. 21, no. 6, Pages 1-9, 2018; (2) Anders SG Andrae, and Tomas Edler. "On global electricity usage of communication technology: trends to 2030." Challenges, vol. 6, no. 1, pp. 117-157, 2015.



Sustainable Cloud Computing

- Components (networks, storage, memory and cooling systems) of CDCs are consuming huge amount of energy.
- To improve energy efficiency of CDC, there is a need for energy-aware resource management technique for management of all the resources (including servers, storage, memory, networks, and cooling systems) in a holistic manner.
- Due to the under-loading/ over-loading of infrastructure resources, the energy consumption in CDCs is not efficient; in fact, most of the energy is consumed while some resources (i.e., networks, storage, memory, processor) are in idle state, increasing the overall cost of cloud services.

NPTEL



Sustainable Cloud Computing

- CSPs are finding other alternative ways to reduce carbon footprints of their CDCs
- Major CSPs (like Google, Amazon, Microsoft and IBM) are planning to power their datacenters using renewable energy sources.
- Future CDCs are required to provide cloud services with minimum emissions of carbon footprints and heat release in the form of greenhouse gas emissions.



Sustainable Cloud Computing

To enable sustainable cloud computing, datacenters can be relocated based on:

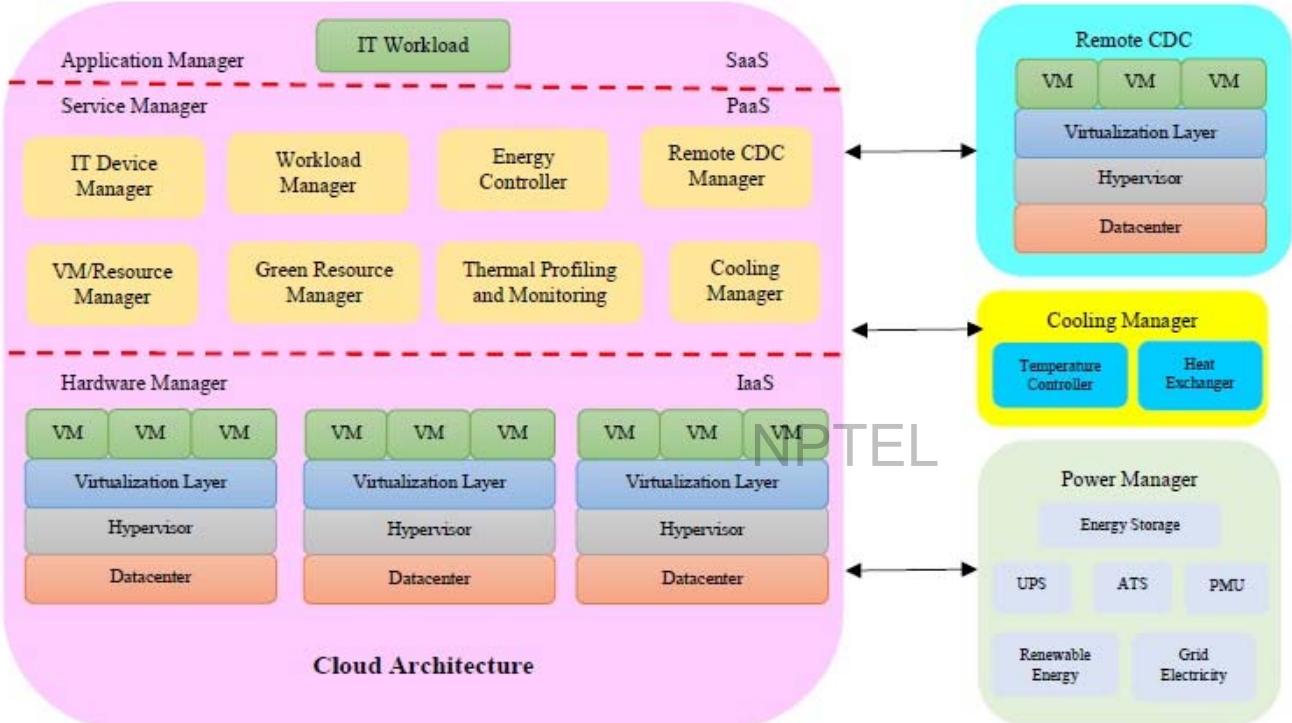
- opportunities for waste heat recovery
- accessibility of green resource and
- proximity of free cooling resources

NPTEL

- To resolve these issues and substantially reduce energy consumption of CDCs, there is a need for cloud computing architectures that can provide sustainable cloud services through holistic management of resources.

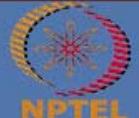


Sustainable Cloud Computing



Sustainable Cloud Computing - A Conceptual Model

Ref: Rajkumar Buyya and Sukhpal Singh Gill. "Sustainable Cloud Computing: Foundations and Future Directions." Business Technology & Digital Transformation Strategies, Cutter Consortium, Vol. 21, no. 6, Pages 1-9, 2018;



Sustainable Cloud Computing – Conceptual Model

Conceptual model for sustainable cloud computing in the form of layered architecture, which offers holistic management of cloud computing resources, to make cloud services more energy-efficient and sustainable.

- **Cloud Architecture:** This component is divided into three different sub-components: Software as a Service, Platform as a Service and Infrastructure as a Service.
- **Cooling Manager:** Thermal alerts will be generated if temperature is higher than the threshold value and heat controller will take an action to control the temperature with minimal impact on the performance of the CDC.

Ref: Rajkumar Buyya and Sukhpal Singh Gill. "Sustainable Cloud Computing: Foundations and Future Directions." Business Technology & Digital Transformation Strategies, Cutter Consortium, Vol. 21, no. 6, Pages 1-9, 2018;



Sustainable Cloud Computing

- **Power Manager:** It controls the power generated from renewable energy resources and fossil fuels (grid electricity). If there is execution of deadline oriented workloads, then grid energy can be used to maintain the reliability of cloud services. Automatic Transfer Switch (ATS) is used to manage the energy coming from both sources (renewable energy and grid electricity). Further, Power Distribution Unit is used to transfer the electricity to all the CDCs and cooling devices.
- **Remote CDC:** VMs and workloads can be migrated to a remote CDC to balance the load effectively.

NPTEL



Reliability and Sustainability - Issues

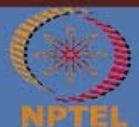
Energy

- To reduce energy consumption of cloud datacenter
- To reduce under loading and overloading of resources which improves load balancing
- To minimize heat concentration and dissipation in cloud datacenter
- To reduce carbon footprints to make environment more eco-friendly
- To improve bandwidth and computing capacity
- To improve storage management like disk-drives

Reliability

- To identify system failures and their reasons to manage the risks
- To reduce SLA violation and service delay
- To protect critical information from security attacks
- To make point to point communication using encryption and decryption
- To provide secure VM migration mechanism
- To improve capability of the system
- To reduce Turn of Investment (ToI)

NPTEL



Implication of Reliability on Sustainability

- Improving energy utilization, which reduces electricity bills and operational costs to enables sustainable cloud computing.
- However, to provide reliable cloud services, the business operations of different cloud providers are replicating services, which needs additional resources and increases energy consumption.
- Thus, a trade-off between energy consumption and reliability is required to provide cost-efficient cloud services.
- Existing energy efficient resource management techniques consume a huge amount of energy while executing workloads, which decreases resources leased from cloud datacenters.
- Dynamic Voltage and Frequency Scaling (DVFS) based energy management techniques reduced energy consumption, but response time and service delay are increased due to the switching of resources between high scaling and low scaling modes.

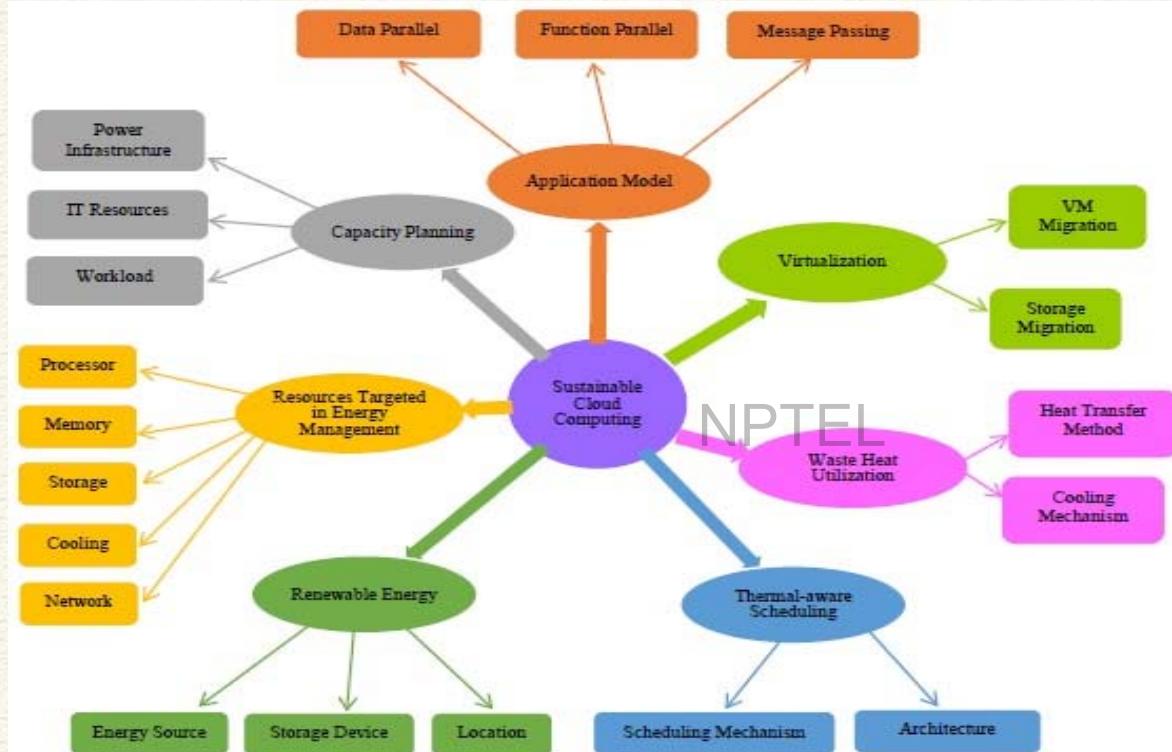


Implication of Reliability on Sustainability

- Reliability of the system component is also affected by excessive turning on/off servers.
- Power modulation decreases the reliability of server components like storage devices, memory etc.
- By reducing energy consumption of CDCs, we can improve the resource utilization, reliability and performance of the server.
- There is a need of new energy-aware resource management techniques to reduce power consumption without affecting the reliability of cloud services.



Sustainable Cloud Computing – Components



Ref: Rajkumar Buyya and Sukhpal Singh Gill. "Sustainable Cloud Computing: Foundations and Future Directions." Business Technology & Digital Transformation Strategies, Cutter Consortium, Vol. 21, no. 6, Pages 1-9, 2018;



Sustainable Cloud Computing – Components

- **Application Model:**
 - In sustainable cloud computing, the application model plays a vital role and the efficient structure of an application can improve the energy efficiency of cloud datacenters.
 - Applications models can be data parallel, function parallel and message passing.
- **Resources Targeted in Energy Management:**
 - Energy consumption of processor, memory, storage, network and cooling of cloud datacenters is typically reported as 45%, 15%, 10%, 10% and 20% respectively
 - Power regulation approaches increase energy consumption during workload execution, which affects the resource utilization of CDCs.
 - DVFS attempts to solve the problem of resource utilization but switching of resources between high scaling and low scaling modes increases response time and service delay, which may violate the SLA.
 - Putting servers in sleeping mode or turning on/off servers may affects the availability/ reliability of the system components.
 - Thus improving energy efficiency of cloud datacenters affects the resource utilization, reliability and performance of the server.



Sustainable Cloud Computing – Components

- **Thermal-aware Scheduling**

- Components of thermal-aware scheduling are architecture and scheduling mechanisms. Architecture can be single-core or multi-core while scheduling mechanism can be reactive or proactive.
- Heating problem during execution of workloads reduces the efficiency of cloud datacenters. To solve the heating problem of CDCs, thermal-aware scheduling is designed to minimize the cooling set-point temperature, hotspots and thermal gradient
- Existing thermal-aware techniques focused on reducing Power Usage Efficiency (PUE) can be found, but a reduction in PUE may not reduce the Total Cost of Ownership (TCO).

- **Virtualization**

- During the execution of workloads, VM migration is performed to balance the load effectively to utilize renewable energy resources in decentralized CDCs.
- Due to the lack of on-site renewable energy, the workloads to the other machines distributed geographically.
- VM technology also offers migration of workloads from renewable energy based cloud datacenters to the cloud datacenters utilizing the waste heat at another site.
- To balance the workload demand and renewable energy, VM based workload migration and consolidation techniques provide virtual resources using few physical servers.

NPTEL



Sustainable Cloud Computing – Components

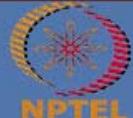
- **Capacity Planning**

- Cloud service providers must involve an effective and organized capacity planning to attain the expected return-on-investment (ROI). The capacity planning can be done for power infrastructure, IT resources and workloads.
- There is a need to consider important utilization parameters per application to maximize the utilization of resources through virtualization by finding the applications, which can be merged. Merging of applications improves resource utilization and reduces capacity cost.
- For efficient capacity planning, cloud workloads should be analysed before execution to finish its execution for deadline-oriented workloads.
- There is also a need of effective capacity planning for data storage and their processing effectively at lower cost.

NPTEL

- **Renewable Energy**

- Renewable energy source (e.g. solar or wind), the energy storage device and the location (off-site or on-site) are important factors, which can be optimized. Carbon Usage Efficiency (CUE) can be reduced by adding more renewable energy resources.
- Major challenges of renewable energy are unpredictability and high capital.
- Workload migration and energy-aware load balancing techniques attempt to address the issue of unpredictability in supply of renewable energy
- Cloud datacenters are required to place nearer the renewable energy sources to make cost effective.



Sustainable Cloud Computing – Components

- **Waste Heat Utilization**

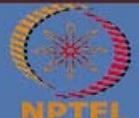
- The cooling mechanism and heat transfer model plays an important role to utilize waste heat effectively.
- Due to consumption of large amounts of energy, CDCs are acting as a heat generator. The vapor-absorption based cooling systems of CDCs can use waste heat then it utilizes the heat while evaporating.
- Vapor-absorption based free cooling techniques may help in reducing the cooling expenses. The energy efficiency of CDCs can be improved by reducing the energy usage in cooling.

NPTEL



Sustainable Cloud Computing

- The ever-increasing demand for cloud computing services that are deployed across multiple cloud datacenters harnesses significant amount of power, resulting in not only high operational cost but also high carbon emissions
- The next generation of cloud computing must be energy efficient and sustainable to fulfill end-user requirements
- In sustainable cloud computing, the CDCs are powered by renewable energy resources by replacing the conventional fossil fuel-based grid electricity or brown energy to effectively reduce carbon emissions
- Sustainability with high performance and reliability is one of the primary goals



REFERENCES

- Rajkumar Buyya and Sukhpal Singh Gill. "Sustainable Cloud Computing: Foundations and Future Directions." *Business Technology & Digital Transformation Strategies*, Cutter Consortium, Vol. 21, no. 6, Pages 1-9, 2018;
- Anders SG Andrae, and Tomas Edler. "On global electricity usage of communication technology: trends to 2030." *Challenges*, vol. 6, no. 1, pp. 117-157, 2015.
- Zhenhua Liu, Yuan Chen, Cullen Bash, Adam Wierman, Daniel Gmach, Zhikui Wang, Manish Marwah, and Chris Hyser. "Renewable and cooling aware workload management for sustainable datacenters." *ACM SIGMETRICS Performance Evaluation Review*, vol. 40, no. 1, pp. 175-186, 2012.
- Sukhpal Singh Gill and Rajkumar Buyya. 2018. A Taxonomy and Future Directions for Sustainable Cloud Computing: 360 Degree View. *ACM Comput. Surv.* 51, 5, Article 104 (December 2018), 33 pages.



*Thank
you*



NPTEL





NPTEL ONLINE CERTIFICATION COURSES

Cloud Computing

Prof. Soumya K Ghosh

**Department of Computer Science
and Engineering**

Module 11: Cloud Computing Paradigms

Lecture 55: Sustainable Cloud Computing - II

CONCEPTS COVERED

- Sustainable Computing
- Sustainable Cloud Computing

NPTEL



KEYWORDS

- Sustainable Cloud Computing
- Sustainable Cloud Computing - Taxonomy
- Energy Management
- Carbon Footprint

NPTEL



Sustainable Cloud Computing - II

NPTEL

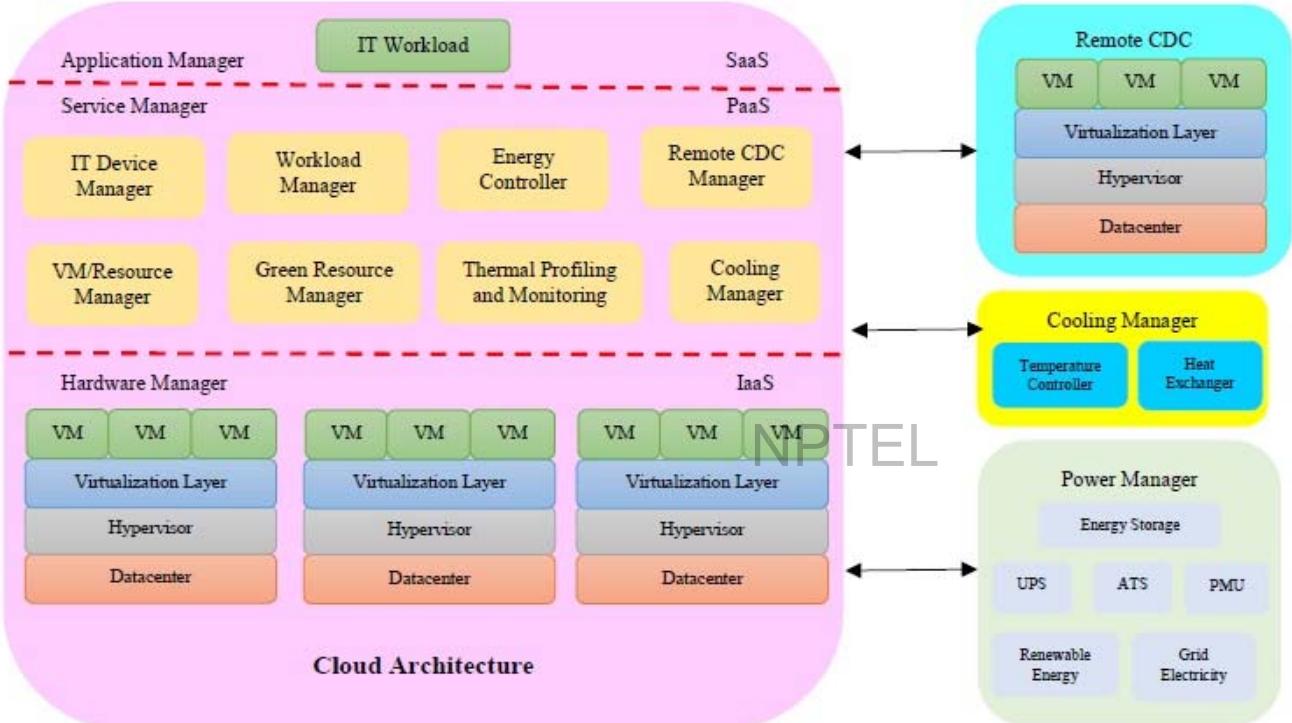


Sustainable Cloud Computing

- Cloud Service Providers (CSPs) rely heavily on the Cloud Data Centers (CDCs) to support the ever-increasing demand for their computational and application services.
- The financial and carbon footprint related costs of running such large infrastructure negatively impacts the sustainability of cloud services. **NPTEL**
- Focus on minimizing the energy consumption and carbon footprints and ensuring reliability of the CDCs – goal of Sustainable Cloud Computing



Sustainable Cloud Computing

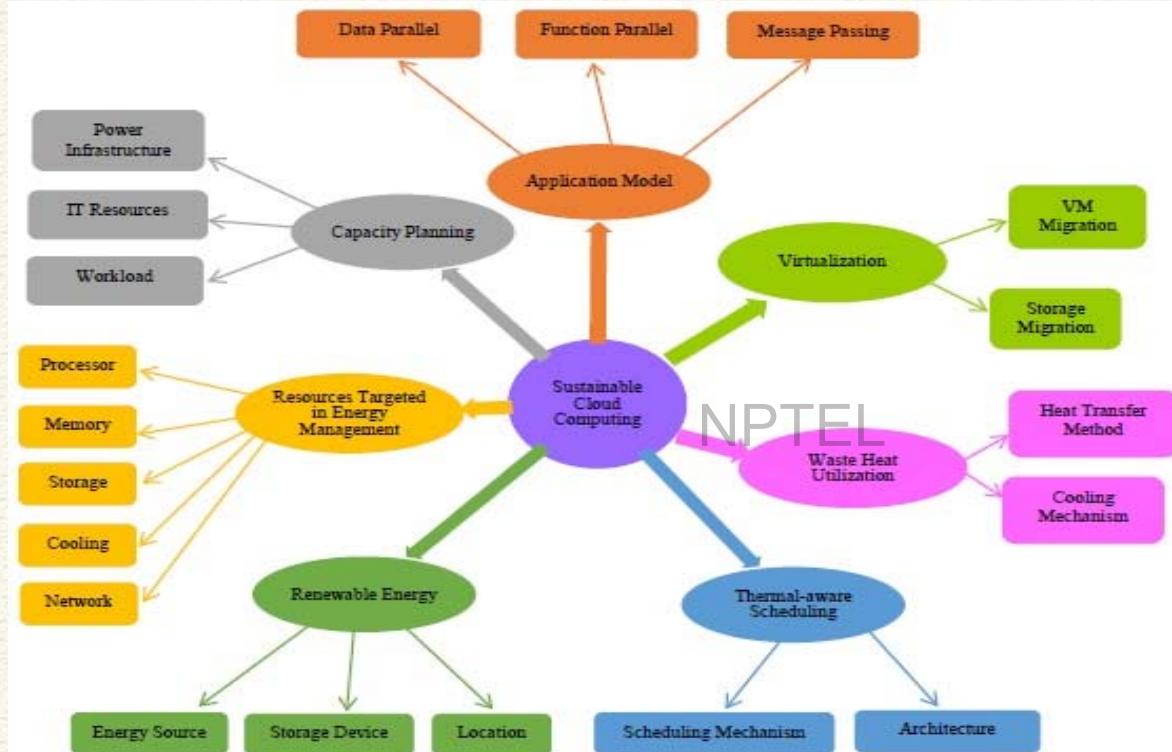


Sustainable Cloud Computing - A Conceptual Model

Ref: Rajkumar Buyya and Sukhpal Singh Gill. "Sustainable Cloud Computing: Foundations and Future Directions." Business Technology & Digital Transformation Strategies, Cutter Consortium, Vol. 21, no. 6, Pages 1-9, 2018;



Sustainable Cloud Computing – Components



Ref: Rajkumar Buyya and Sukhpal Singh Gill. "Sustainable Cloud Computing: Foundations and Future Directions." Business Technology & Digital Transformation Strategies, Cutter Consortium, Vol. 21, no. 6, Pages 1-9, 2018;



Sustainable Cloud Computing – Components

- **Application Model:**
 - In sustainable cloud computing, the application model plays a vital role and the efficient structure of an application can improve the energy efficiency of cloud datacenters.
 - Applications models can be data parallel, function parallel and message passing.
- **Resources Targeted in Energy Management:**
 - Energy consumption of processor, memory, storage, network and cooling of cloud datacenters is typically reported as 45%, 15%, 10%, 10% and 20% respectively
 - Power regulation approaches increase energy consumption during workload execution, which affects the resource utilization of CDCs.
 - DVFS attempts to solve the problem of resource utilization but switching of resources between high scaling and low scaling modes increases response time and service delay, which may violate the SLA.
 - Putting servers in sleeping mode or turning on/off servers may affects the availability/ reliability of the system components.
 - Thus improving energy efficiency of cloud datacenters affects the resource utilization, reliability and performance of the server.



Sustainable Cloud Computing – Components

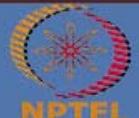
- **Thermal-aware Scheduling**

- Components of thermal-aware scheduling are architecture and scheduling mechanisms. Architecture can be single-core or multi-core while scheduling mechanism can be reactive or proactive.
- Heating problem during execution of workloads reduces the efficiency of cloud datacenters. To solve the heating problem of CDCs, thermal-aware scheduling is designed to minimize the cooling set-point temperature, hotspots and thermal gradient
- Existing thermal-aware techniques focused on reducing Power Usage Efficiency (PUE) can be found, but a reduction in PUE may not reduce the Total Cost of Ownership (TCO).

- **Virtualization**

- During the execution of workloads, VM migration is performed to balance the load effectively to utilize renewable energy resources in decentralized CDCs.
- Due to the lack of on-site renewable energy, the workloads to the other machines distributed geographically.
- VM technology also offers migration of workloads from renewable energy based cloud datacenters to the cloud datacenters utilizing the waste heat at another site.
- To balance the workload demand and renewable energy, VM based workload migration and consolidation techniques provide virtual resources using few physical servers.

NPTEL



Sustainable Cloud Computing – Components

- **Capacity Planning**

- Cloud service providers must involve an effective and organized capacity planning to attain the expected return-on-investment (ROI). The capacity planning can be done for power infrastructure, IT resources and workloads.
- There is a need to consider important utilization parameters per application to maximize the utilization of resources through virtualization by finding the applications, which can be merged. Merging of applications improves resource utilization and reduces capacity cost.
- For efficient capacity planning, cloud workloads should be analysed before execution to finish its execution for deadline-oriented workloads.
- There is also a need of effective capacity planning for data storage and their processing effectively at lower cost.

NPTEL

- **Renewable Energy**

- Renewable energy source (e.g. solar or wind), the energy storage device and the location (off-site or on-site) are important factors, which can be optimized. Carbon Usage Efficiency (CUE) can be reduced by adding more renewable energy resources.
- Major challenges of renewable energy are unpredictability and high capital.
- Workload migration and energy-aware load balancing techniques attempt to address the issue of unpredictability in supply of renewable energy
- Cloud datacenters are required to place nearer the renewable energy sources to make cost effective.

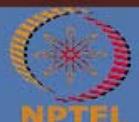


Sustainable Cloud Computing – Components

- **Waste Heat Utilization**

- The cooling mechanism and heat transfer model plays an important role to utilize waste heat effectively.
- Due to consumption of large amounts of energy, CDCs are acting as a heat generator. The vapor-absorption based cooling systems of CDCs can use waste heat then it utilizes the heat while evaporating.
- Vapor-absorption based free cooling techniques may help in reducing the cooling expenses. The energy efficiency of CDCs can be improved by reducing the energy usage in cooling.

NPTEL



Sustainable Cloud Computing – Taxonomy

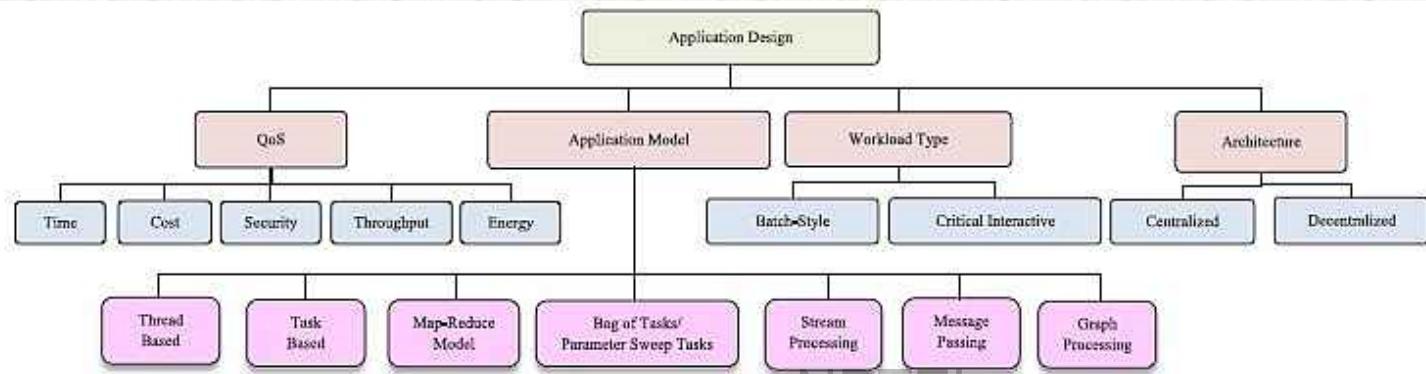
- With huge growth of Internet of Things (IoT)-based applications, the use of cloud services is increasing exponentially.
- Thus, cloud computing must be energy efficient and sustainable to fulfill the ever-increasing end-user needs.
- Research initiatives on sustainable cloud computing can be categorized as follows:
 - application design
 - sustainability metrics
 - capacity planning
 - energy management
 - Virtualization
 - thermal-aware scheduling
 - cooling management
 - renewable energy
 - waste heat utilization

NPTEL

Ref: Sukhpal Singh Gill and Rajkumar Buyya. 2018. A Taxonomy and Future Directions for Sustainable Cloud Computing: 360 Degree View. ACM Comput. Surv. 51, 5, Article 104 (December 2018), 33 pages. <https://doi.org/10.1145/3241038>



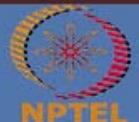
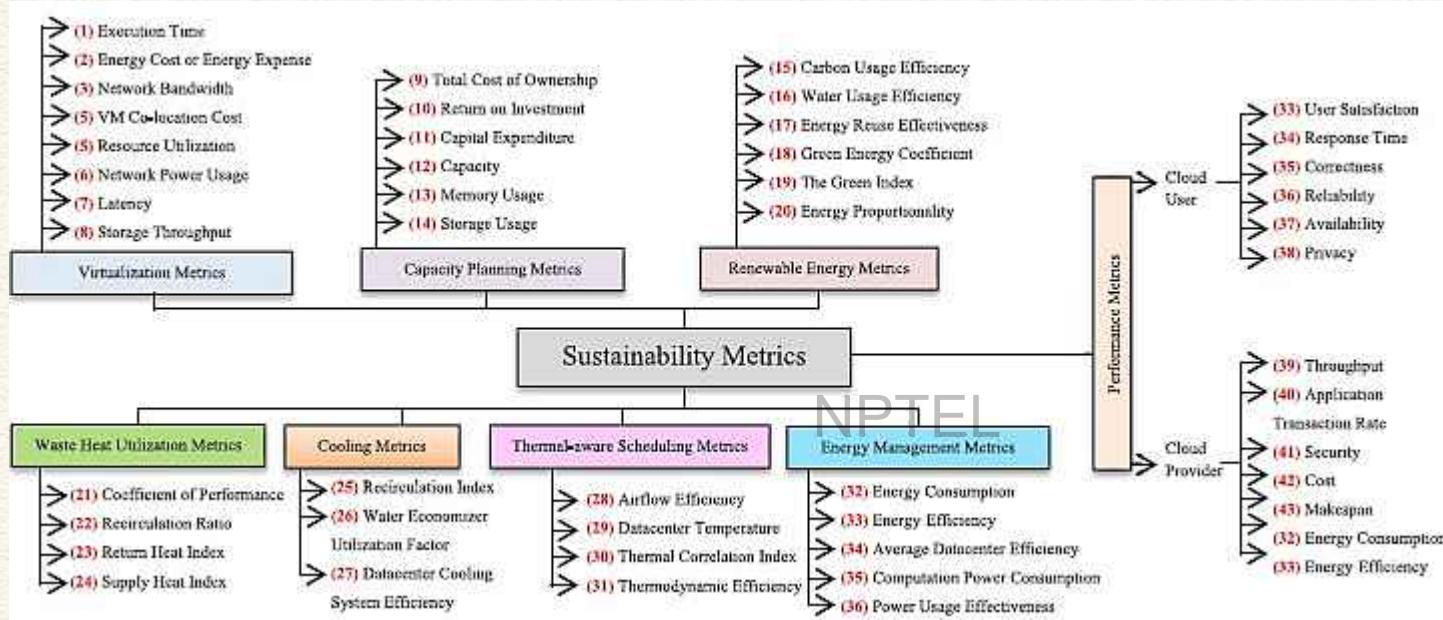
Application Design



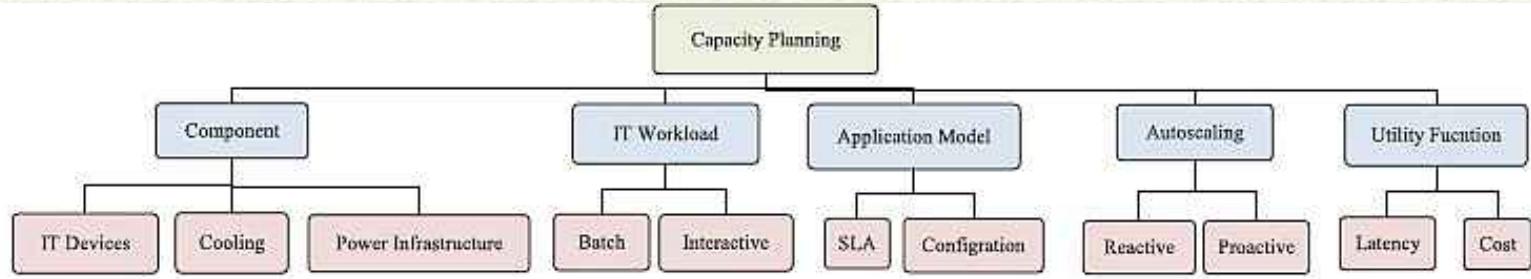
- Design of an application plays a vital role and the efficient structure of an application can improve energy efficiency of CDCs.
- The resource manager and scheduler follow different approaches for application modelling
- To make the infrastructure sustainable and environmentally eco-friendly, there is a need for green ICT-based innovative applications



Sustainability Metrics

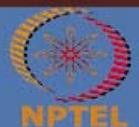


Capacity Planning

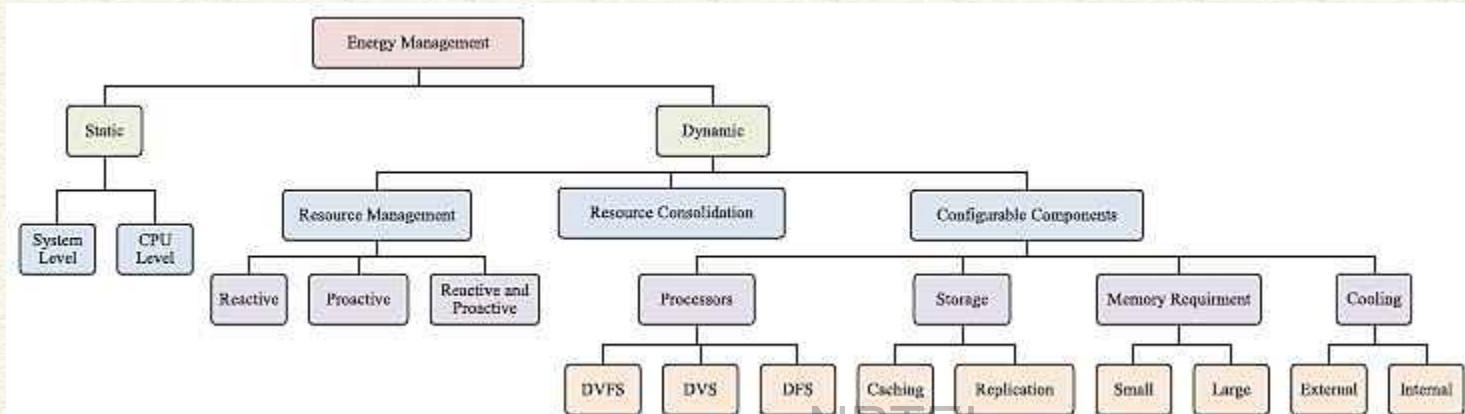


NPTEL

- CSPs must initiate effective and organized capacity planning to enable sustainable computing.
- Capacity planning can be done for power infrastructure, IT infrastructure, and cooling mechanism.



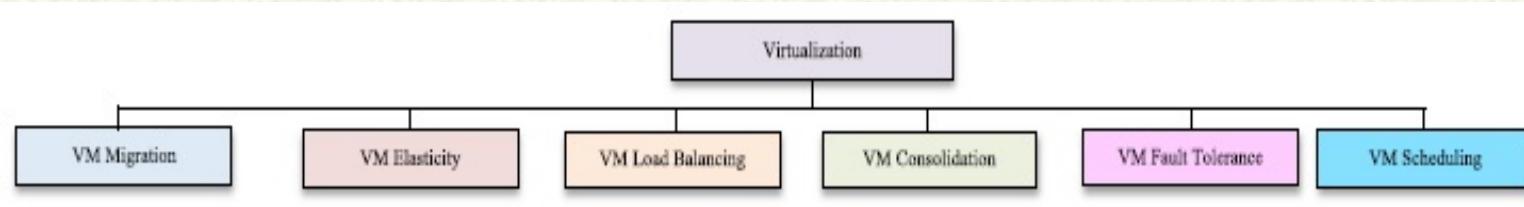
Energy Management



- Energy management in sustainable computing is an important factor for CSPs
- Improving energy use reduces electricity bills and operational costs to enable sustainable cloud computing.
- Essential requirements of sustainable CDCs are optimal software system design, optimized air ventilation, and installing temperature monitoring tools for adequate resource utilization, which improves energy efficiency



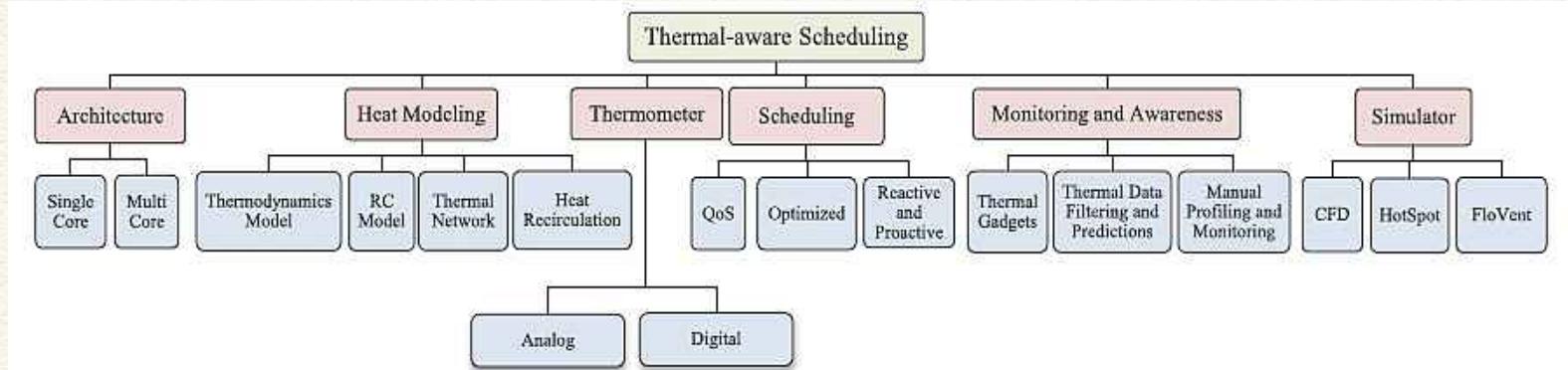
Virtualization



- Virtualization technology is an important part of sustainable CDCs to support energy-efficient VM migration, VM elasticity, VM load balancing, VM consolidation, VM fault tolerance, and VM scheduling
- Operational costs can be reduced by using VM scheduling to manage cloud resources using efficient dynamic provisioning of resources



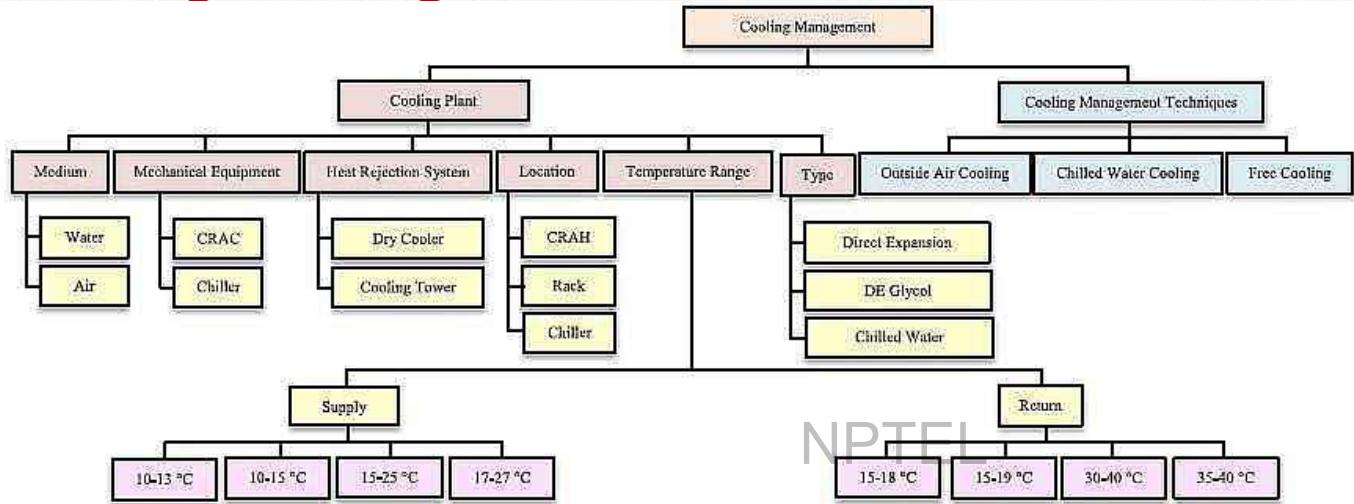
Thermal-aware Scheduling



- CDCs consist of a chassis and racks to place the servers to process the IT workloads.
- To maintain the temperature of datacenters, cooling mechanisms are needed.
- Servers produce heat during execution of IT workload. The processor is an important component of a server and consumes the most electricity.
- Both cooling and computing mechanisms consume a huge amount of electricity. Proper management is needed.



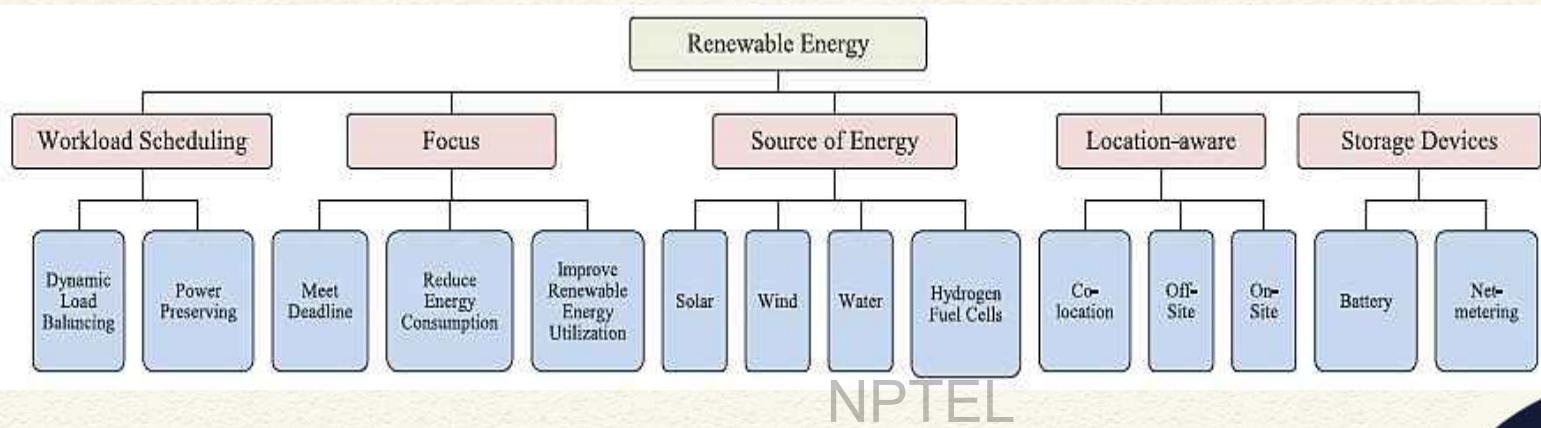
Cooling Management



- The increasing demand for computation, networking, and storage expands the complexity, size, and energy density of CDCs exponentially, which consumes a large amount of energy and produces a huge amount of heat.
- To make CDCs more energy efficient and sustainable, we need an effective cooling management system, which can maintain the temperature of CDCs



Renewable Energy

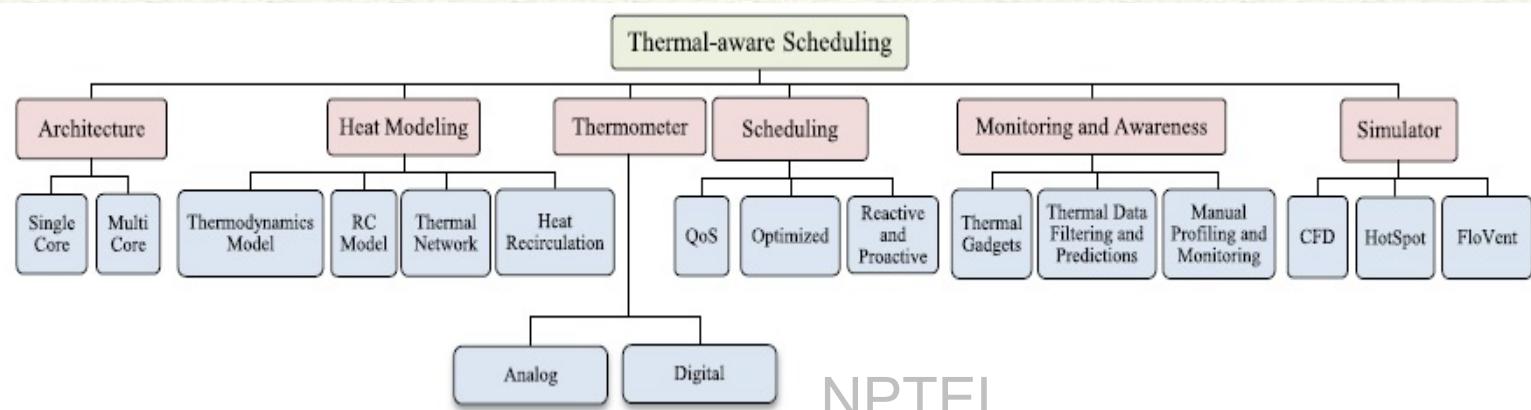


NPTEL

- Sustainable computing needs energy-efficient workload execution by using renewable energy resources to reduce carbon emissions
- Green energy resources, such as solar, wind, and water generate energy with nearly zero carbon-dioxide emissions



Waste Heat Utilization



NPTEL

- Reuse of waste heat is becoming a solution for fulfilling energy demand in energy conservation systems
- The vapor-absorption-based cooling systems can use waste heat, and remove the heat while evaporating.
- Vapor-absorption-based free cooling mechanisms can make the value of PUE (Power Usage Effectiveness) ideal by neutralizing cooling expenses.



Sustainable Cloud Computing

- The ever-increasing demand for cloud computing services that are deployed across multiple cloud datacenters harnesses significant amount of power, resulting in not only high operational cost but also high carbon emissions
- The next generation of cloud computing must be energy efficient and sustainable to fulfill end-user requirements
- In sustainable cloud computing, the CDCs are powered by renewable energy resources by replacing the conventional fossil fuel-based grid electricity or brown energy to effectively reduce carbon emissions
- Sustainability with high performance and reliability is one of the primary goals



REFERENCES

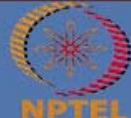
- Rajkumar Buyya and Sukhpal Singh Gill. "Sustainable Cloud Computing: Foundations and Future Directions." Business Technology & Digital Transformation Strategies, Cutter Consortium, Vol. 21, no. 6, Pages 1-9, 2018
- Sukhpal Singh Gill and Rajkumar Buyya. 2018. A Taxonomy and Future Directions for Sustainable Cloud Computing: 360 Degree View. ACM Comput. Surv. 51, 5, Article 104 (December 2018), 33 pages.
- Anders SG Andrae, and Tomas Edler. "On global electricity usage of communication technology: trends to 2030." Challenges, vol. 6, no. 1, pp. 117-157, 2015.
- Zhenhua Liu, Yuan Chen, Cullen Bash, Adam Wierman, Daniel Gmach, Zhikui Wang, Manish Marwah, and Chris Hyser. "Renewable and cooling aware workload management for sustainable datacenters." ACM SIGMETRICS Performance Evaluation Review, vol. 40, no. 1, pp. 175-186, 2012.
- Sukhpal Singh Gill, Inderveer Chana, Maninder Singh and Rajkumar Buyya. 2018. RADAR: Self-Configuring and Self-Healing in Resource Management for Enhancing Quality of Cloud Services, Concurrency and Computation: Practice and Experience (CCPE), 2018.



*Thank
you*



NPTEL





NPTEL ONLINE CERTIFICATION COURSES

Cloud Computing

Prof. Soumya K Ghosh

**Department of Computer Science
and Engineering**

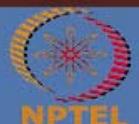
Module 12: Cloud Computing Paradigms

Lecture 56: Cloud Computing in 5G Era

CONCEPTS COVERED

- 5G Network
- Cloud Computing in 5G

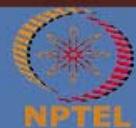
NPTEL



KEYWORDS

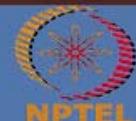
- Spatial Data
- Spatial Cloud Computing

NPTEL



Cloud Computing in 5G

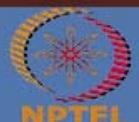
NPTEL



5G Network

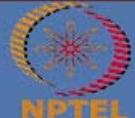
- 5G is the 5th generation mobile network. It is a new global wireless standard after 1G, 2G, 3G, and 4G networks. 5G enables a new kind of network that is designed to connect virtually everyone and everything together including machines, objects, and devices.
- 5G wireless technology is meant to deliver higher multi-Gbps peak data speeds, ultra low latency, more reliability, massive network capacity, increased availability, and a more uniform user experience to more users. Higher performance and improved efficiency empower new user experiences and connects new industries.

NPTEL



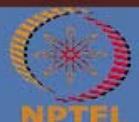
Different Generations

- **First generation - 1G** - 1980s: 1G delivered analog voice.
- **Second generation - 2G** - Early 1990s: 2G introduced digital voice (e.g. CDMA- Code Division Multiple Access).
- **Third generation - 3G** - Early 2000s: 3G brought mobile data (e.g. CDMA2000).
- **Fourth generation - 4G LTE** - 2010s: 4G LTE ushered in the era of mobile broadband.
- 1G, 2G, 3G, and 4G all led to **5G**, which is designed to provide more connectivity than was ever available before.
- **5G** is a unified, more capable air interface. It has been designed with an extended capacity to enable next-generation user experiences, empower new deployment models and deliver new services.
- With high speeds, superior reliability and negligible latency, 5G is all set to expand the mobile ecosystem into new realms.
- 5G will impact Cloud Computing paradigm in a big way.

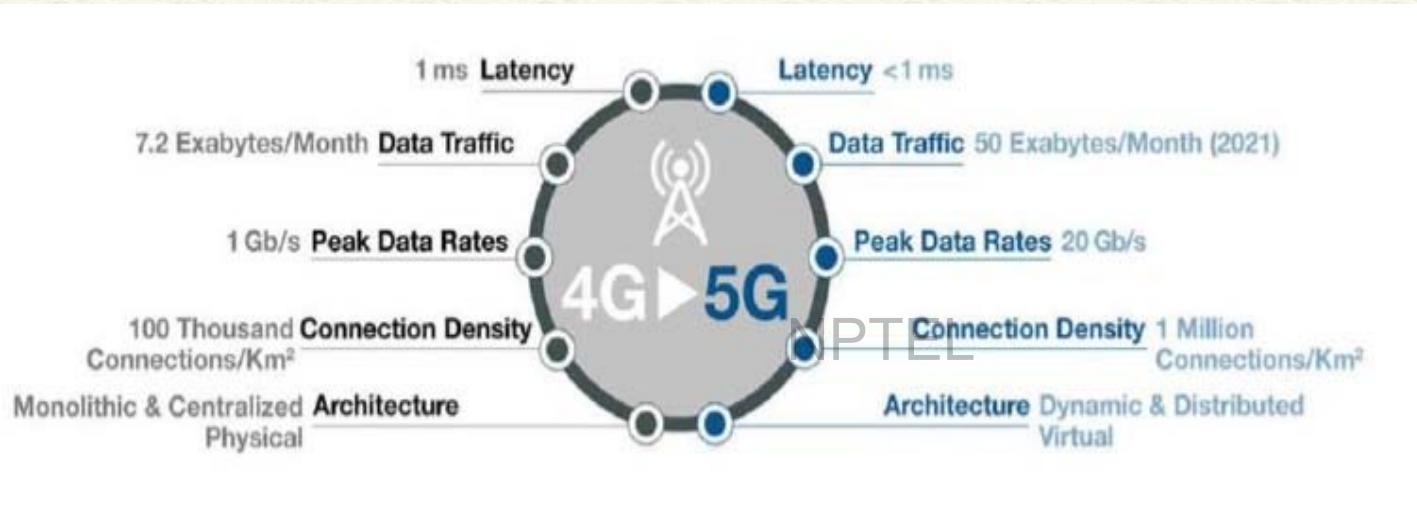


Evolution of Mobile Networks

	1G	2G	3G	4G	5G
Approximate deployment date	1980s	1990s	2000s	2010s	2020s
Theoretical download speed	2kbit/s	384kbit/s	56Mbit/s	1Gbit/s	10Gbit/s
Latency	N/A	629 ms	212 ms	60-98 ms	< 1 ms



4G vs 5G Features



Use of 5G

- 5G is designed for forward compatibility—the ability to flexibly support future services.
- 5G is used across three main types of connected services.
- **Enhanced mobile broadband**
In addition to making our smartphones better, 5G mobile technology can usher in new immersive experiences such as VR and AR with faster, more uniform data rates, lower latency, and lower cost-per-bit.
- **Mission-critical communications**
5G can enable new services that can transform industries with ultra-reliable, available, low-latency links like remote control of critical infrastructure, vehicles, and medical procedures.
- **Massive IoT**
5G is meant to seamlessly connect a massive number of embedded sensors in virtually everything through the ability to scale down in data rates, power, and mobility—providing extremely lean and low-cost connectivity solutions.

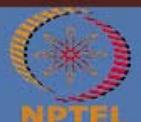
NPTEL



5G Network - Features

- **Enhanced mobile broadband (eMBB)** – enhanced indoor and outdoor broadband, enterprise collaboration, augmented and virtual reality.
- **Massive machine-type communications (mMTC)** – IoT, asset tracking, smart agriculture, smart cities, energy monitoring, smart home, remote monitoring.
- **Ultra-reliable and low-latency communications (URLLC)** – autonomous vehicles, smart grids, remote patient monitoring and telehealth, industrial automation.

NPTEL



5G and Cloud Computing

- 5G is the perfect companion to cloud computing both in terms of its distribution and the diversity of compute and storage capabilities.
- On-premises and edge data centers will continue to close the gap between resource-constrained low-latency devices and distant cloud data centers, leading to driving the need for heterogeneous and distributed computing architectures.
- In this evolving computing paradigm, service providers should look to provide full end-to-end orchestration, with defined service layer agreements, in a self-service and automated way.
- *Network as a Platform* for enterprise services
- Service orchestration will play a key role moving forward, enabling industrial applications to interact with the network resources in advanced ways such as selecting location, quality of service, or influencing the traffic routing to deliver on application demands.



5G and Cloud Computing

- Two key aspects in the relationship between 5G technologies and cloud computing.
 - First, further development of cloud computing has to meet the 5G needs. This is reflected by growing roles of edge, mobile edge, and fog computing in the cloud computing realm.
 - Second aspect is that 5G technologies are undergoing “cloudification” through network softwarization”, NFV, SDN, etc.
 - Both technology types influence the developments of each other.
- 5G deployments bring up discussions about the convergence of computing, cloud, and IoT that takes us to the era of hyper-connectivity.



Edge Computing in 5G

- 5G is the next generation cellular network that aspires to achieve substantial improvement on quality of service, such as higher throughput and lower latency.
- Edge computing is an emerging technology that enables the evolution to 5G by bringing cloud capabilities near to the end users (or user equipment, UEs) in order to overcome the intrinsic problems of the traditional cloud, such as high latency etc.
- Edge computing is preferred to cater for the wireless communication requirements of next generation applications, such as augmented reality and virtual reality, which are interactive in nature.
 - These highly interactive applications are computationally-intensive and have high quality of service (QoS) requirements, including low latency and high throughput.
 - Further, these applications are expected to generate a massive amount



Edge Computing in 5G

- 5G is expected to cater following needs of today's network traffic
 - Handle massive amount of data generated by mobile devices/ IoTs
 - Stringent QoS requirements are imposed to support highly interactive applications, requiring ultra-low latency and high throughput
 - Heterogeneous environment must be supported to allow interoperability of a diverse range of end-user equipment, QoS requirements, network types etc.

NPTEL



Edge Computing in 5G - Applications

- Healthcare
- Entertainment and multimedia applications
- Virtual reality, augmented reality, and mixed reality
- Tactile internet
- Internet of Things
- Factories of the future
- Emergency response
- Intelligent Transportation System

NPTEL

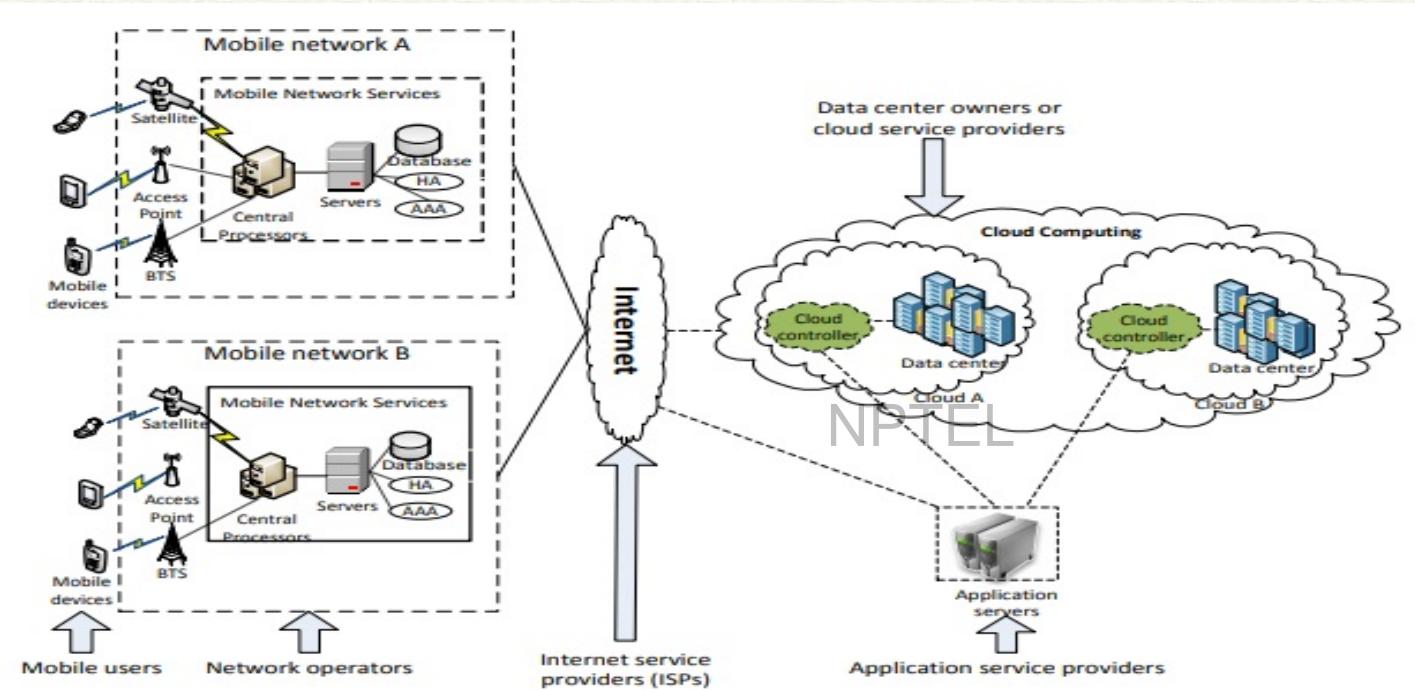


5G and Mobile Cloud Computing (MCC)

- MCC is a cloud computing system including mobile devices and delivering applications to the mobile devices.
- Key features of MCC for 5G networks include sharing resources for mobile applications and improved reliability as data is backed up and stored in the cloud.
- As data processing is offloaded by MCC from the devices to the cloud, fewer device resources are consumed by applications.
- Compute-intensive processing of mobile users' requests is off-loaded from mobile networks to the cloud. Mobile devices are connected to mobile networks via base stations (e.g., base transceiver station, access point, or satellite).

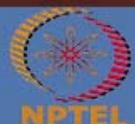


Mobile Cloud Computing (MCC)



REFERENCES

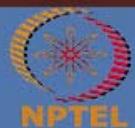
- Qualcomm: <https://www.qualcomm.com/5g/what-is-5g>
- Ericsson: <https://www.ericsson.com/en/blog/2021/2/5g-and-cloud>
- N. Hassan, K. A. Yau and C. Wu, "Edge Computing in 5G: A Review," in IEEE Access, vol. 7, pp. 127276-127289, 2019, doi: 10.1109/ACCESS.2019.2938534
- Setting the Scene for 5G: Opportunities & Challenges. International Telecommunication Union, 2018 **NPTEL**
- Securing 4G, 5G and Beyond with Fortinet: <https://www.fortinet.com/solutions/mobile-carrier.html>
- How 5G Transforms Cloud Computing, Dell Technologies, https://education.dell EMC.com/content/dam/dell-emc/documents/en-us/2020KS_Gloukhovtsev_How_5G_Transforms_Cloud_Computing.pdf



*Thank
you*



NPTEL





NPTEL ONLINE CERTIFICATION COURSES

Cloud Computing

Prof. Soumya K Ghosh

**Department of Computer Science
and Engineering**

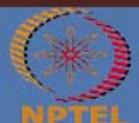
Module 12: Cloud Computing Paradigms

Lecture 57: CPS and Cloud Computing

CONCEPTS COVERED

- Cyber Physical System (CPS)
- CPS and Cloud Computing

NPTEL



KEYWORDS

- Cyber Physical System (CPS)

NPTEL



CPS and Cloud Computing

NPTEL

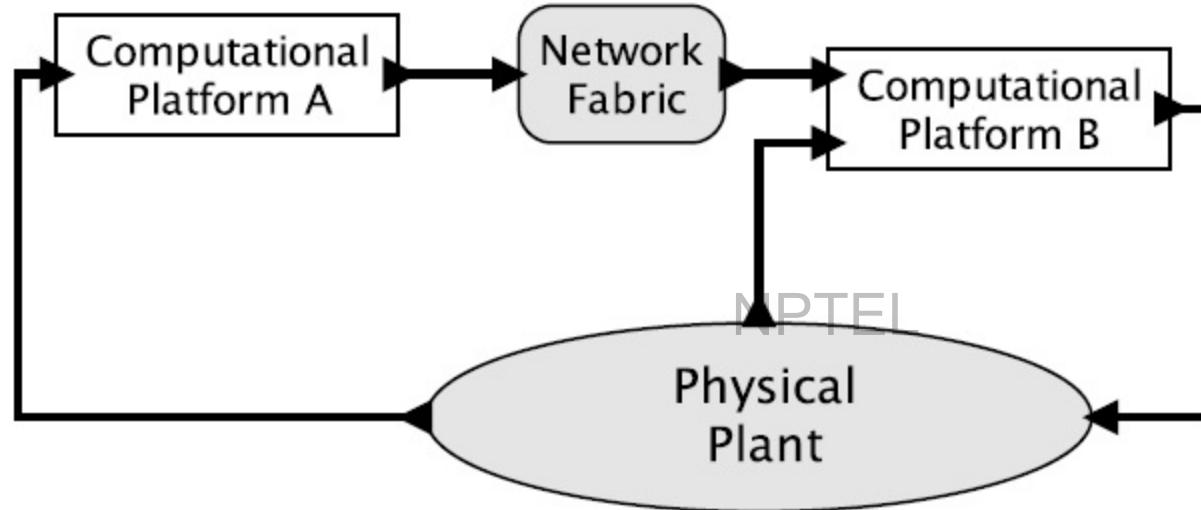


Cyber-Physical System (CPS)

- A cyber-physical system (CPS) is an orchestration of computers and physical systems. Embedded computers monitor and control physical processes, usually with feedback loops, where physical processes affect computations and vice versa.
- The term “cyber-physical systems” emerged around 2006, when it was coined by Helen Gill at the National Science Foundation , USA
- CPS is about the intersection, not the union, of the physical and the cyber. It combines engineering models and methods from mechanical, environmental, civil, electrical, biomedical, chemical, aeronautical and industrial engineering with the models and methods of computer science.
- Applications of CPS include automotive systems, manufacturing, medical devices, military systems, assisted living, traffic control and safety, process control, power generation and distribution, energy conservation etc.



Cyber-Physical System (CPS)



Cyber-Physical System (CPS)

- CPS describes a broad range of complex, multi-disciplinary, physically-aware next generation engineered system that integrates embedded computing technologies (cyber part) into the physical world.
- In cyber-physical systems, physical and software components are deeply intertwined, able to operate on different spatial and temporal scales, exhibit multiple and distinct behavioral modalities, and interact with each other in ways that change with context.
- CPS involves transdisciplinary approaches, merging theory of cybernetics, mechatronics, design and process science.
- Cyber + Physical + Computation + Dynamics + Communication + Security + Safety

NPTEL



Cyber-Physical System (CPS)

- Cyber physical systems (CPS) are an emerging discipline that involves engineered computing and communicating systems interfacing the physical world.
- Ongoing advances in science and engineering improve the tie between computational and physical elements by means of intelligent mechanisms, increasing the adaptability, autonomy, efficiency, functionality, reliability, safety, and usability of cyber-physical systems.
- Potential applications of cyber-physical systems are in several areas, including: *intervention* (e.g., collision avoidance); *precision* (e.g., robotic surgery and nano-level manufacturing); *operation in dangerous or inaccessible environments* (e.g., search and rescue, firefighting, and deep-sea exploration); *coordination* (e.g., air traffic control, war fighting); *efficiency* (e.g., zero-net energy buildings); and *augmentation of human capabilities* (e.g. in healthcare monitoring and delivery).
- Typical examples of CPS include : smart grid, autonomous automobile systems, medical monitoring, industrial control systems, robotics systems, and automatic pilot avionics.

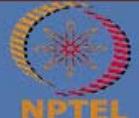
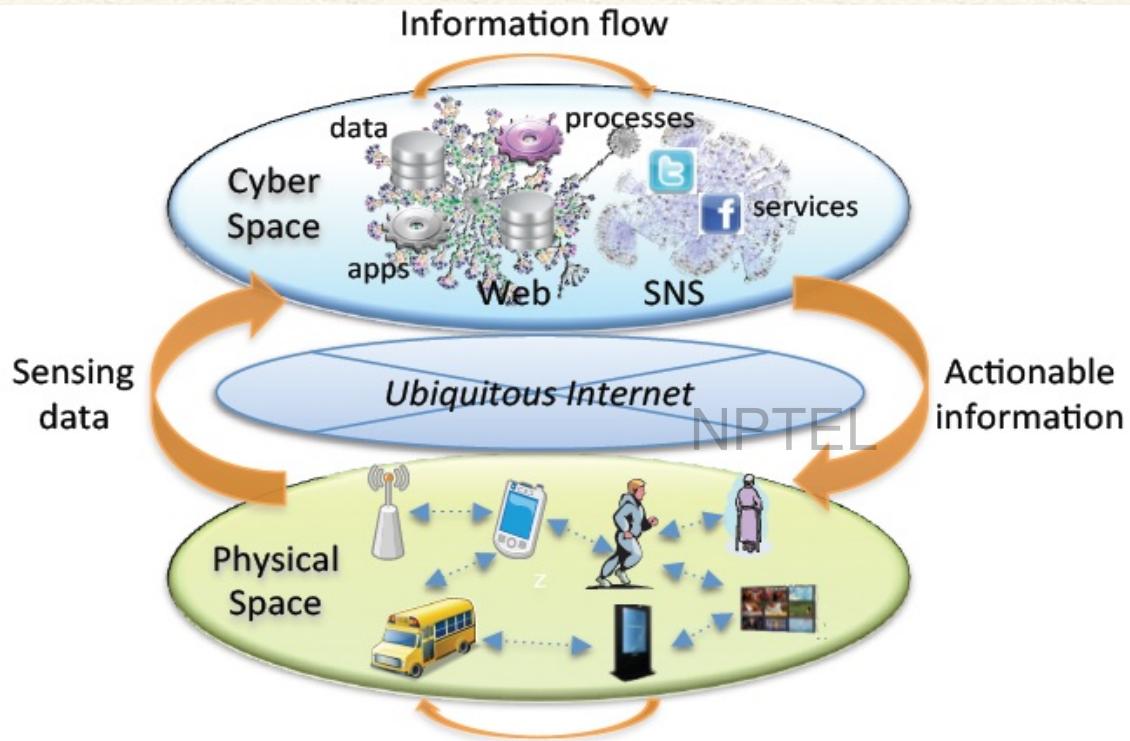


Cyber-Physical System (CPS)

- The interlinked networks of sensors, actuators and processing devices create a vast network of connected computing resources, things and humans.
- A CPS is the “integration of computation with physical processes” and uses sensors and actuators to link the computational systems to the physical world.
- CPS can be viewed as “computing as a physical act” where the real world is monitored through sensors that transfer sensing data into the cyberspace where cyber applications and services use the data to affect the physical environment
- ***Cloud Computing Services*** provide a flexible platform for realizing the goals of CPS



Cyber-Physical System (CPS)



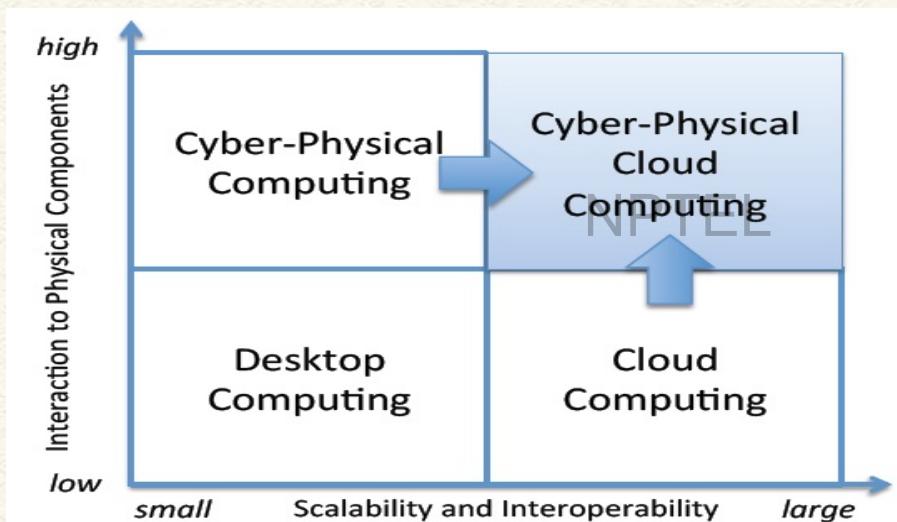
Cyber-Physical System (CPS)

- The interlinked networks of sensors, actuators and processing devices create a vast network of connected computing resources, things and humans that we will refer to as a Smart Networked Systems and Societies (SNSS).
- A CPS is the “integration of computation with physical processes” and uses sensors and actuators to link the computational systems to the physical world.
- CPS can be viewed as “computing as a physical act” where the real world is monitored through sensors that transfer sensing data into the cyberspace where cyber applications and services use the data to affect the physical environment
- ***Cloud Computing Services*** provide a flexible platform for realizing the goals of CPS
- A Cyber-Physical Cloud Computing (CPCC) architectural framework is defined as “a system environment that can rapidly build, modify and provision cyber-physical systems composed of a set of cloud computing based sensor, processing, control, and data services.”



CPS and Cloud - Cyber-Physical Cloud Computing (CPCC)

- A Cyber-Physical Cloud Computing (CPCC) architectural framework can be defined as “a system environment that can rapidly build, modify and provision cyber-physical systems composed of a set of cloud computing based sensor, processing, control, and data services.”



Ref: A Vision of Cyber-Physical Cloud Computing for Smart Networked Systems, NIST Report NIST, USA, August 2013



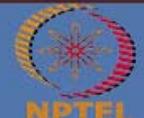
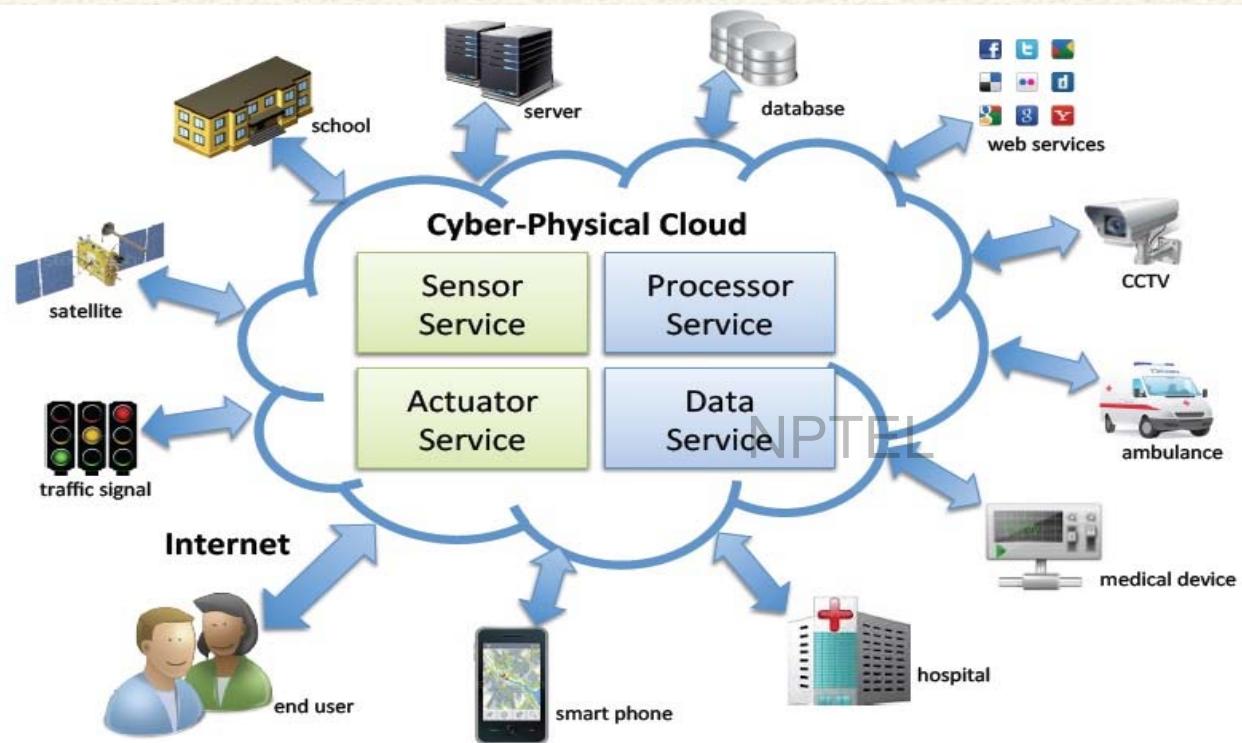
CPCC Benefits

- Efficient use of resources
- Modular composition
- Rapid development and scalability
- Smart adaptation to environment at every scale
- Reliable and resilient architecture

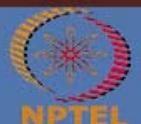
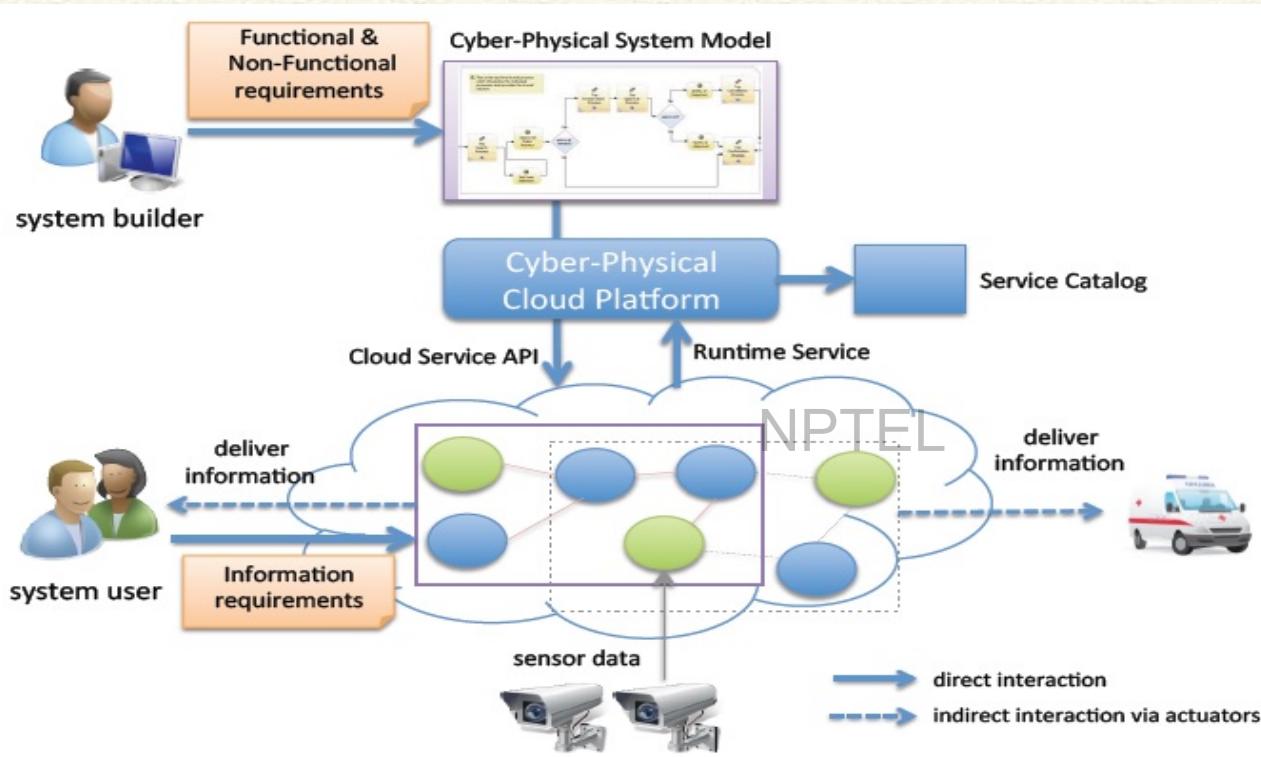
NPTEL



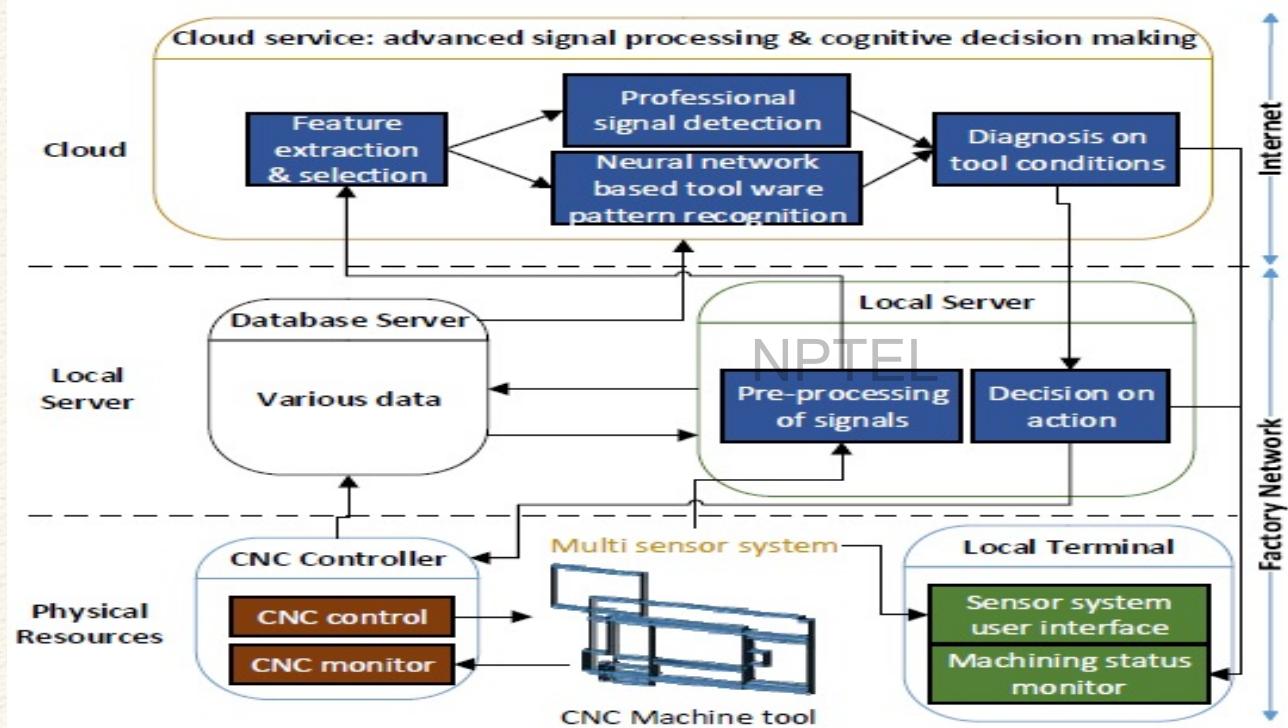
CPS and Cloud



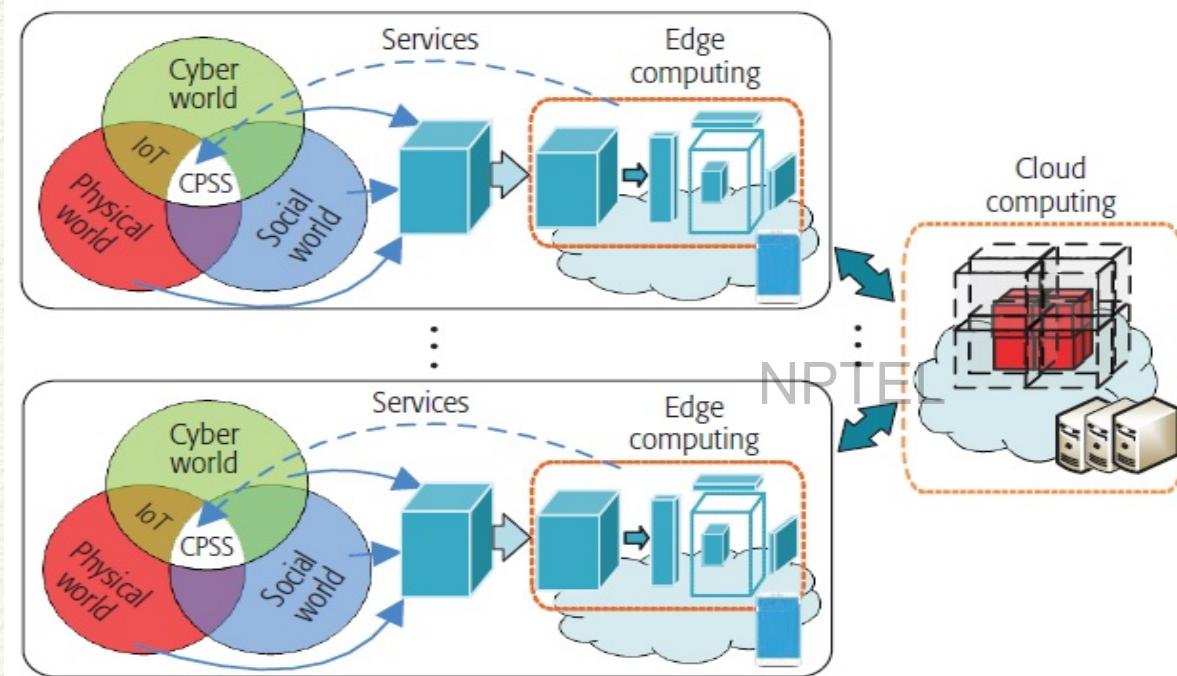
High level CPCC Scenario



A Cloud-based CPS architecture for Intelligent Monitoring of Machining Processes



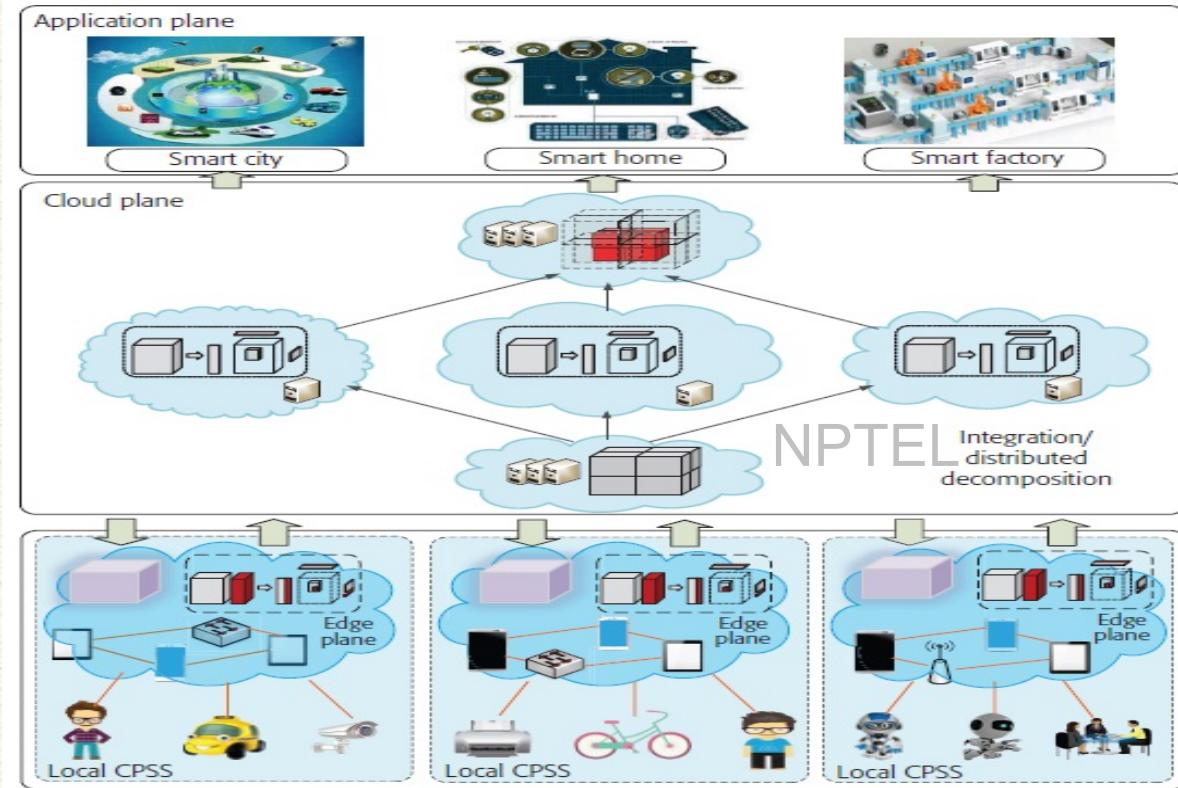
Cloud-Edge Computing Framework for CPS



Ref: X. Wang, L. T. Yang, X. Xie, J. Jin and M. J. Deen, "A Cloud-Edge Computing Framework for Cyber-Physical-Social Services," in IEEE Communications Magazine, vol. 55, no. 11, pp. 80-85, Nov. 2017, doi: 10.1109/MCOM.2017.1700360



Cloud-Edge Computing Framework for CPS



REFERENCES

- https://en.wikipedia.org/wiki/Cyber-physical_system
- A Vision of Cyber-Physical Cloud Computing for Smart Networked Systems, NIST Report NIST, USA, August 2013
- Architecture of Cyber-Physical Systems Based on Cloud, Shaojie Luo, Lichen Zhang, Nannan Guo, Proceedings of IEEE 5th Intl Conference on Big Data Security on Cloud (BigDataSecurity), 2019
- Lee EA. The Past, Present and Future of Cyber-Physical Systems: A Focus on Models. Sensors. 2015; 15(3):4837-4869. <https://doi.org/10.3390/s150304837>
- X. Wang, L. T. Yang, X. Xie, J. Jin and M. J. Deen, "A Cloud-Edge Computing Framework for Cyber-Physical-Social Services," in IEEE Communications Magazine, vol. 55, no. 11, pp. 80-85, Nov. 2017, doi: 10.1109/MCOM.2017.1700360.



*Thank
you*



NPTEL



NPTEL



NPTEL ONLINE CERTIFICATION COURSES

Cloud Computing

Prof. Soumya K Ghosh

**Department of Computer Science
and Engineering**

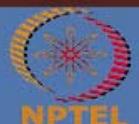
Module 12: Cloud Computing Paradigms

Lecture 58: Case Study I (Spatial Cloud Computing)

CONCEPTS COVERED

- Spatial Data
- Spatial Cloud
- Spatial Analysis on Cloud

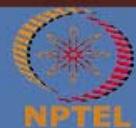
NPTEL



KEYWORDS

- Spatial Data
- Spatial Cloud Computing

NPTEL



Spatial Analysis on Cloud

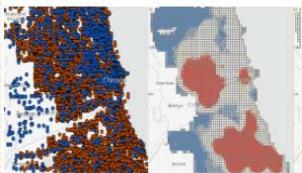
NPTEL



Spatial Data and Analysis

- Spatial (or Geospatial) data is information that describes objects, events or other features with a location on or near the surface of the earth.
- Geospatial data typically combines location information (usually coordinates on the earth) and attribute information (the characteristics of the object, event or phenomena concerned) with temporal information (the time or life span at which the location and attributes exist).

Whenever we look at a map, we inherently start turning that map into information by analyzing its contents—finding patterns, assessing trends, or making decisions.



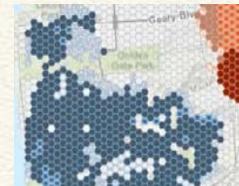
Crime Studies



Drought Analysis



Finding optimal paths

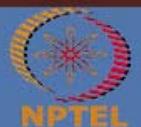


Predictions

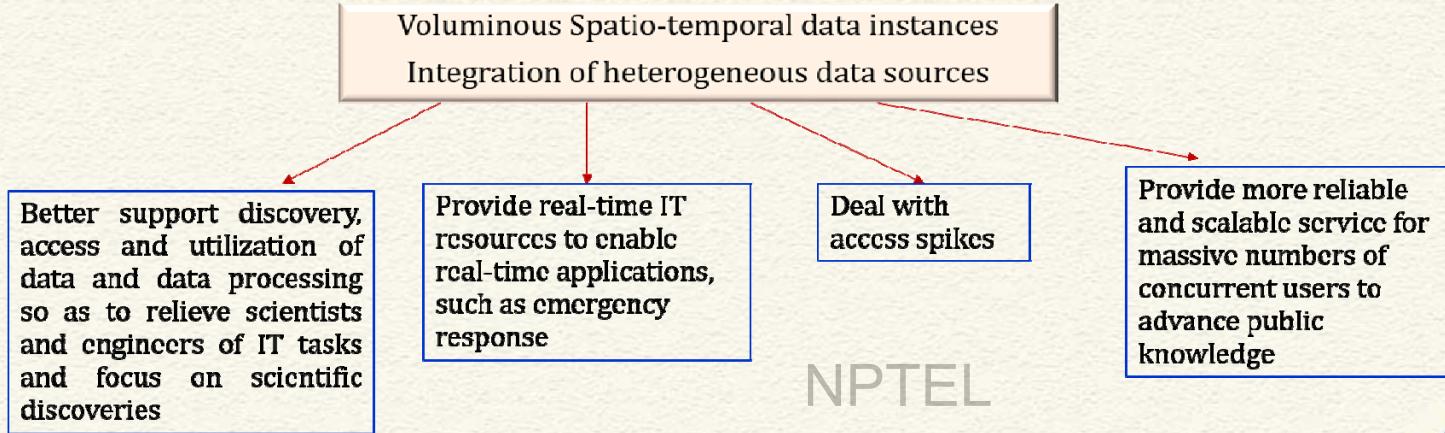


Spatial Analysis

- Attempt to solve location-oriented problems and better understanding of where and what is occurring in surrounding world/ region.
 - Beyond mapping - study the characteristics of places/ regions and the relationships between them
- Spatial analysis lends new perspectives to any decision-making
- Spatial analysis lets you pose questions and derive answers on spatial data.
- Help to derive new information and make informed decisions.
- The organizations that use spatial analysis in their work are wide-ranging—local and state governments, national agencies, businesses of all kinds, utility companies, colleges and universities, NGOs...



Spatial Analysis - Challenges



NPTEL

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



Spatial Analytics + Cloud Computing

Emergence of cloud computing provides a potential solution with an *elastic, on-demand computing platform to integrate – observation systems, parameter extracting algorithms, phenomena simulations, analytical visualization and decision support*, and to provide social impact and user feedback

- *Search, access and utilize* geospatial data
- *Configure computing infrastructure* to enable the computability of intensive simulation models disseminate and utilize research results for massive numbers of concurrent users
- *Adopt spatiotemporal principles* to support spatiotemporal intensive applications

Spatial cloud computing refers to the cloud computing paradigm that is driven by geospatial sciences, and optimized by spatiotemporal principles for enabling geospatial science discoveries and cloud computing within distributed computing environment

NPTEL



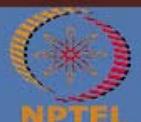
Spatial Cloud

- It supports shared resource pooling which is useful for participating organizations with common or shared goals
 - Network, Servers, Apps, Services, Storages and Databases
- Choice of various deployment, service and business models to best suit organization goals
- Managed services prevent data loss from frequent outages, minimizing financial risks, while increasing efficiency

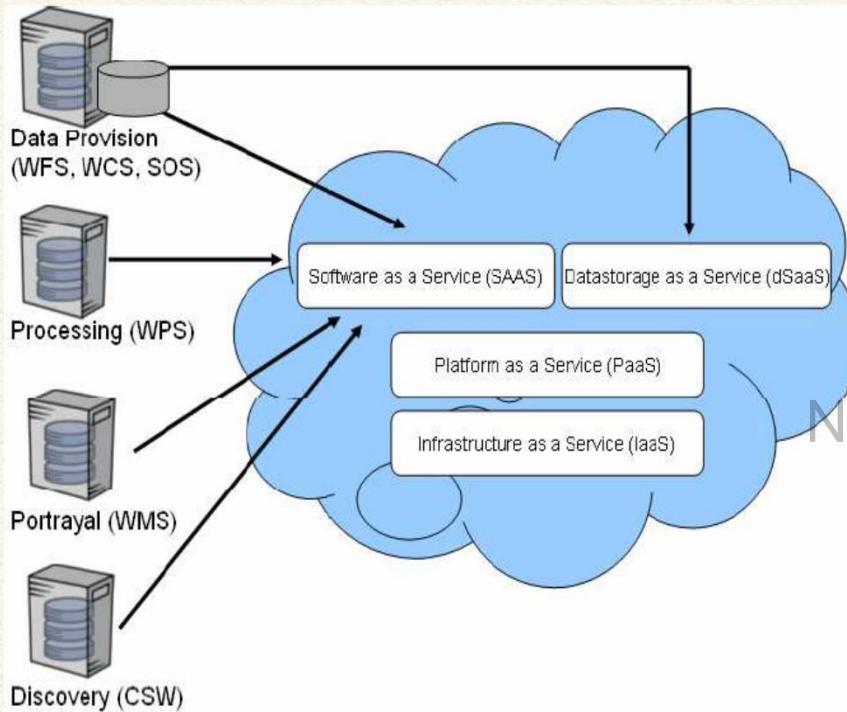


Spatial Cloud - Advantages

- **Easy to Use-** Infrastructure deployment with click of mouse, API and Network.
- **Scalability-** Infrastructure requirement is based on application, nothing to purchase.
- **Cost-** Optimized as it is resource usage based
- **Reliability-** Based on Enterprise grade Hardware; can subscribe to multiple clouds.
- **Risk-** Change instantly (even OS).



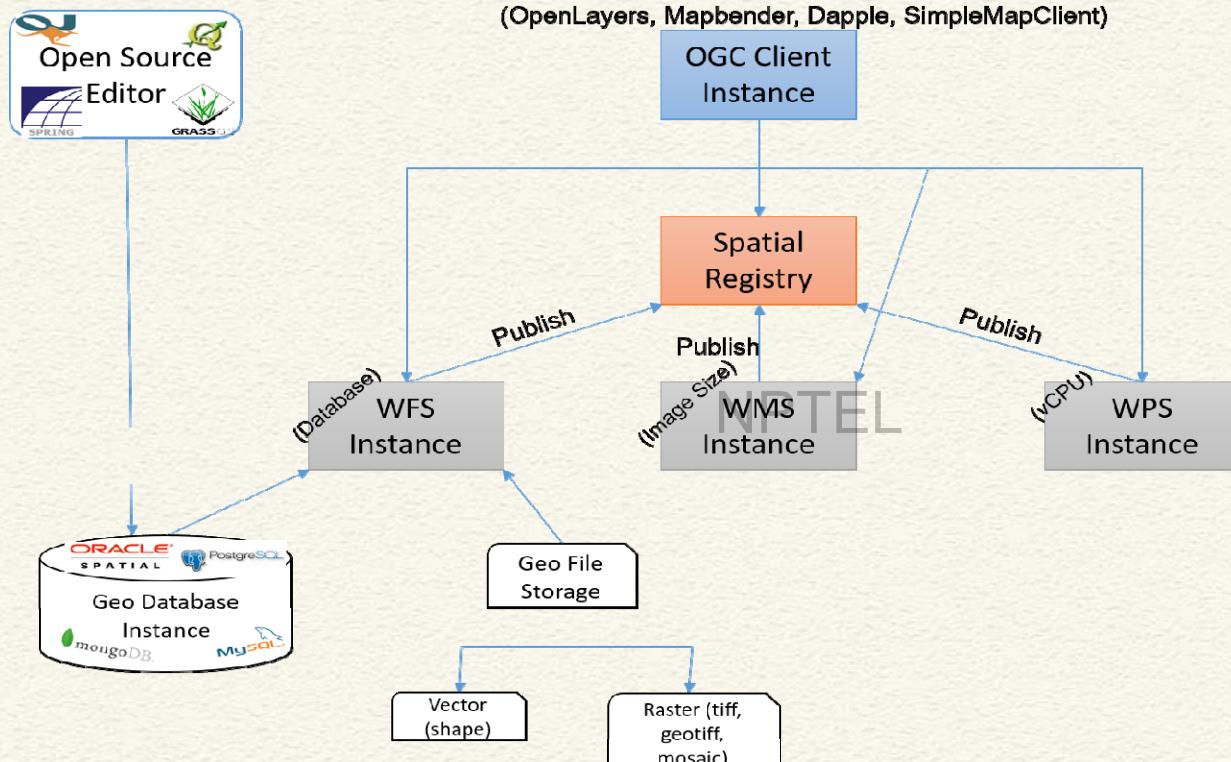
Spatial Cloud – Typical Architecture



- **Private and public organization wants to share their spatial data**
 - Different requirement of geospatial data space and network bandwidth
- **Easy access of spatial services**
- **GIS decisions are made easier**
 - Integrate latest databases
 - Merge disparate systems
 - Exchange information internally and externally

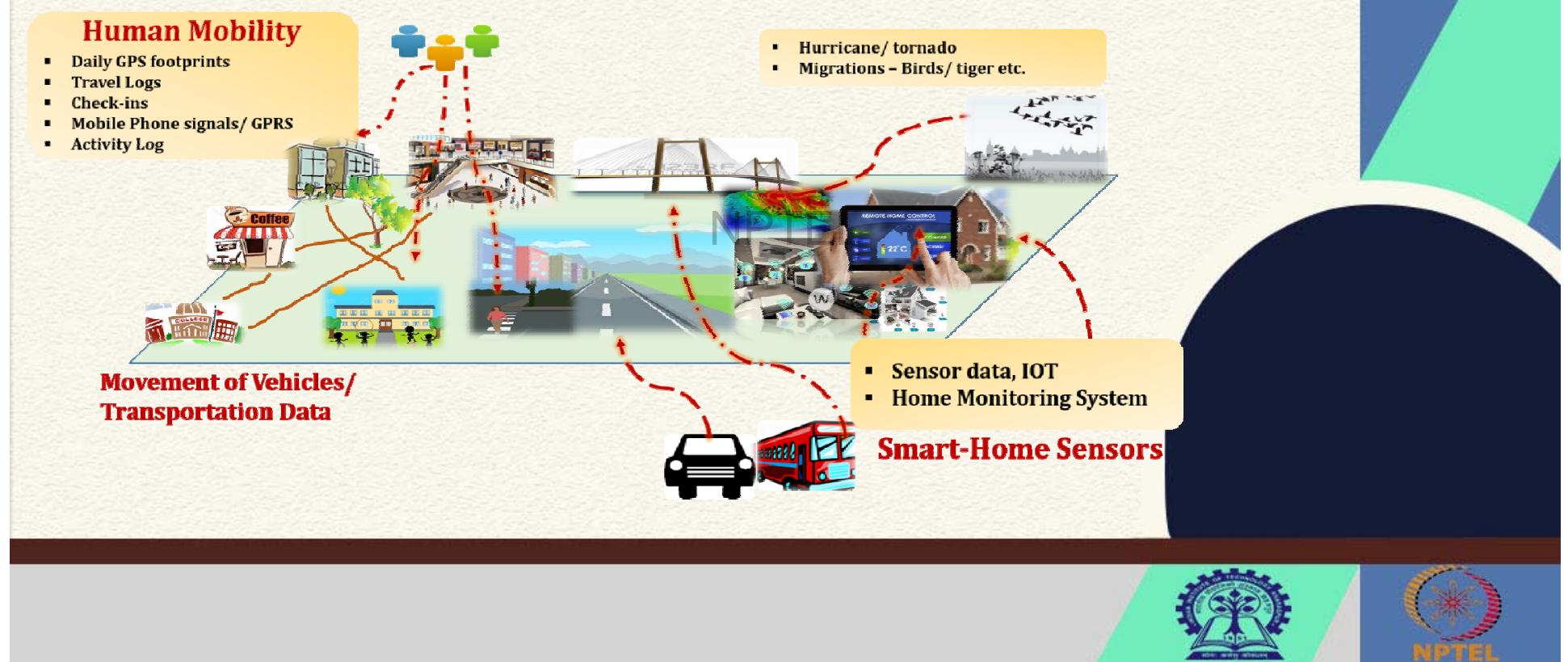


Spatial Cloud – Typical Architecture

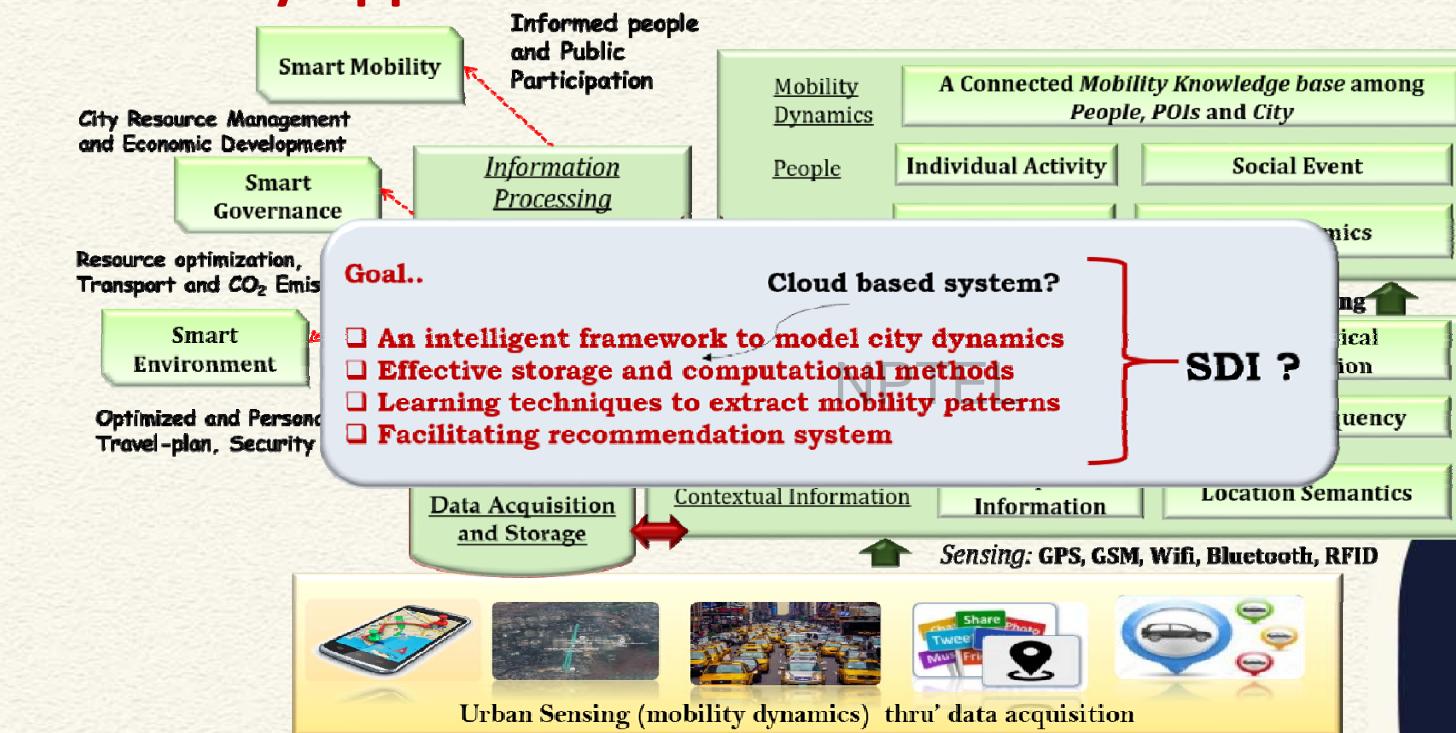


Mobility Analytics

(Utilize Cloud platform for computation and storage)



A general framework of Trajectory Trace Mining for Smart-City Applications



A Trajectory Cloud for enabling Efficient Mobility Services



Spatial Trajectory

- ❑ A **spatial trajectory** is a trace generated by a moving object in geographical spaces, usually represented by a series of chronologically ordered points
- ❑ Sequence of time stamped locations (latitude, longitude):
$$<(lat_1, lon_1), t_1>, <(lat_2, lon_2), t_2>, \dots <(lat_n, lon_n), t_n>$$

Semantic Trajectory

- ❑ “**Human movement follows an intent**” – How to capture the implicit knowledge/information?
- ❑ For better understanding additional information (stay-point information, activity performed at stay points?) are appended



NPTEL

Traj-Cloud for analyzing Urban Dynamics

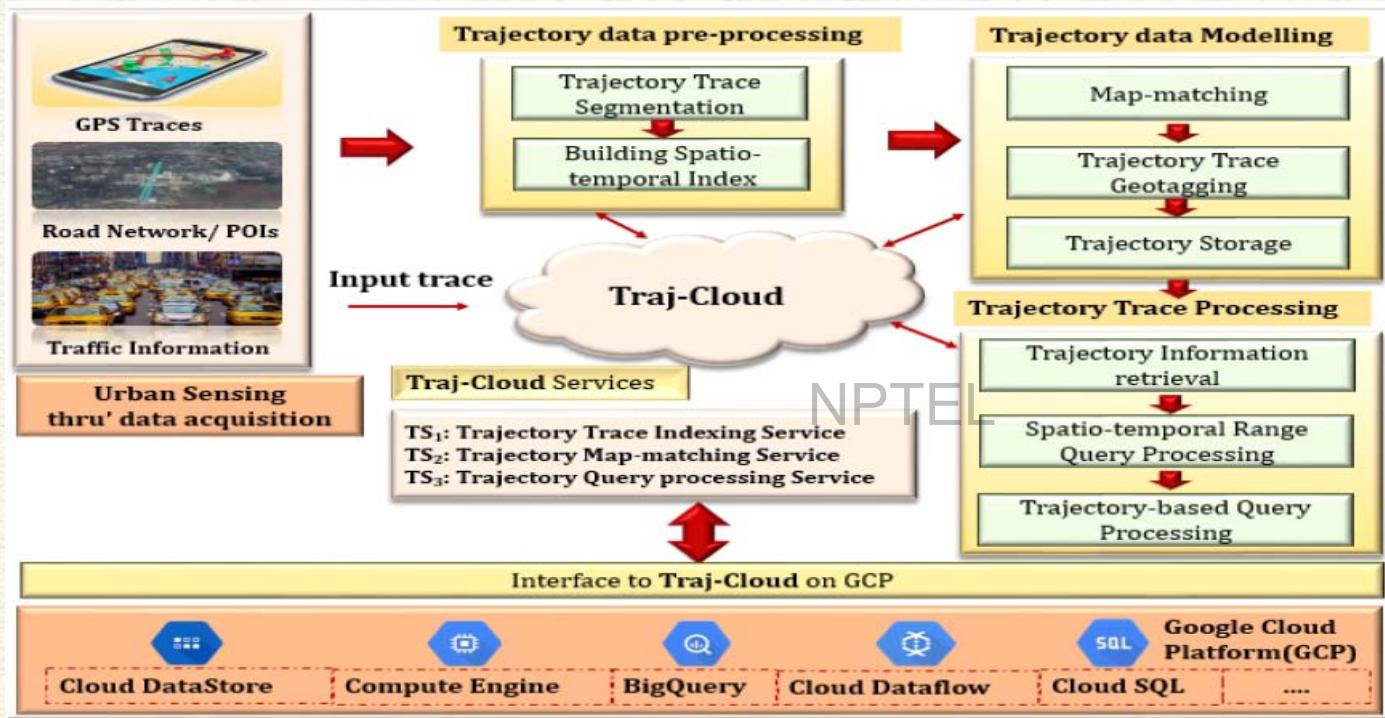
- Mobility trace analysis has a significant role in mapping the *urban dynamics*.
 - This analysis helps in location-based service-provisioning and facilitates an effective transportation resource planning.
- Key aspect of the intelligent transportation system (ITS) is efficient mobility analytics to understand the movement behaviours of the people.
- Analysing mobility traces and providing location-aware service is a challenging task.

NPTEL

- An end-to-end cloud-based framework may facilitate efficient location-based service provisioning.
- It helps to minimize the service-waiting time and service-provisioning time of location-based services such as food delivery or medical emergency



Traj-Cloud for analyzing Urban Dynamics



Traj-Cloud Services

Trajectory data Indexing Service (TS₁):

- *Input:* GPS trajectory trace (G) and other semantic information, such as, geotagged locations or road network
- *Output:* Spatio-temporal indices of input traces and storage of the information
- *GCP Component:* Google BigQuery and Cloud SQL storage.

Trajectory Map-matching Service (TS₂):

- *Input:* GPS trajectory trace (G) and road network (R)
- *Output:* Projection of G into the corresponding R utilizing the MapReduce based platform to effectively handle huge data load in near real-time.
- *GCP Component:* Google Compute Engine

Trajectory Query Service (TS₃):

- *Input:* GPS trajectory trace (G) log, Trajectory point and range Query Q
- *Output:* Trajectory Trace (Point or Line shape)
- *GCP Component:* Google Compute Engine and Cloud SQL



REFERENCES

- S. Ghosh and S. K. Ghosh, "Traj-Cloud: A Trajectory Cloud for enabling Efficient Mobility Services," *2019 11th International Conference on Communication Systems & Networks (COMSNETS)*, 2019, pp. 765-770, doi: 10.1109/COMSNETS.2019.8711428.
- Shreya Ghosh, "Semantic Analysis of Trajectory Traces to Explore Human Movement Behaviour", PhD Thesis, Department of Computer Science and Engineering, Indian Institute of Technology Kharagpur, India, 2021.



*Thank
you*



NPTEL





NPTEL ONLINE CERTIFICATION COURSES

Cloud Computing

Prof. Soumya K Ghosh

**Department of Computer Science
and Engineering**

Module 12: Cloud Computing Paradigms

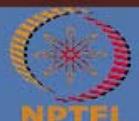
Lecture 59: Case Study II (Internet of Health Things)

(Part-A)

CONCEPTS COVERED

- Cloud-Fog-Edge-IoT Framework
- Internet of Health Things (IoHT)
- Case Study on Cloud-Fog-Edge-IoHT

NPTEL



KEYWORDS

- Internet of Health Things (IoHT)

NPTEL



Cloud-Fog-Edge Computing for Internet of Health Things (IoHT)

NPTEL

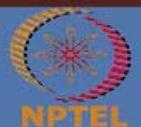
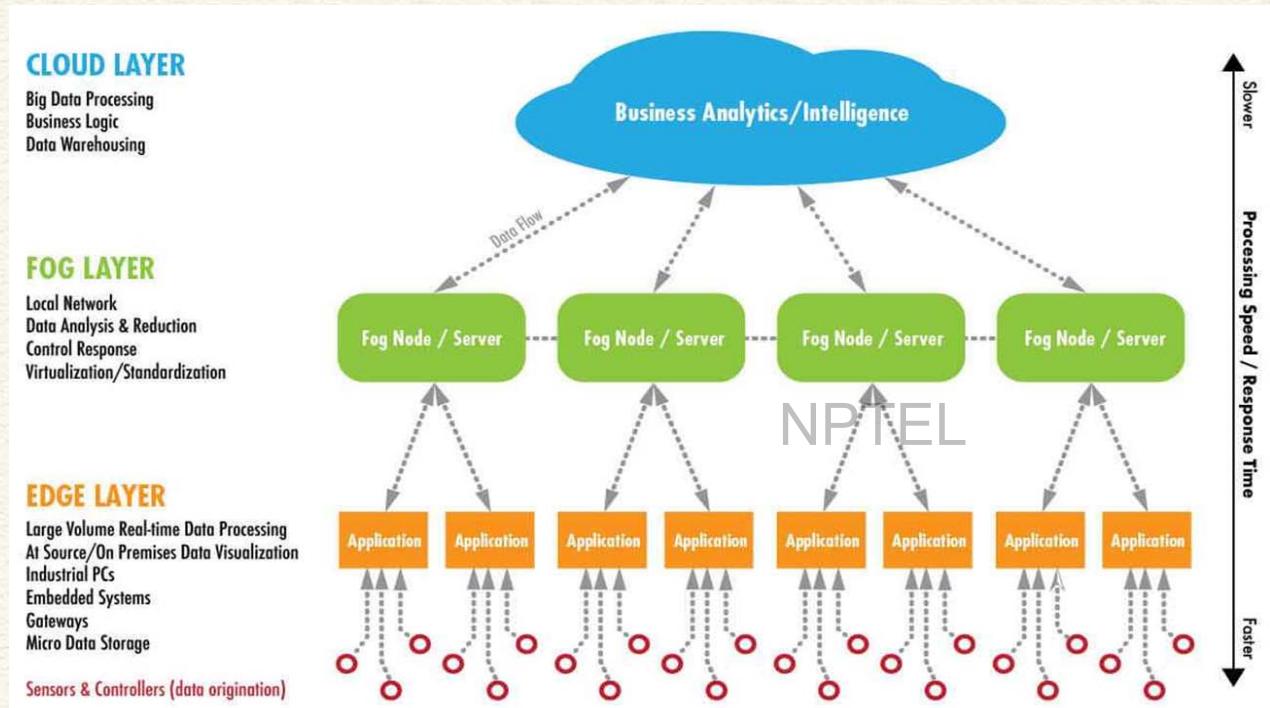


Fog Computing

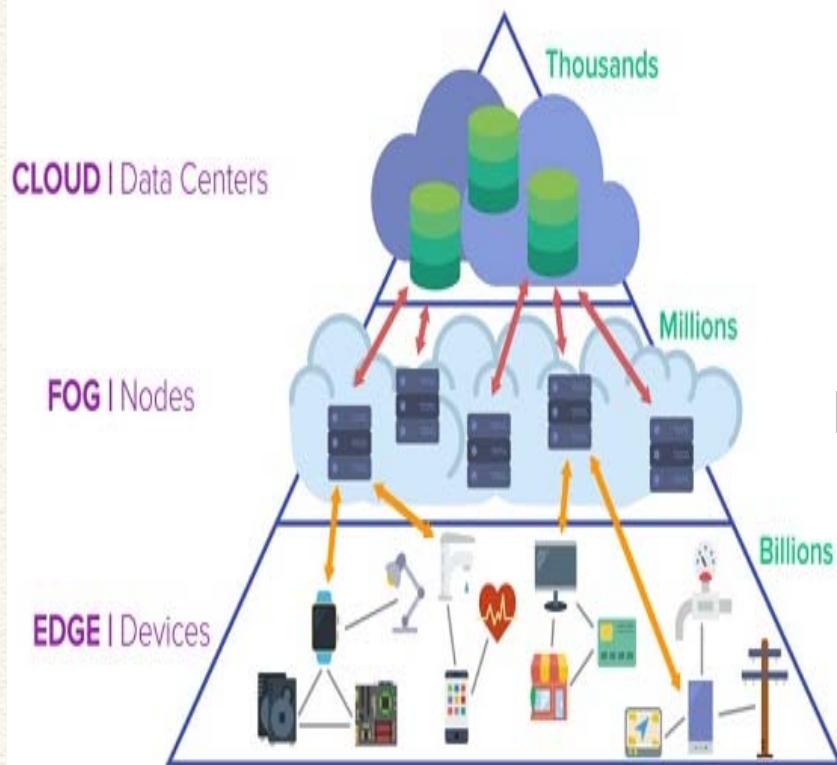
- Fog Computing takes the cloud closer to the data producing sensor devices. Devices such as routers, servers, switches act as fog nodes if their processing power is employed for data processing and result generation.
- Use of Fog technology for real time applications
- Aim is to develop a Fog Based Healthcare model based on data collected by IoT based health sensor.
- Collected data will be processed at Edge devices to reduce latency, network usage and overall cost incurred at the cloud.
- The performance to be evaluated using simulator tool as well as actual hardware



Cloud-Fog-Edge-IoT



Cloud-Fog-Edge Hierarchy



Cloud Limitations

- Latency
- Large volume of data being generated.
- Bandwidth requirement

IoT Device Limitations

- Processing
- Storage
- Power requirement

NPTEL

Fog-Edge Computing

- Reduced **latency** supports Real-time applications
- Less **network congestion**
- Reduced **cost of execution** at cloud
- Better handling of colossal data generated by sensors
- More of data location awareness



Cloud-Fog-Edge-IoHT

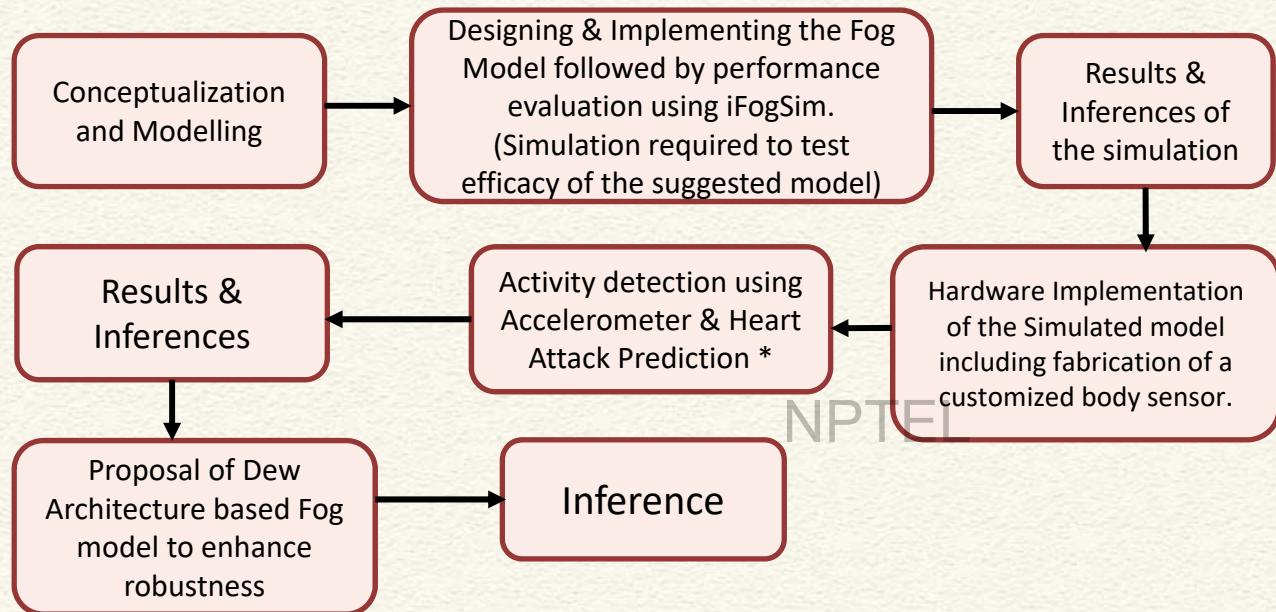
Objectives

- To design a Fog-Edge Computing based health model to reduce latency, network usage and cost incurred at the cloud.
- To test the designed fog model using iFogSim simulator.
- To develop a customized wearable device for collection of health parameters.
- To implement the proposed model over hardware and test its efficacy.
- To study dew based computing and study its efficacy in the proposed health scenario

NPTEL



Overall Workflow

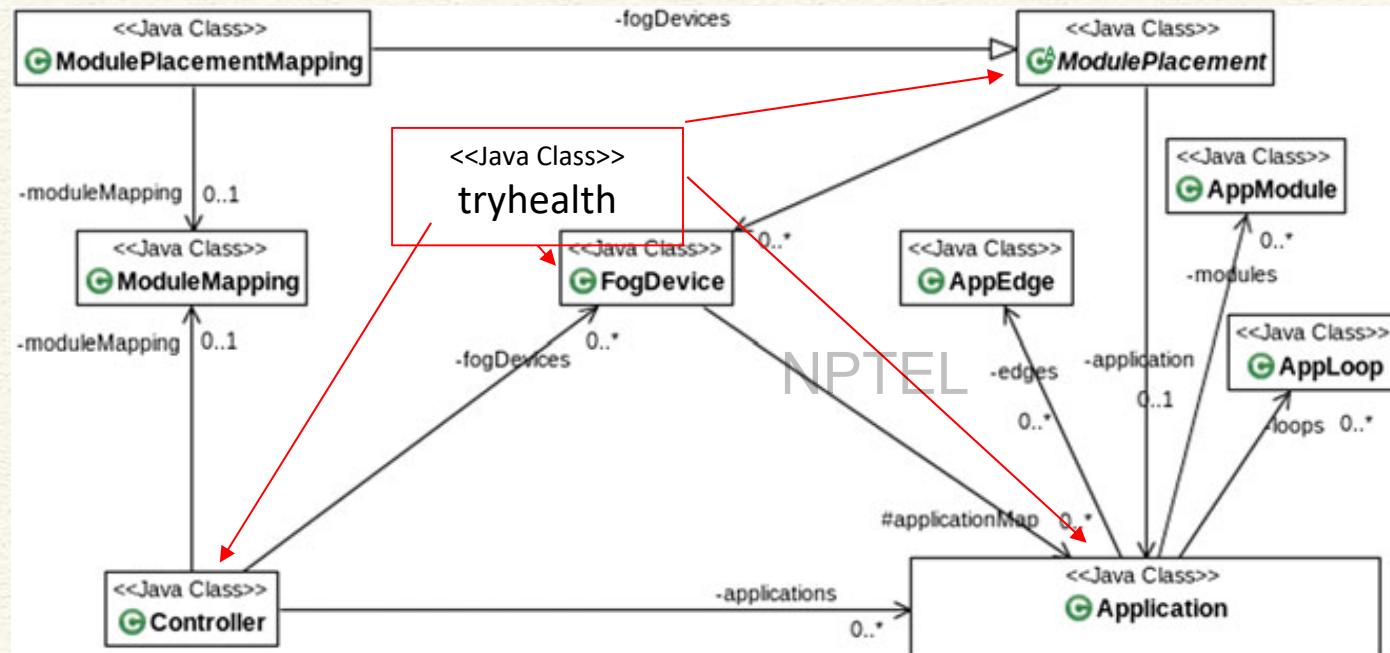


Note:

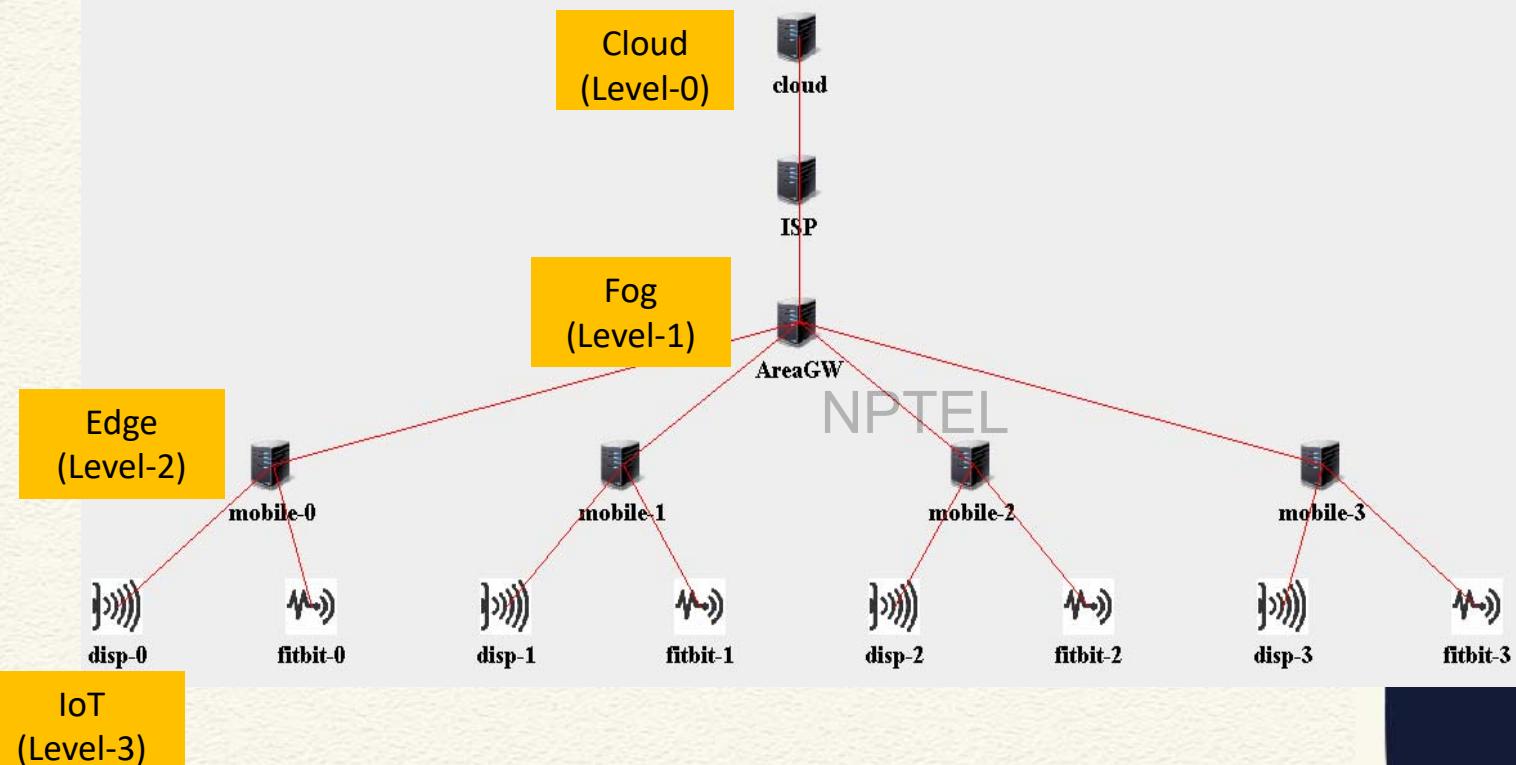
*Heart Attack Prediction algorithm has no medical/ clinical implication and has been used only for demonstration purposes.



Simulation using iFogSim

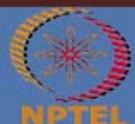


Hierarchical Network Topology Model



REFERENCES

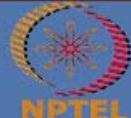
- Anish Poonia, MTech Dissertation, IIT Kharagpur, Fog Computing For Internet of Health Things, 2020
- Anish Poonia, Shreya Ghosh, Akash Ghosh, Shubha Brata Nath, Soumya K. Ghosh, Rajkumar Buyya, CONFRONT: Cloud-fog-dew based monitoring framework for COVID-19 management, Internet of Things, Elsevier, Volume 16, 2021
- Cisco White Paper. 2015. Fog Computing and the Internet of Things: Extend the Cloud to Where the Things Are.
- Gupta H, Vahid Dastjerdi A, Ghosh SK, Buyya R. iFogSim: A toolkit for modeling and simulation of resource management techniques in the Internet of Things, Edge and Fog computing environments. *Softw Pract Exper.* 2017;47:1275-296. <https://doi.org/10.1002/spe.2509>
- Luiz Bittencourt et al., The Internet of Things, Fog and Cloud continuum: Integration and challenges, Internet of Things, Volumes 3–4, 2018, Pages 134-155, ISSN 2542-6605, <https://doi.org/10.1016/j.iot.2018.09.005>



*Thank
you*



NPTEL





NPTEL ONLINE CERTIFICATION COURSES

Cloud Computing

Prof. Soumya K Ghosh

**Department of Computer Science
and Engineering**

Module 12: Cloud Computing Paradigms

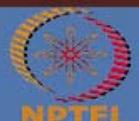
Lecture 60: Case Study II (Internet of Health Things)

(Part-B)

CONCEPTS COVERED

- Cloud-Fog-Edge-IoT Framework
- Internet of Health Things (IoHT)
- Case Study on Cloud-Fog-Edge-IoHT

NPTEL



KEYWORDS

- Internet of Health Things (IoHT)

NPTEL



Cloud-Fog-Edge Computing for Internet of Health Things (IoHT)

NPTEL



Cloud-Fog-Edge-IoHT

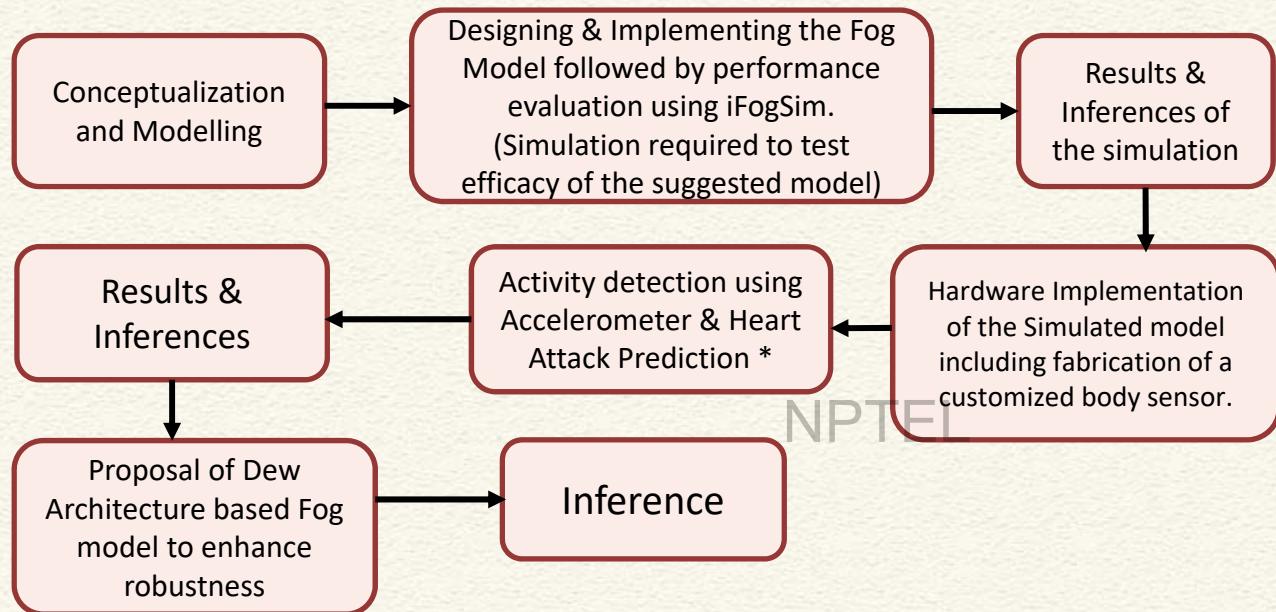
Objectives

- To design a Fog-Edge Computing based health model to reduce latency, network usage and cost incurred at the cloud.
- To test the designed fog model using iFogSim simulator.
- To develop a customized wearable device for collection of health parameters.
- To implement the proposed model over hardware and test its efficacy.
- To study dew based computing and study its efficacy in the proposed health scenario

NPTEL



Overall Workflow

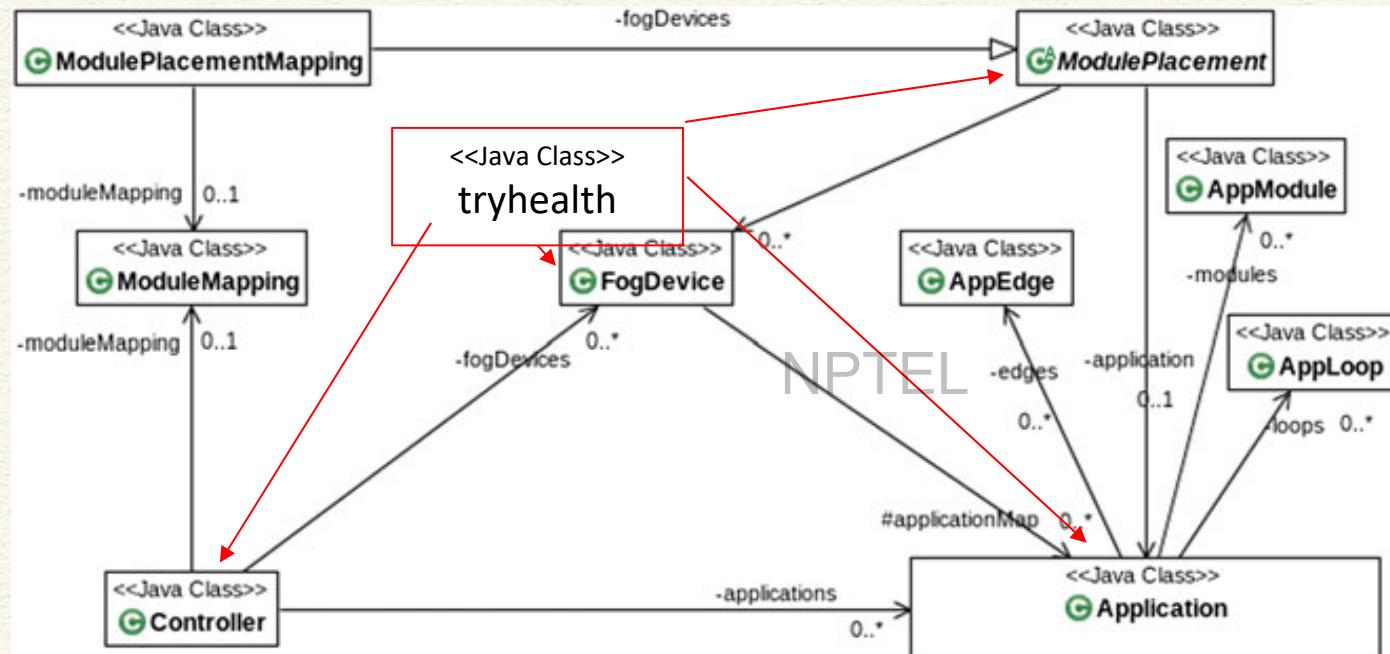


Note:

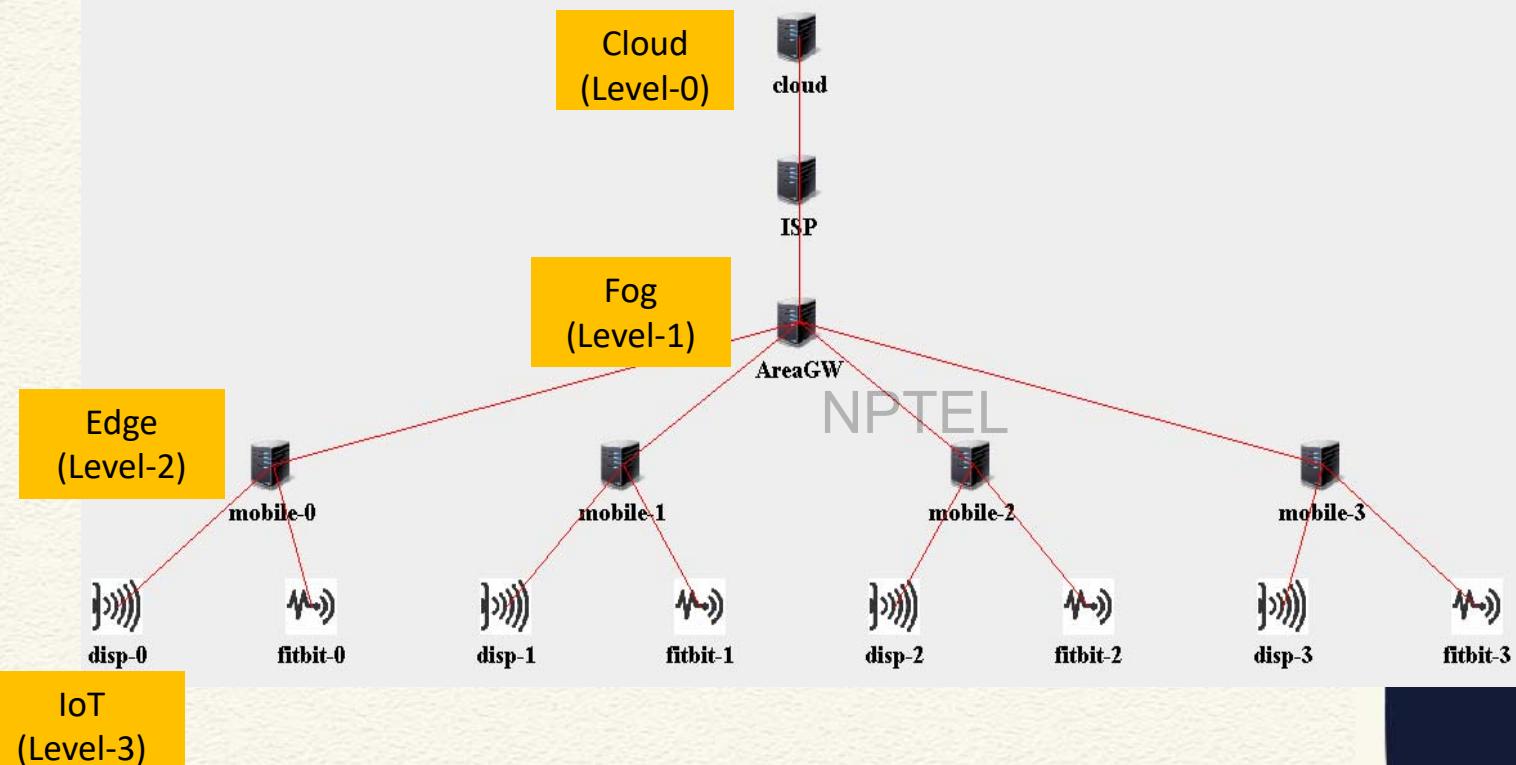
*Heart Attack Prediction algorithm has no medical/ clinical implication and has been used only for demonstration purposes.



Simulation using iFogSim



Hierarchical Network Topology Model



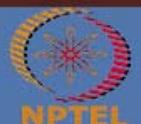
Cloud-Fog-Edge-IoHT – Typical Configuration

Device Configuration

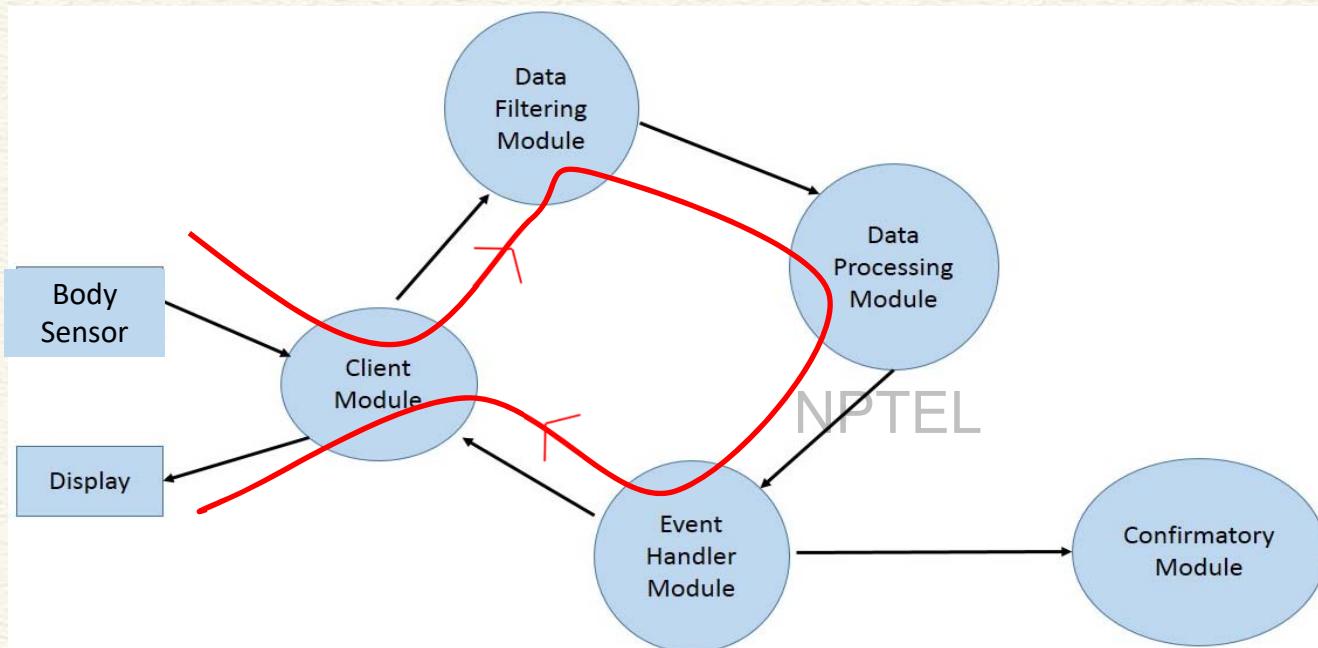
Device	MIPS	RAM (MB)	Up Bw (Kbps)	Down Bw (Kbps)	Level	Cost/MIPS	BusyPower (Watts)	Idle Power (Watts)
Cloud	44800	40000	100	10000	0	0.01	16*103	16*83.25
ISP	2800	4000	10000	10000	1	0	107.339	83.4333
AreaGW	2800	4000	10000	10000	2	0	107.339	83.4333
Mobile	350	1000	10000	270	3	0	87.53 mW	82.44 mW

Latency

Source	Destination	Latency
Body sensor	Mobile	1
Mobile	Area GW	2
Area GW	ISP GW	2
ISP GW	Cloud	100
Mobile	Display	1



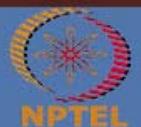
Cloud-Fog-Edge-IoHT – Process Flow



Application Placement

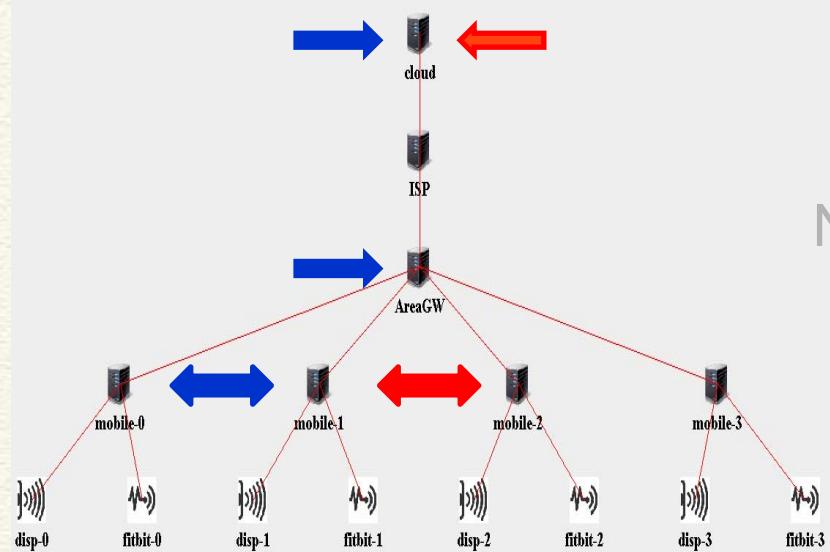
Application Module	Placement in Fog based Model	Placement in Cloudbased Model
Client Module	Mobile (Edge)	Mobile (Edge)
Data Filtering Module	Area Gateway (Fog)	Cloud
Data Processing Module	Area Gateway (Fog)	Cloud
Event Handler Module	Area Gateway (Fog)	Cloud
Confirmatory Module	Cloud	Cloud

NPTEL



Simulation Configuration

Configuration	No. of AreaGW	Total No. of Users
1	1	4
2	2	8
3	4	16
4	8	32
5	16	64

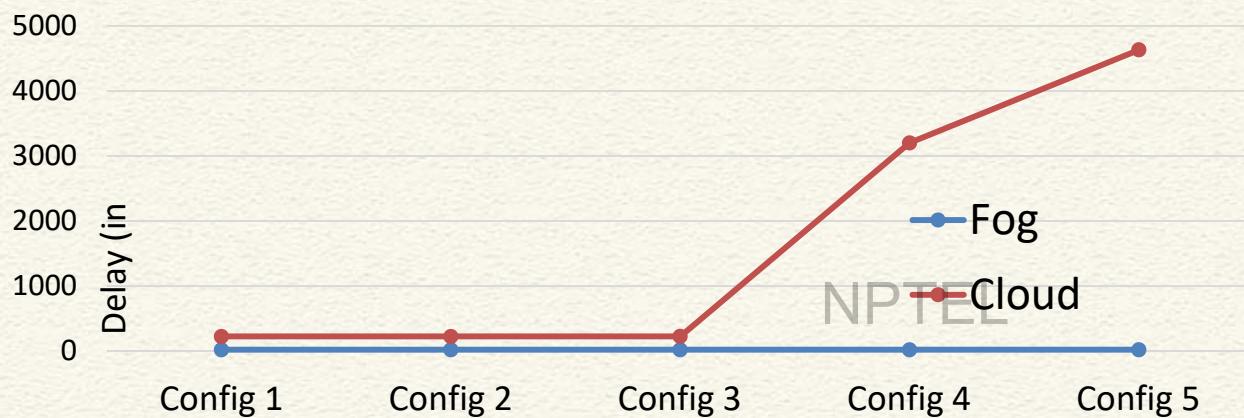


NPTEL



Performance Evaluation - Latency

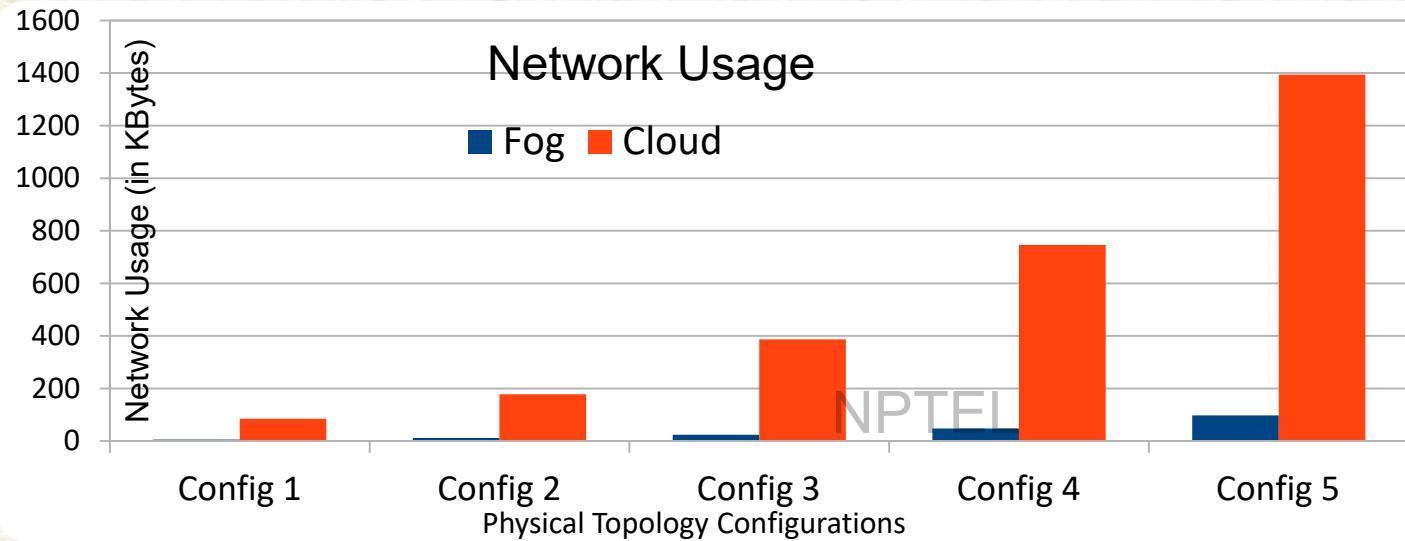
Average Latency of Control Loop



- Fog: Latency is fixed as the application modules which form part of the control loop are located at Area Gateway itself
- Cloud: Modules are located at the Cloud Datacenter



Performance Evaluation – Network Usage

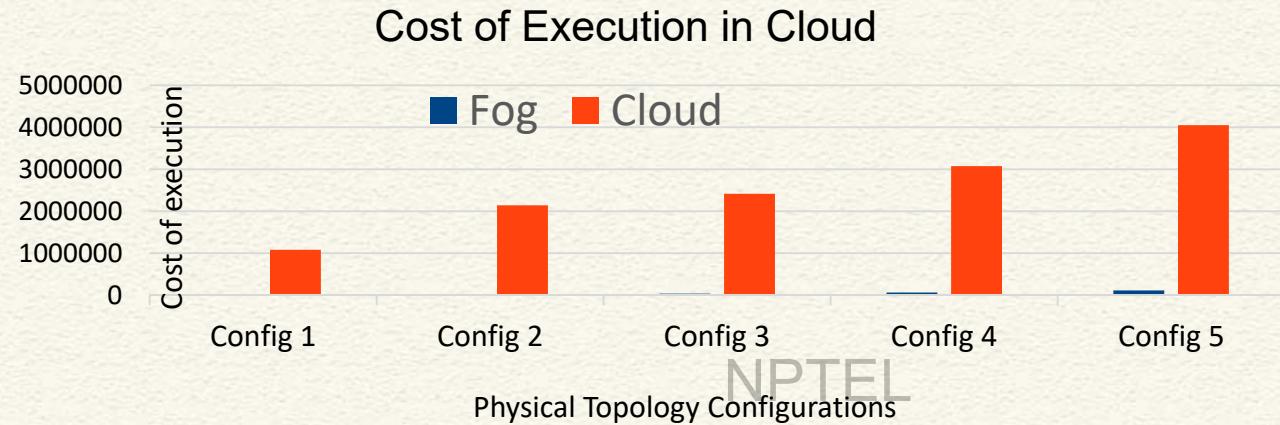


Fog: Network usage is very low as only for few positive cases, the Confirmatory module residing on Cloud is accessed.

Cloud: Network usage is high as all modules are now on Cloud.



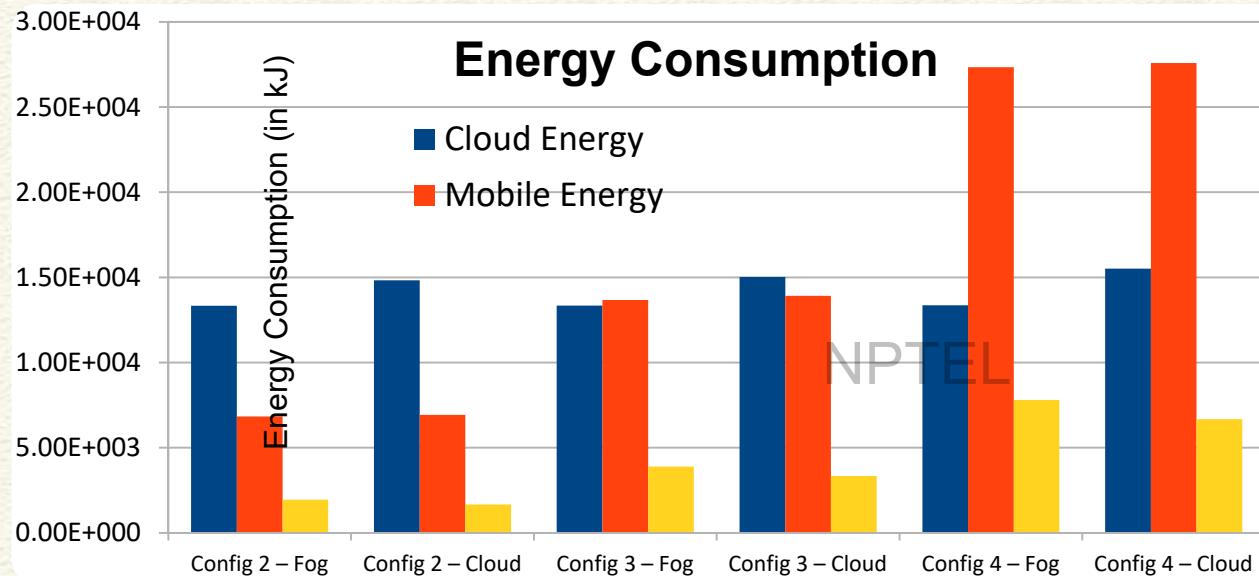
Performance Evaluation – Cost of Execution



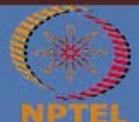
- Fog: Only the resources on Cloud incur cost, other resources are owned by the organization.
- Cloud: More processing at Cloud leads to higher costs in case of Cloud based architecture.



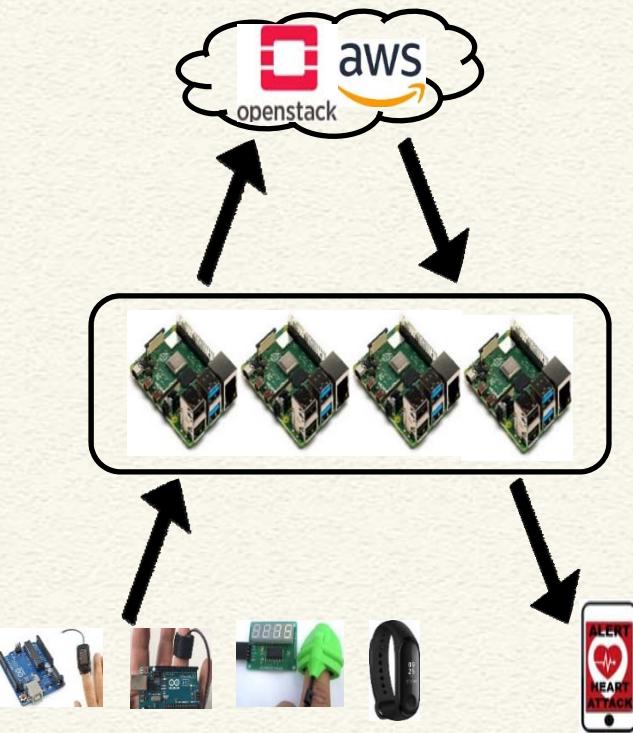
Performance Evaluation – Energy Consumption



- Energy consumption at Mobile devices remains same in Fog as well as Cloud as the load does not change.
- Energy requirement at the fog devices and Datacenter changes as the configuration changes from Fog based to Cloud based architecture owing to shifting of Application modules.



Hardware Implementation



- Simulated model's hardware implementation done using :
 - Customized body sensor
 - Simulated sensor data
 - Raspberry Pi as Fog Devices
 - AWS as Cloud

NPTEL



Hardware Implementation

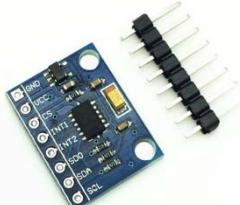
- Customized BP and Pulsemeter

Device has been customized to output serial data at 9600 baud rate in ASCII format.



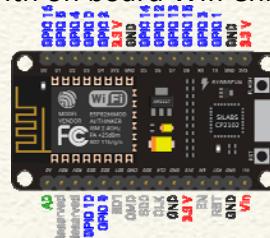
- Accelerometer (ADXL345)

Each value has three components: X-axis, Y-axis and Z-axis



- NodeMCU ESP8266 CP2102 Board

Arduino like Hardware IO with on board Wifi Chip



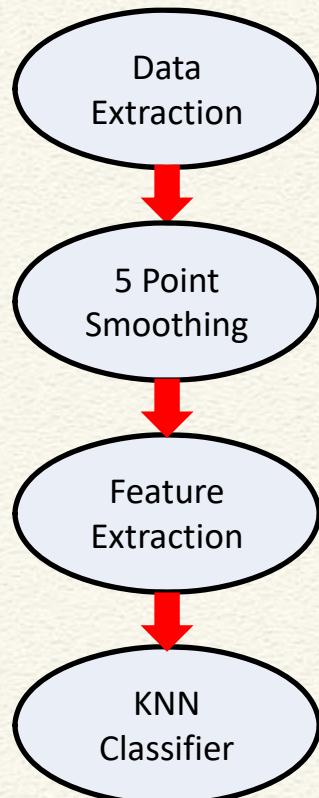
NPTEL

- Raspberry Pi 3 (Fog Device)

64 bit system with 1 GB RAM, wifi and Bluetooth connectivity.



Activity Detection using Accelerometer



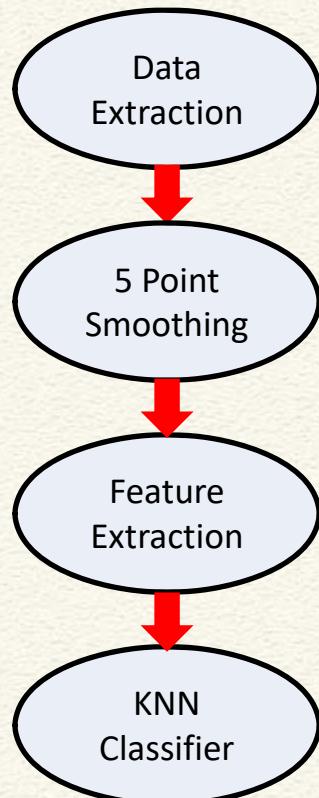
- **Data Extraction** - The collected data has three components: x-axis, y-axis, z-axis.

$$A = \sqrt{x^*x + y^*y + z^*z}$$

- **5-Point smoothing** - To reduce any induced noise, each signal is obtained as an average of five signals; two preceding signals, the signal itself and two succeeding signals



Activity Detection using Accelerometer



- Feature Extraction: Following features were extracted from the filtered signal
Maximum Amplitude
 - Minimum Amplitude
 - Mean Amplitude
 - Standard Deviation in Amplitude
 - Energy in Time Domain
 - Energy in Frequency Domain
- K-Nearest Neighbour Classifier
Feature values are normalized:
$$Y = (x-\text{min})/(\text{max}-\text{min})$$
K=3 (based on 5-fold cross validation using GridSearchCV lib) is used for classification



Case Study: Cardiac Attack Prediction

Cardiac Attack Prediction Logic

```
HeartAttackAlarm      false  
f                  (value returned by KNN) + 1  
p                  (BPM/f)  
s                  (systolic measurement)/f  
d                  (diastolic measurement)/f
```

```
if (p>=170) and (s>=180) and (d>=120 )  
then:
```

```
    HeartAttackAlarm = true  
else
```

```
    Heart AttackAlarm = false
```

```
end if
```

```
return HeartAttackAlarm
```

Health Parameter	Alarm Value
BPM	>=170
Diastolic	>=120
Systolic	>=180

NPTEL

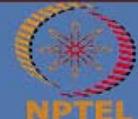
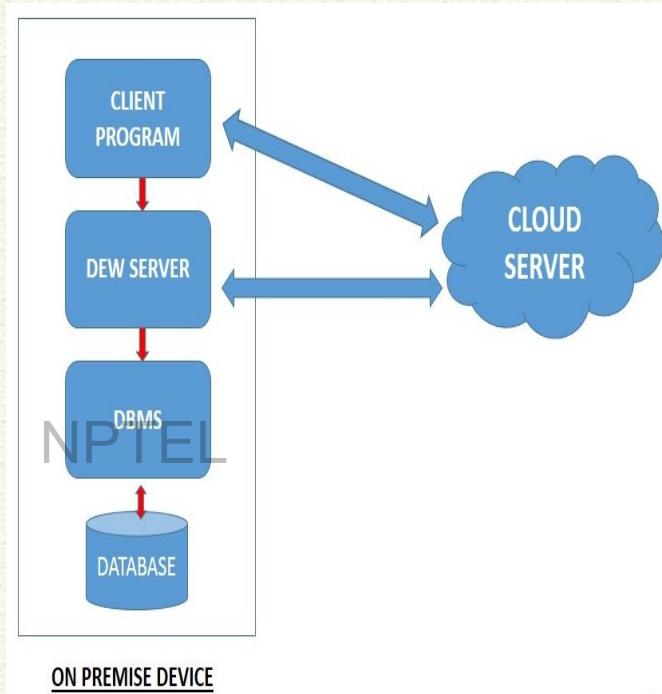
Note:

The proposed approach has no medical or clinical significance / implication and has been proposed strictly for demonstration purpose.

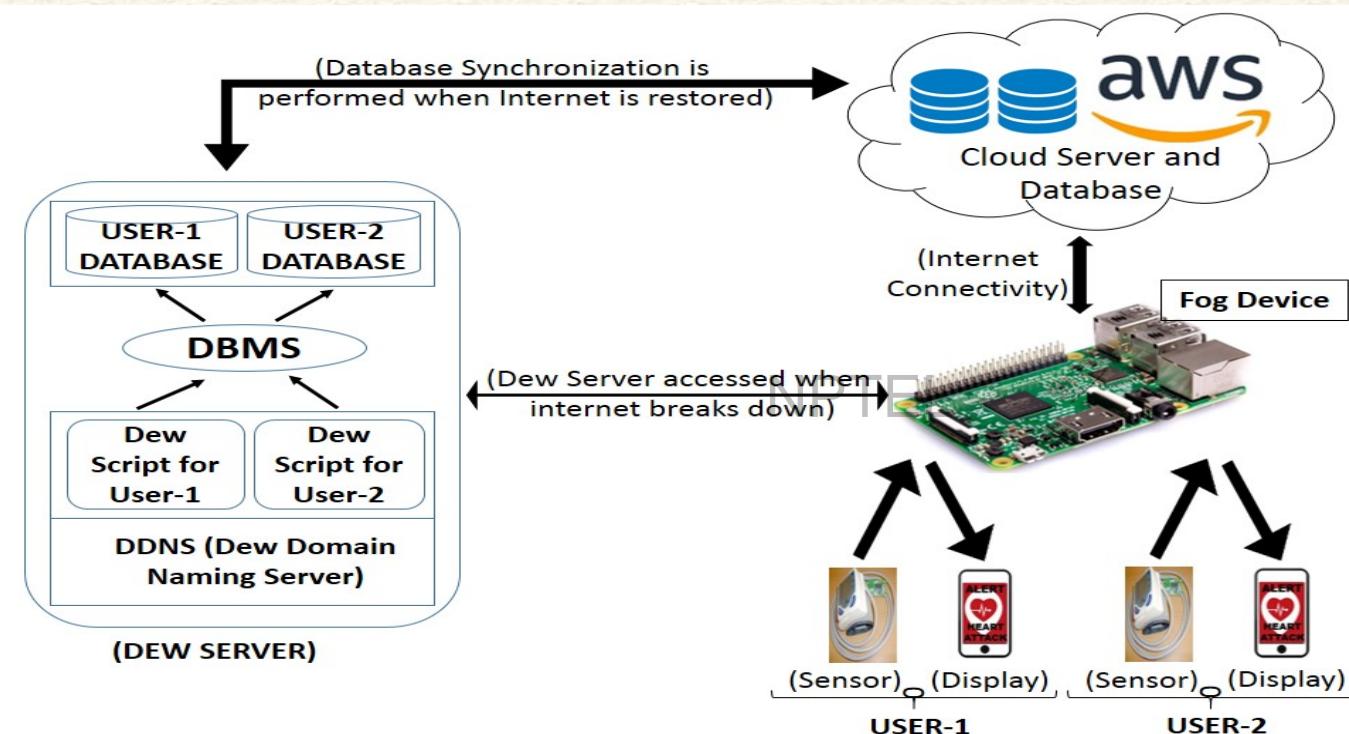


Dew Computing

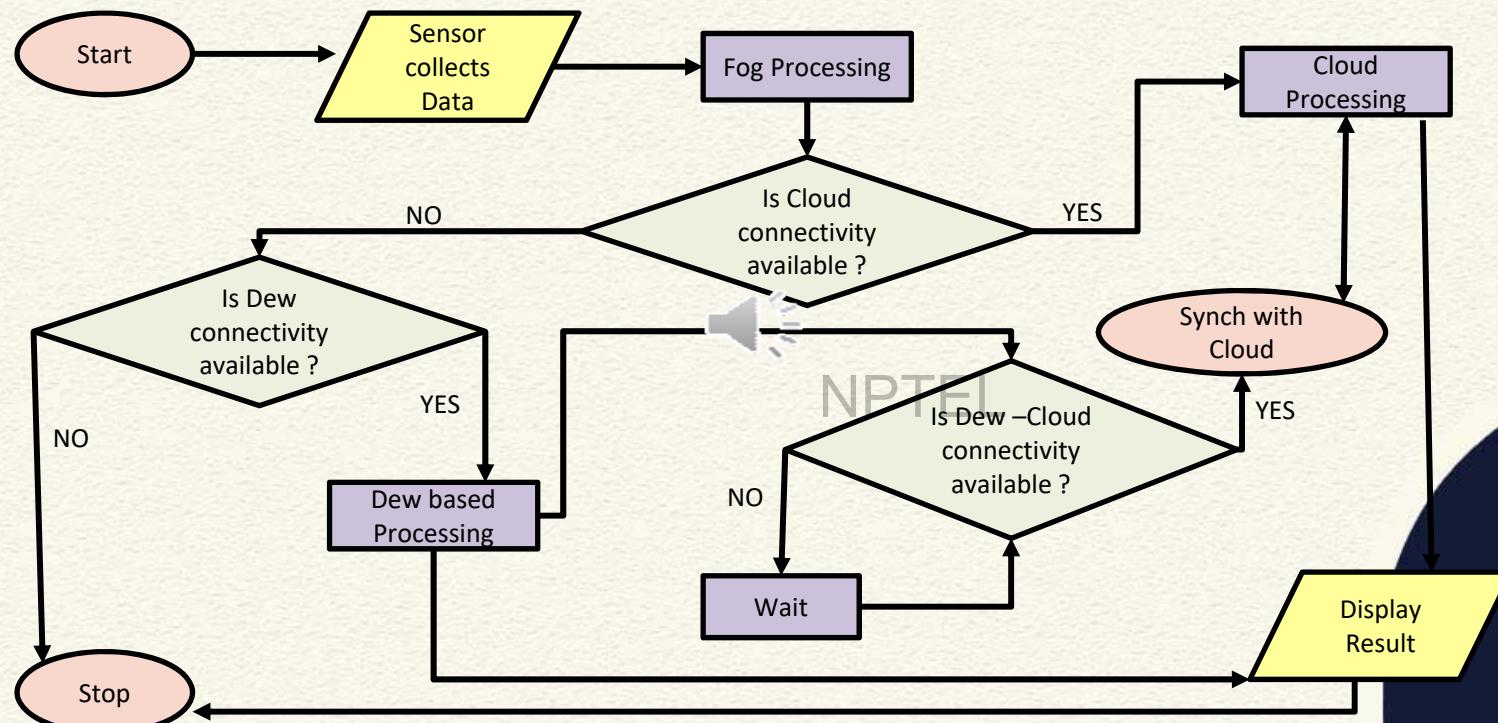
“Dew computing is an *on-premises* computer software-hardware organization paradigm in the cloud computing environment where the on-premises computer provides functionality that is **independent** of cloud services and is also **collaborative** with cloud services. The goal of dew computing is to fully realize the potentials of on-premises computers and cloud services”.



Dew based Cloud-Fog-Edge-IoHT Framework



Dew based Cloud-Fog-Edge-IoHT - Workflow



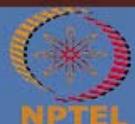
Comparative Study

FEATURE (from health service provider perspective)	CLOUD	Cloud-FOG-Edge	With DEW
On-premise resource utilization	Low	Sub-optimal	Optimal
Connectivity required	Internet	Local	Local
Uptime	Low	High	High
Bandwidth required	High	NPTEL Low	Low
Latency	High	Low	Low
Infrastructure requirement	Low	Moderate	Moderate
Processing power	High	Limited	Limited
Data Storage	High	Low	Moderate



REFERENCES

- Anish Poonia, MTech Dissertation, IIT Kharagpur, Fog Computing For Internet of Health Things, 2020
- Anish Poonia, Shreya Ghosh, Akash Ghosh, Shubha Brata Nath, Soumya K. Ghosh, Rajkumar Buyya, CONFRONT: Cloud-fog-dew based monitoring framework for COVID-19 management, Internet of Things, Elsevier, Volume 16, 2021
- Cisco White Paper. 2015. Fog Computing and the Internet of Things: Extend the Cloud to Where the Things Are.
- Gupta H, Vahid Dastjerdi A, Ghosh SK, Buyya R. iFogSim: A toolkit for modeling and simulation of resource management techniques in the Internet of Things, Edge and Fog computing environments. *Softw Pract Exper.* 2017;47:1275-296. <https://doi.org/10.1002/spe.2509>
- Luiz Bittencourt et al., The Internet of Things, Fog and Cloud continuum: Integration and challenges, Internet of Things, Volumes 3–4, 2018, Pages 134-155, ISSN 2542-6605, <https://doi.org/10.1016/j.iot.2018.09.005>



*Thank
you*



NPTEL

