

**AJAX**

AJAX is an acronym for **Asynchronous JavaScript and XML**. AJAX is a new technique for creating better, faster and interactive web applications with the help of JavaScript, DOM, XML, HTML, CSS etc. AJAX allows you to send and receive data asynchronously without reloading the entire web page. So it is fast.

AJAX allows you to send only important information to the server not the entire page. So only valuable data from the client side is routed to the server side. It makes your application interactive and faster.

Ajax is the most viable Rich Internet Application(RIA) technique so far.

**Where it is used?**

There are too many web applications running on the web that are using AJAX Technology.

Some

are

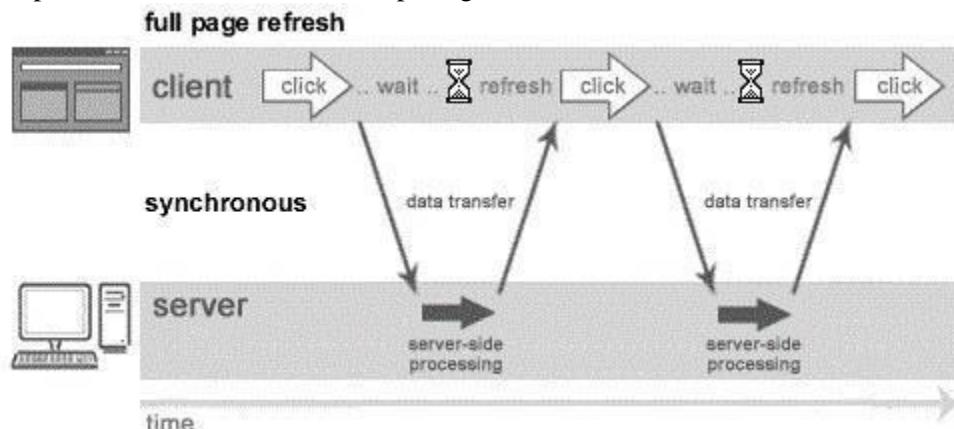
- : 1. Gmail
- 2. Face book
- 3. Twitter
- 4. Google maps
- 5. YouTube etc.,

**Synchronous Vs. Asynchronous Application**

Before understanding AJAX, let's understand classic web application model and AJAX Web application model.

**❖ Synchronous (Classic Web-Application Model)**

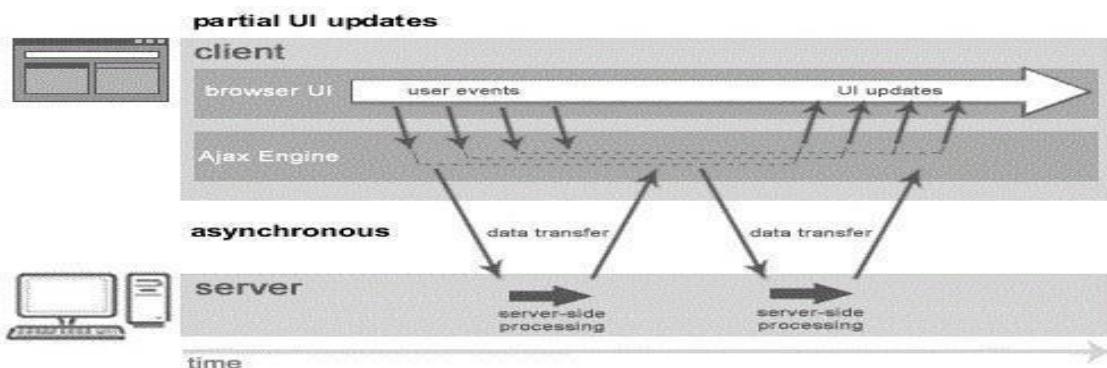
A synchronous request blocks the client until operation completes i.e. browser is not unresponsive. In such case, JavaScript Engine of the browser is blocked.



As you can see in the above image, full page is refreshed at request time and user is blocked until request completes.

**❖ Asynchronous (AJAX Web-Application Model)**

An asynchronous request doesn't block the client i.e. browser is responsive. At that time, user can perform other operations also. In such case, JavaScript Engine of the browser is not blocked.



As you can see in the above image, full page is not refreshed at request time and user gets response from the AJAX Engine. Let's try to understand asynchronous communication by the image given below.

### AJAX Components

AJAX is not a technology but group of inter-related technologies. AJAX Technologies includes:

- ❖ HTML/XHTML and CSS
- ❖ DOM
- ❖ XML or JSON(JavaScript Object Notation)
- ❖ XMLHttpRequest Object
- ❖ JavaScript

- **HTML/XHTML and CSS**

These technologies are used for displaying content and style. It is mainly used for presentation.

- **DOM**

It is used for dynamic display and interaction with data.

- **XML or JSON(Javascript Object Notation)**

For carrying data to and from server. JSON is like XML but short and faster than XML.

- **XMLHttpRequest Object**

For asynchronous communication between client and server.

- **JavaScript**

It is used to bring above technologies together. Independently, it is used mainly for client-side validation.

### Understanding XMLHttpRequest

It is the heart of AJAX technique. An object of XMLHttpRequest is used for asynchronous communication between client and server. It provides a set of useful methods and properties that are used to send HTTP Request to and retrieve data from the web server. It performs following operations:

1. Sends data from the client in the background
2. Receives the data from the server
3. Updates the webpage without reloading it.

- **Methods of XMLHttpRequest object**

Method	Description
void open(method, URL)	Opens the request specifying get or post method and url.
void open(method, URL, async)	Same as above but specifies asynchronous or not.
void open(method, URL, async, username, password)	Same as above but specifies username and password.
void send()	Sends GET request.
void send(string)	Sends POST request.
setRequestHeader(header,value)	It adds request headers.

**Syntax of open() method:**

```
xmlHttp.open("GET","conn.php",true)
```

e); which takes three attributes

1. An HTTP method such as GET ,POST , or HEAD
2. The URL of the Server resource
3. A boolean Flag that indicates whether the request should be asynchronously(true) or synchronously(false)

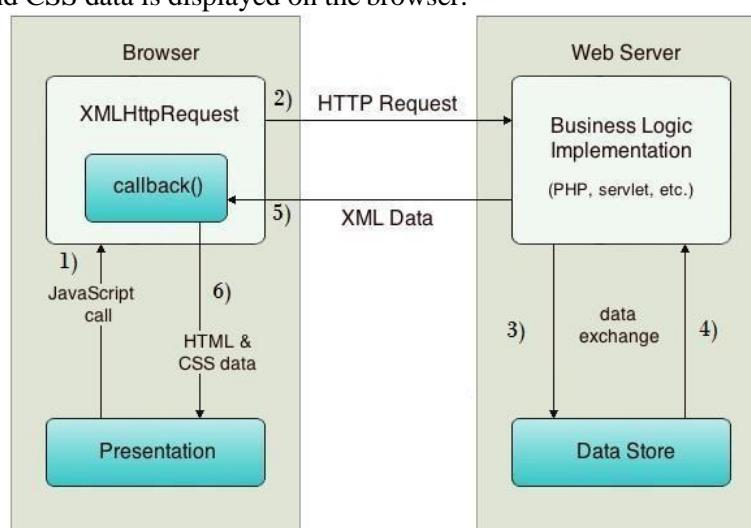
**Properties of XMLHttpRequest Object:**

Property	Description
readyState	Represents the state of the request. It ranges from 0 to 4.  <b>0 UN INITIALIZED</b> – After creating XMLHttpRequest Object before calling <i>open()</i> method. <b>1 CONNECTION ESTABLISHED</b> – <i>open()</i> is called but <i>send()</i> is not called. <b>2 REQUEST SENT</b> - <i>send()</i> is called. <b>3 PROCESSING</b> - Downloading data; <i>responseText</i> holds the data. <b>4 DONE</b> - The operation is completed successfully.
onReadyStateChange	It is called whenever <i>readystate</i> attribute changes. It must not be used with synchronous requests.
reponseText	Returns response as TEXT.
responseXML	Returns response as XML

**How AJAX Works?**

AJAX communicates with the server using XMLHttpRequest object. Let's understand the flow of AJAX with the following figure:

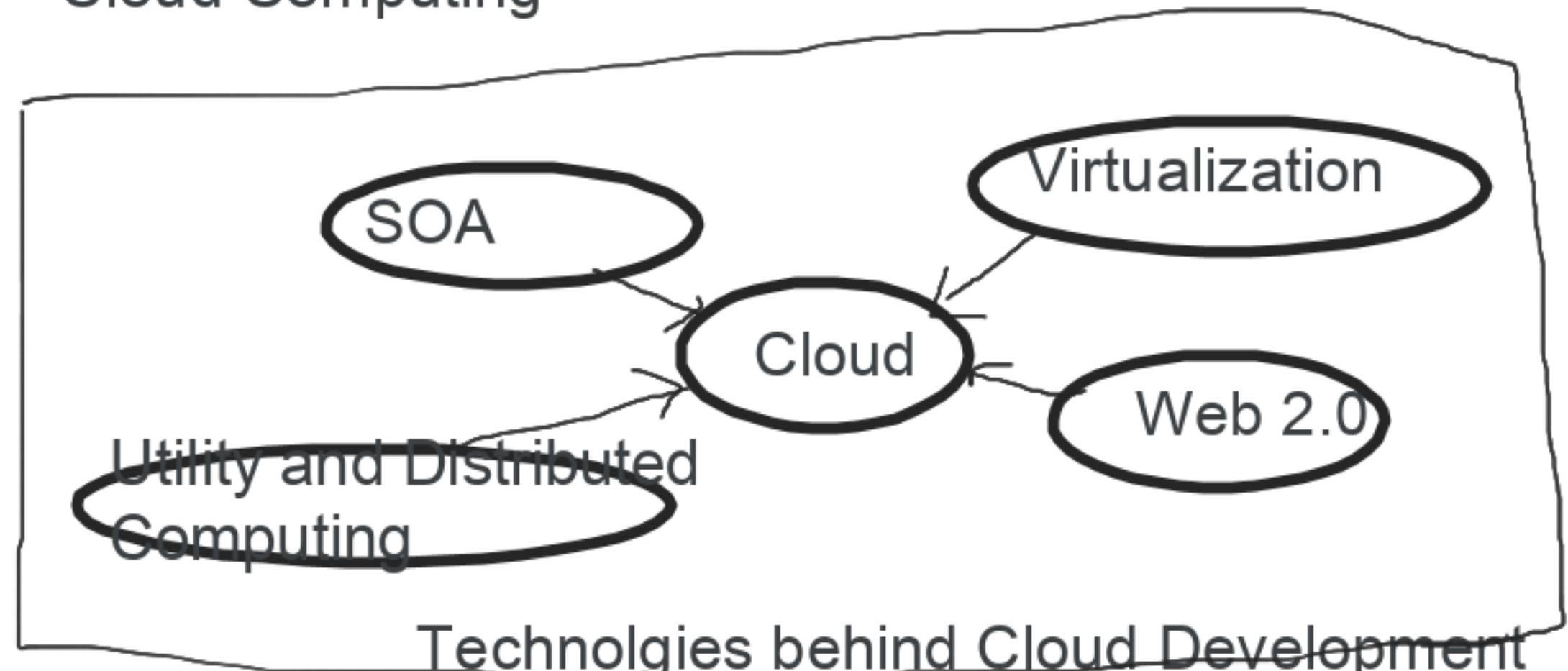
1. User sends a request from the UI and a javascript call goes to XMLHttpRequest object.
2. HTTP Request is sent to the server by XMLHttpRequest object.
3. Server interacts with the database using JSP, PHP, Servlet, ASP.net etc.
4. Data is retrieved.
5. Server sends XML data or JSON data to the XMLHttpRequest callback function.
6. HTML and CSS data is displayed on the browser.



Trends of computing Models

- \* Distributed computing
- \* Grid computing
- \* Cluster computing
- \* Utility computing
- \* Cloud computing

## Cloud Computing



## Cloud Service Model

SaaS: Google docs,CRM

PaaS: google app engine,Azure

IaaS: Amazon EC2(elastic  
comput cloud) ,AWS S3(simple  
storage service).

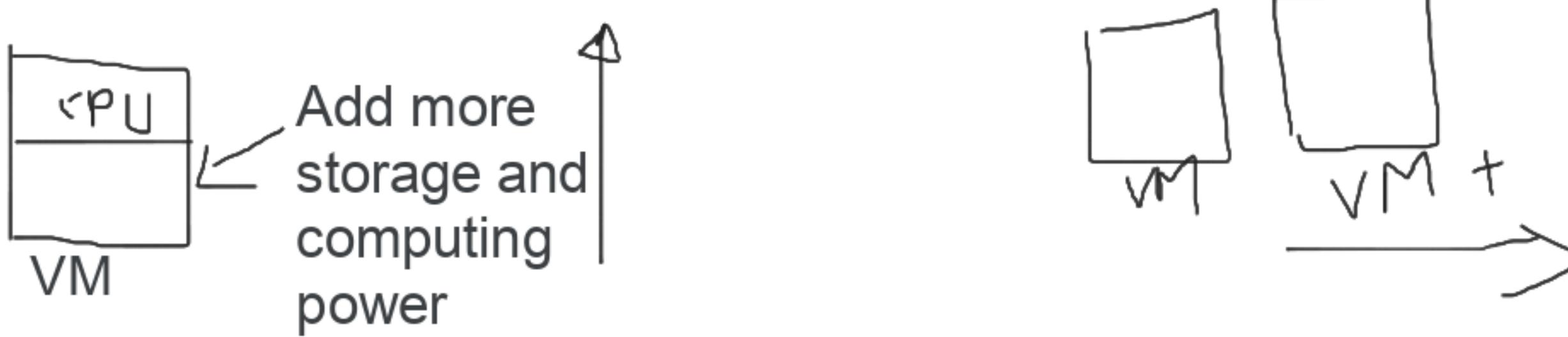
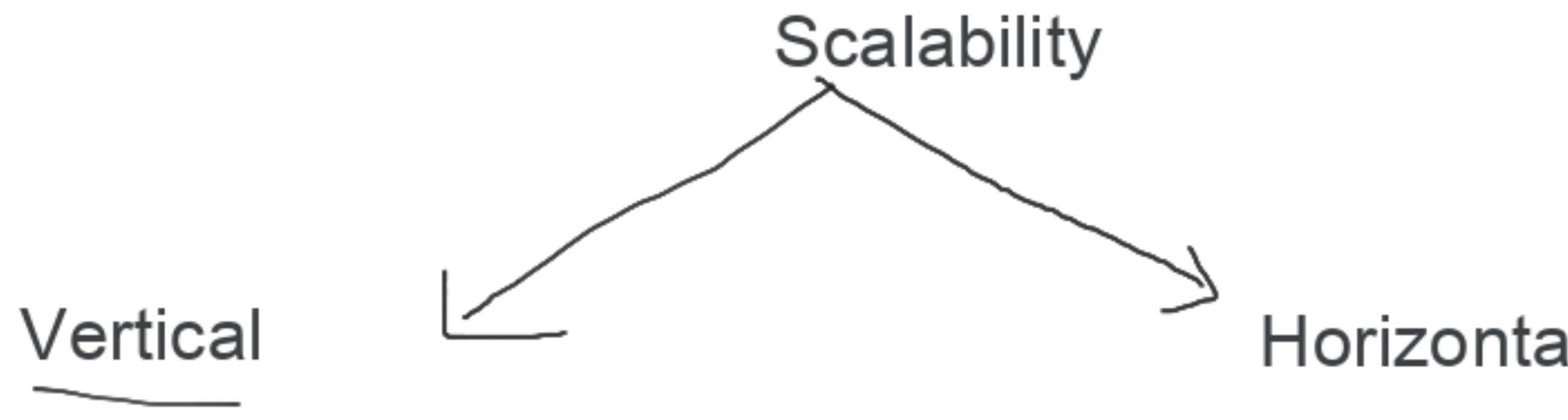
SaaS:Storage as a Service.

BaaS

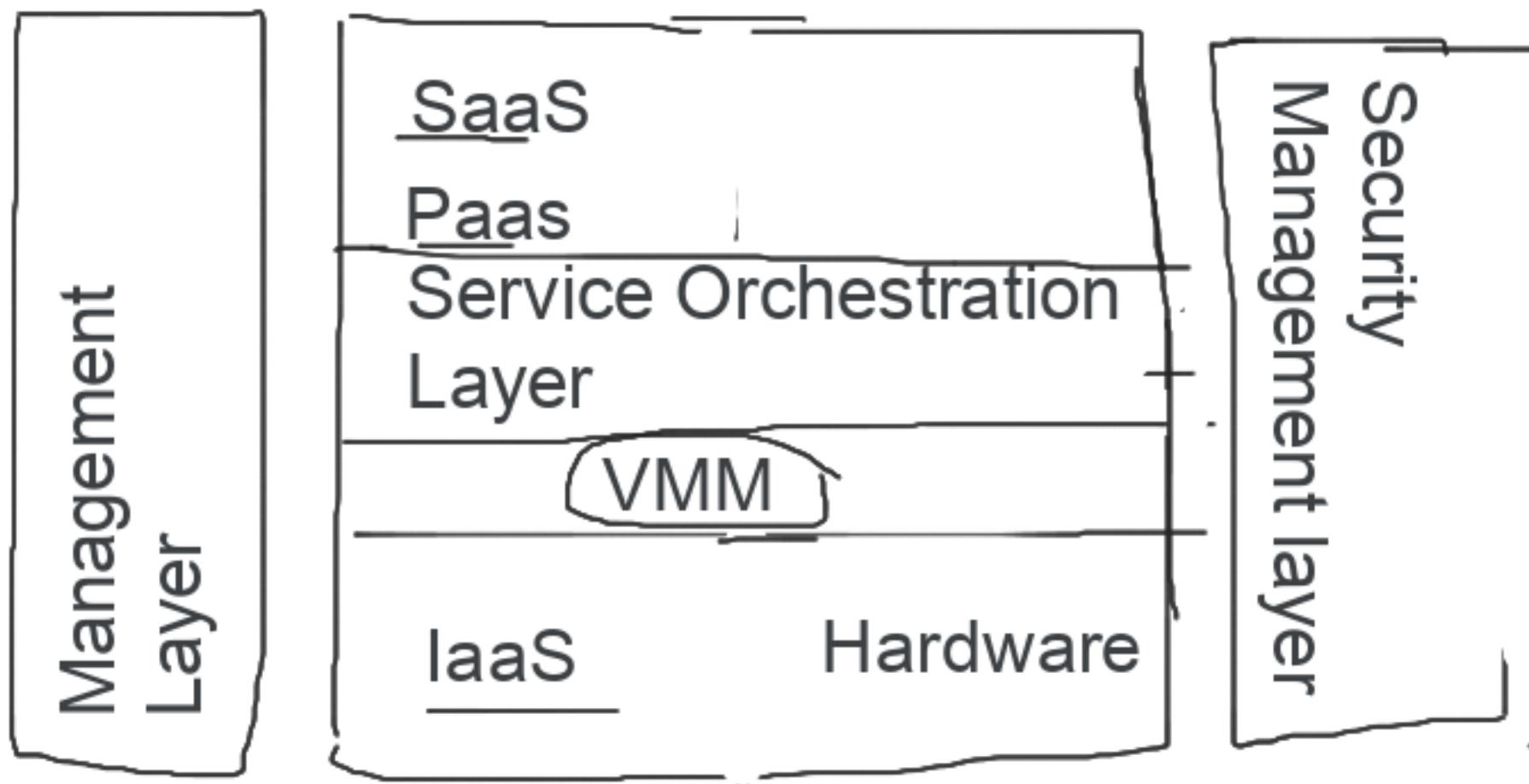
XaaS:Everything as a service

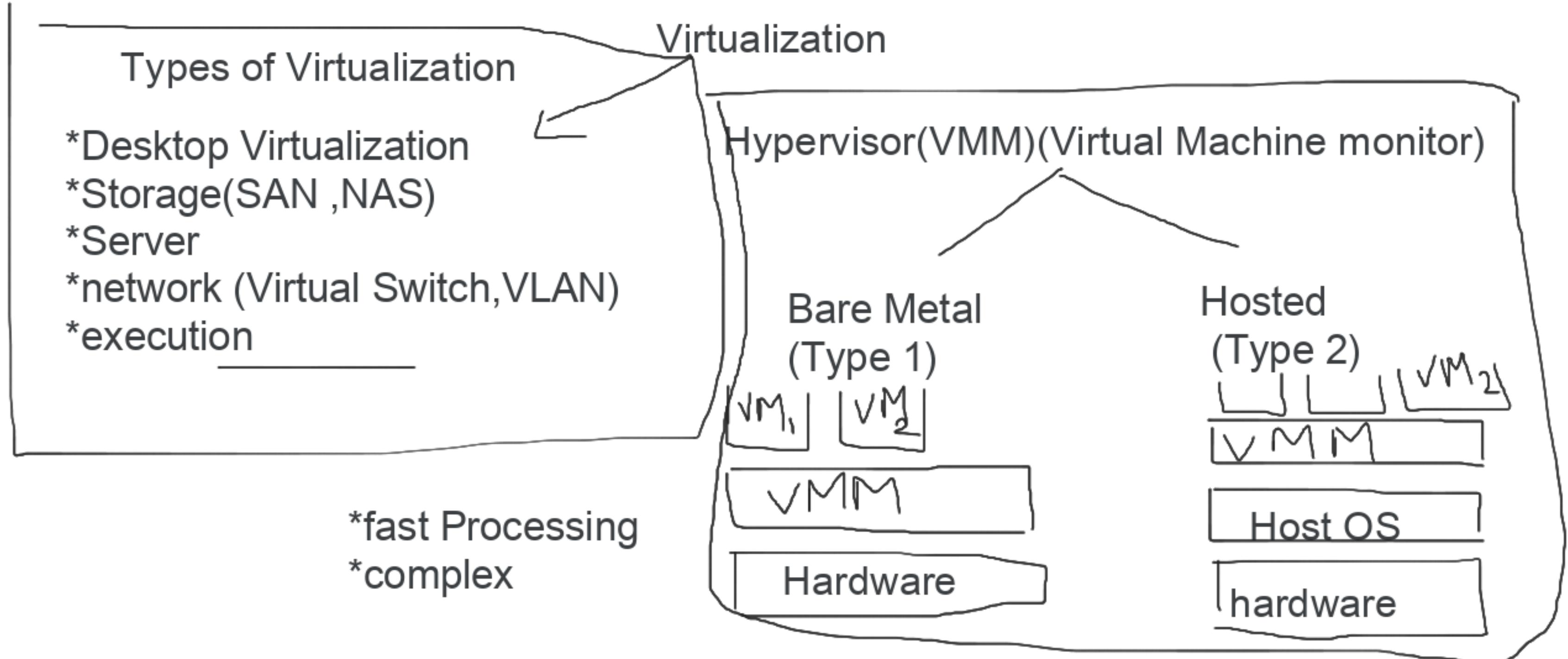
## Deployment model

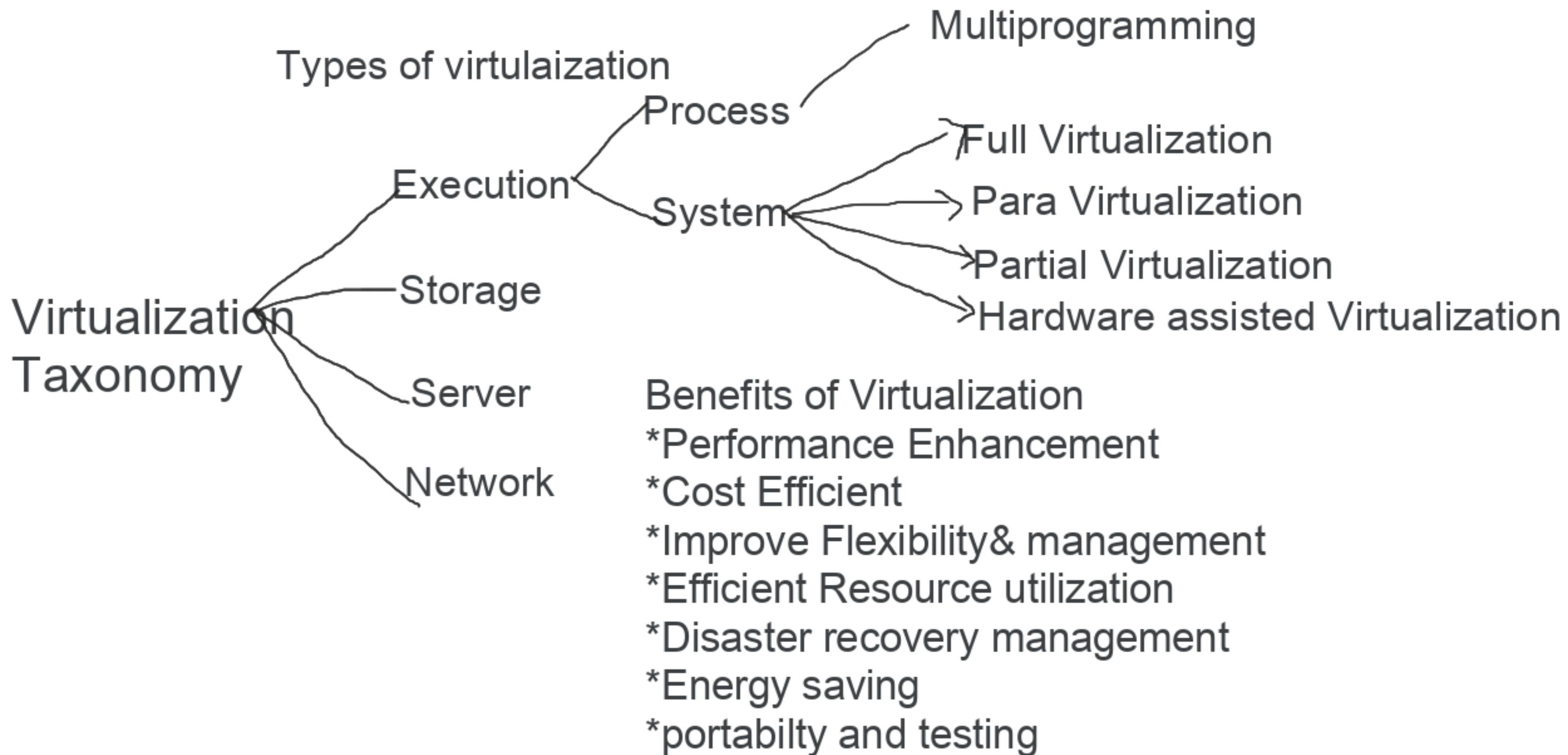
- \*Public cloud
- \*Private cloud.
- \*Hybrid Cloud
- \*Community Cloud .



## Cloud reference Model







## Full virtualization

- \*Guest Operating System get Modified
  - \*Run on raw hardware
  - \*privledge instruction(system call) easily interpreted.
  - \*Secure and complex
  - \*Cost efficient
- VMWARE ESXi

## Para Virtualization

- \*No modification of OS
  - \*explicit system call arec executed
  - \*simple
  - \* helpful in performance critical application
- XEN

# Virtual Machine Migration







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  
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  
NPTEL ONLINE  
CERTIFICATION COURSES

# Distributed Computing

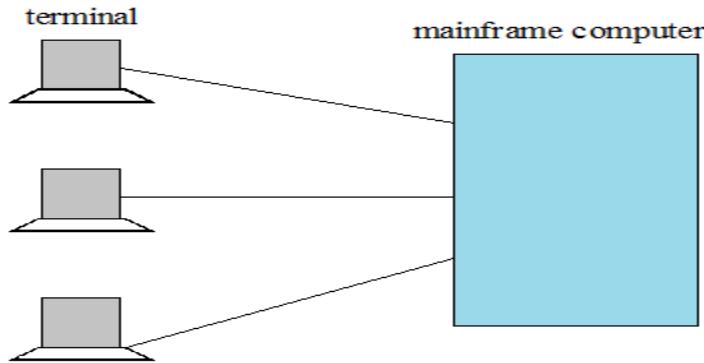


IIT KHARAGPUR



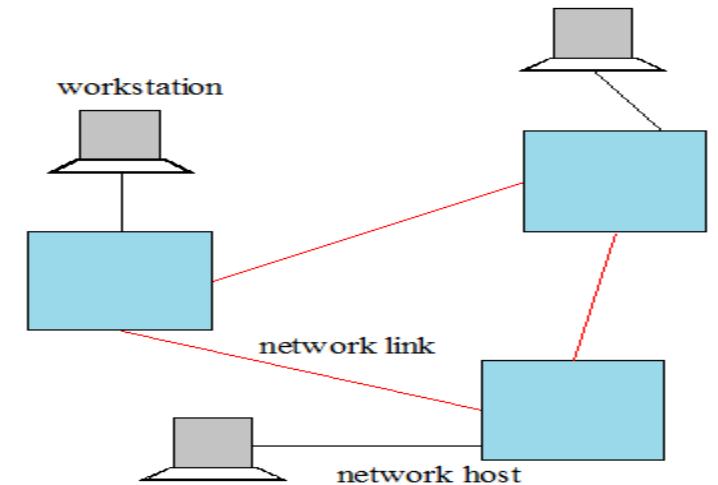
NPTEL  
NPTEL ONLINE  
CERTIFICATION COURSES

# Centralized vs. Distributed Computing



Centralized Computing

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



Distributed Computing



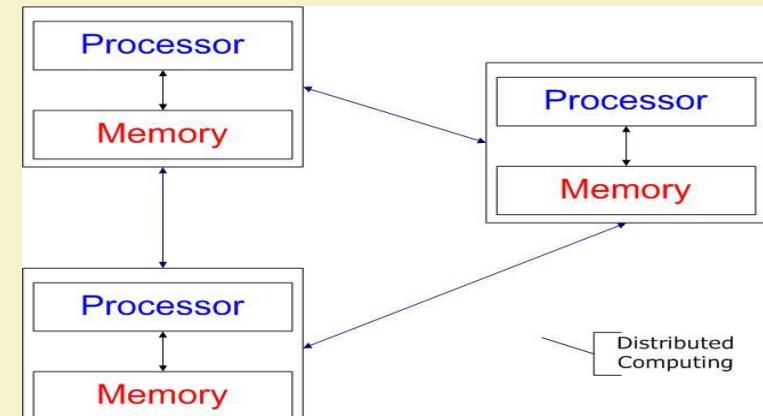
IIT KHARAGPUR



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!!



IIT KHARAGPUR



NPTEL  
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

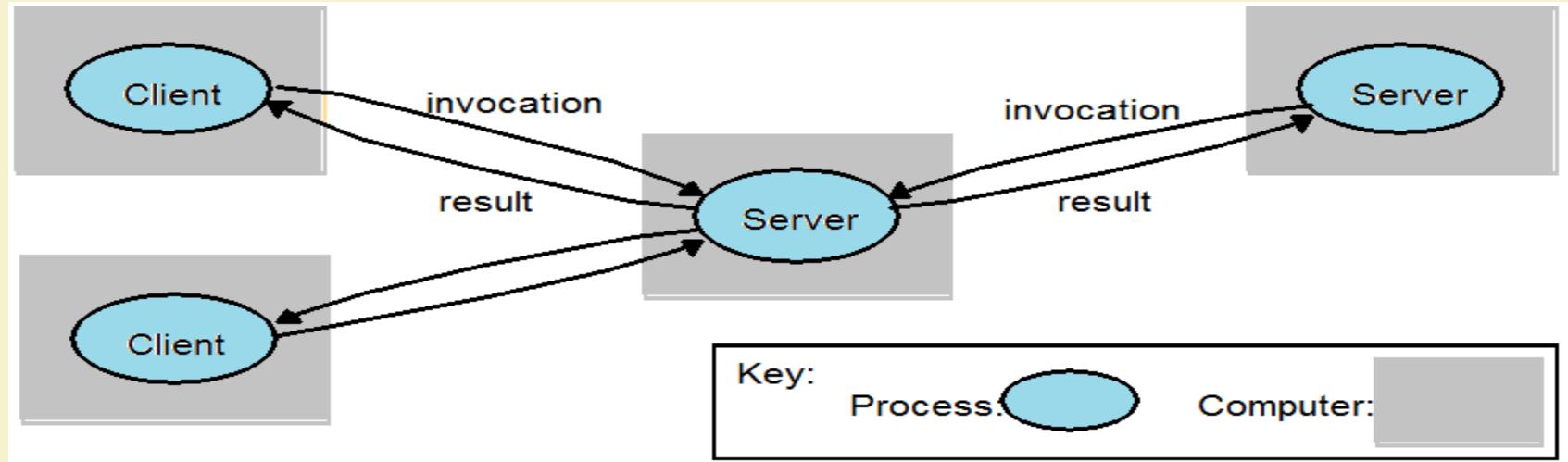


IIT KHARAGPUR



NPTEL ONLINE  
CERTIFICATION COURSES

# Clients invoke individual servers

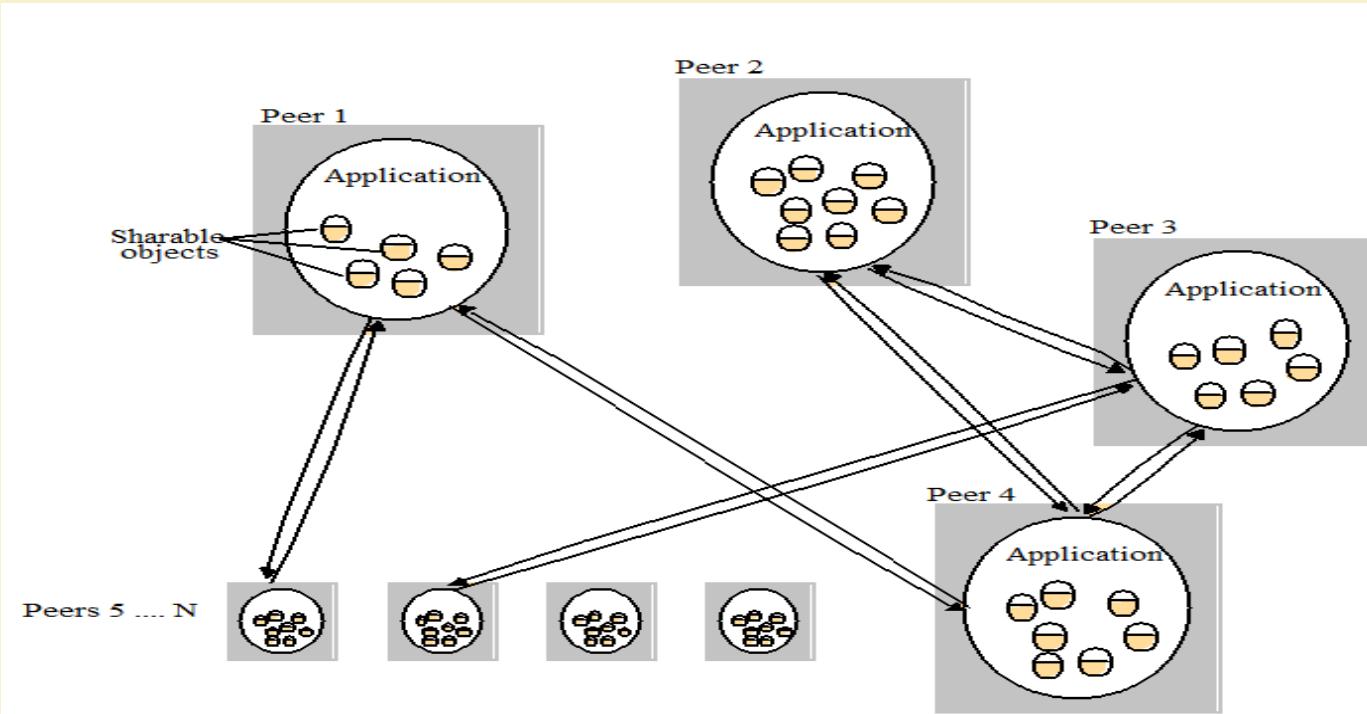


IIT KHARAGPUR



NPTEL  
NPTEL ONLINE  
CERTIFICATION COURSES

# A typical distributed application based on peer processes



IIT KHARAGPUR



NPTEL  
NPTEL ONLINE  
CERTIFICATION COURSES

# Grid Computing



IIT KHARAGPUR



NPTEL ONLINE  
CERTIFICATION COURSES

# Grid Computing?

- [Pcwebopedia.com](http://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

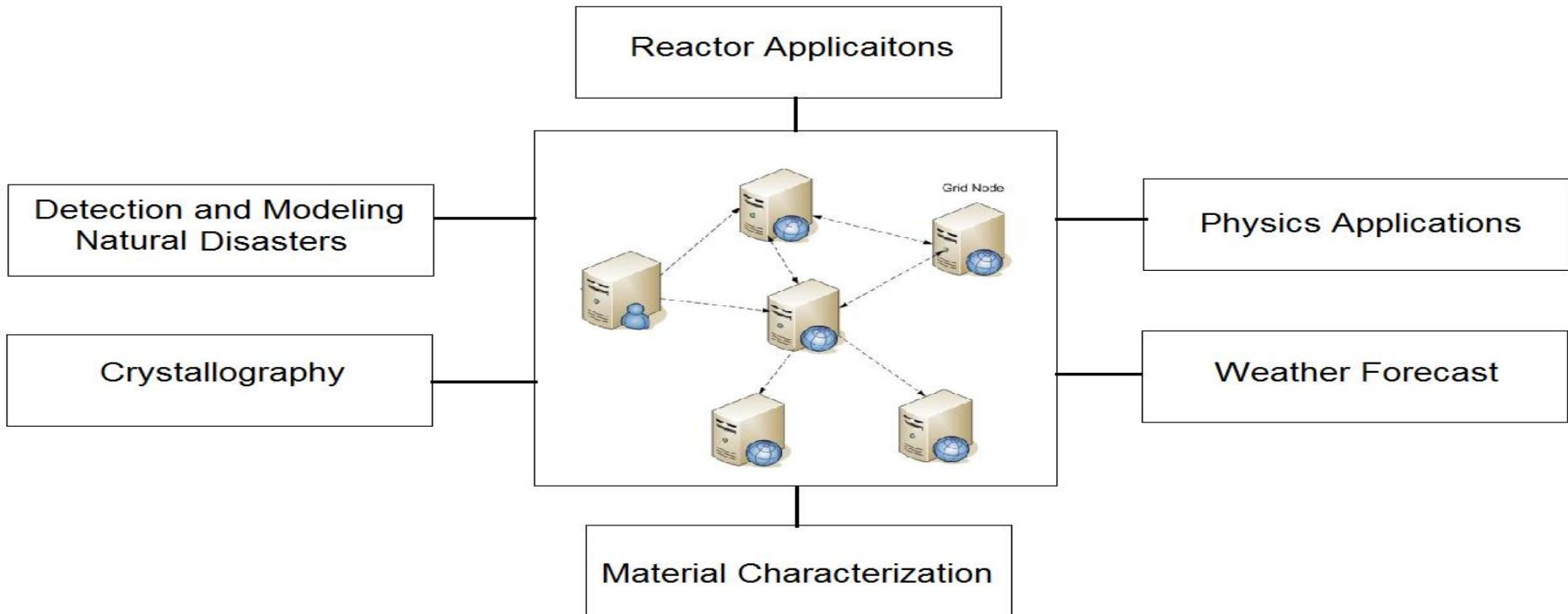
## When v use

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.

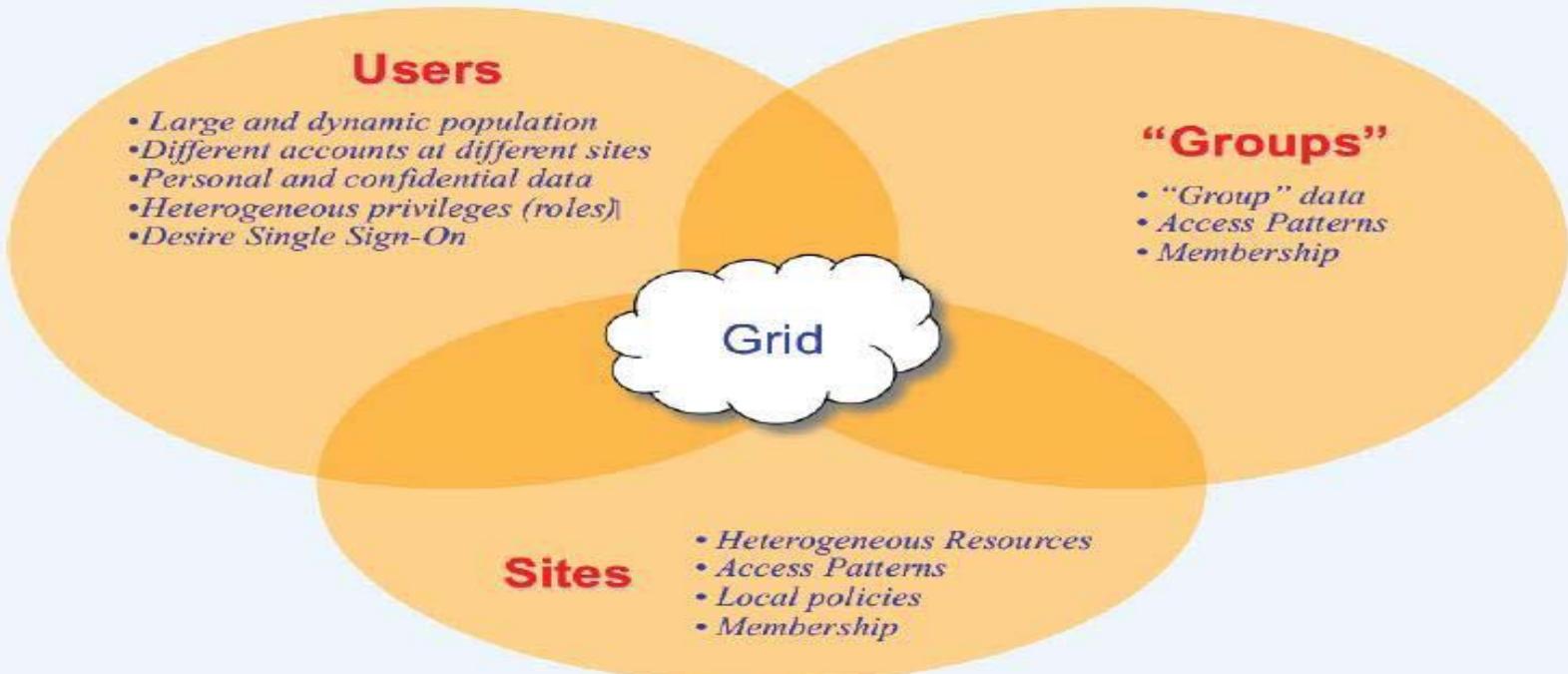


IIT KHARAGPUR



NPTEL ONLINE  
CERTIFICATION COURSES

# 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



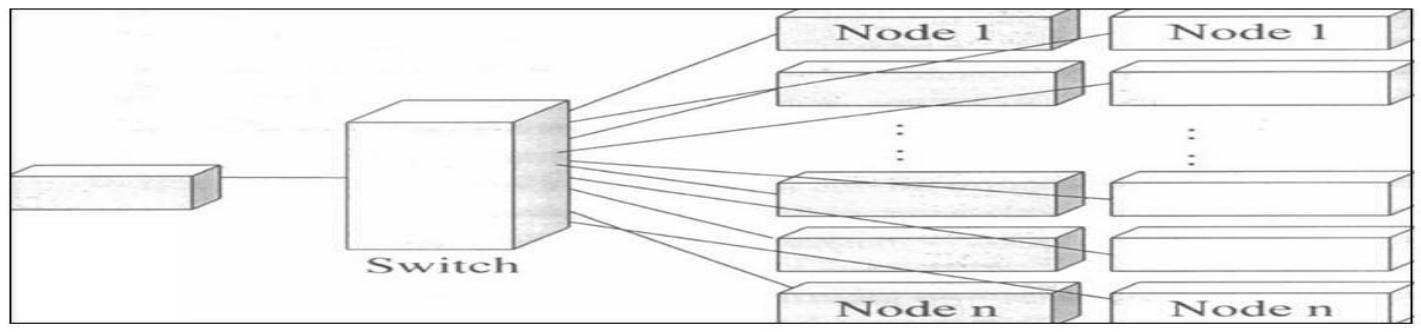
IIT KHARAGPUR



NPTEL 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)

# Utility Computing



IIT KHARAGPUR



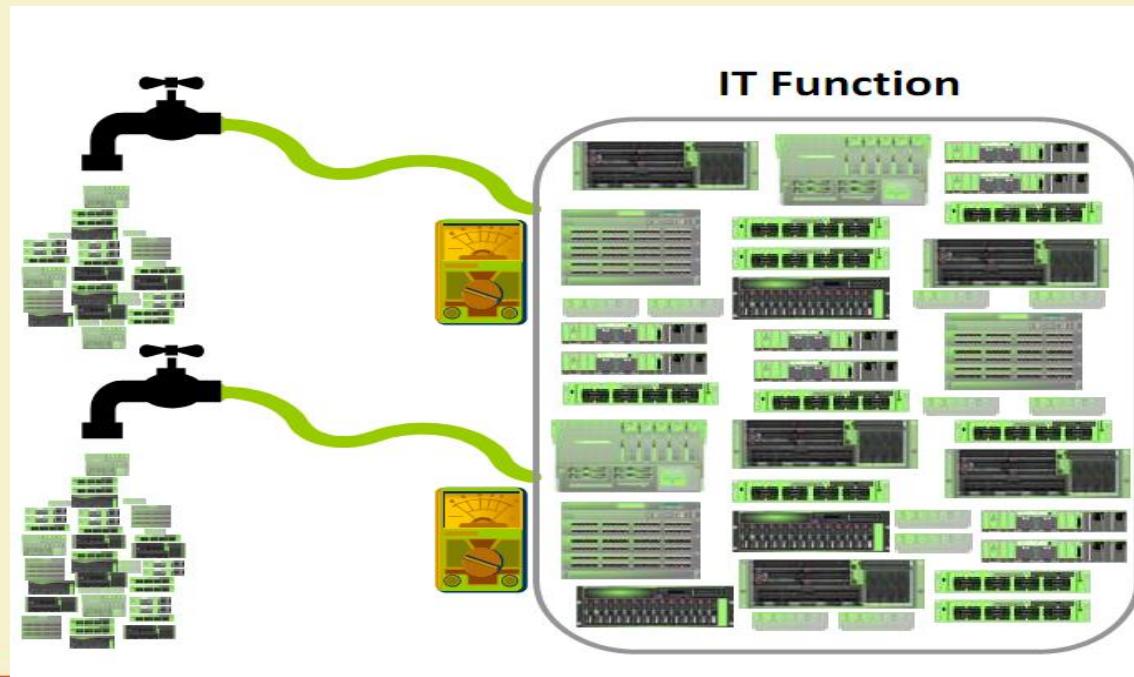
NPTEL  
NPTEL ONLINE  
CERTIFICATION COURSES

# “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



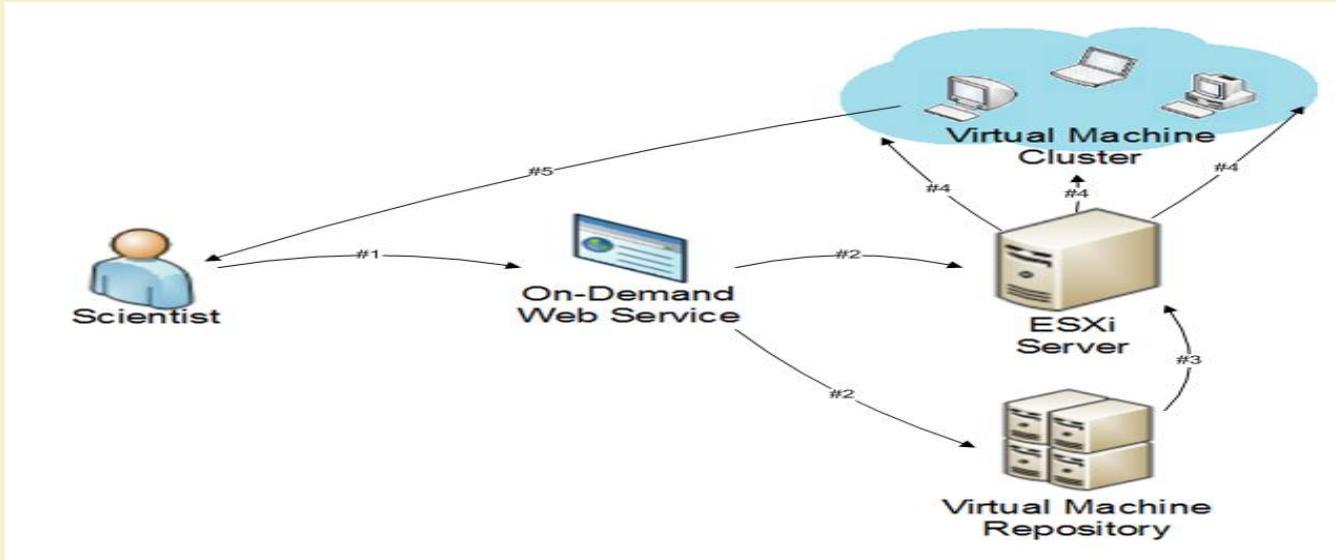
IIT KHARAGPUR



NPTEL  
NPTEL ONLINE  
CERTIFICATION COURSES

# Utility Computing Example

## On-Demand Cyber Infrastructure



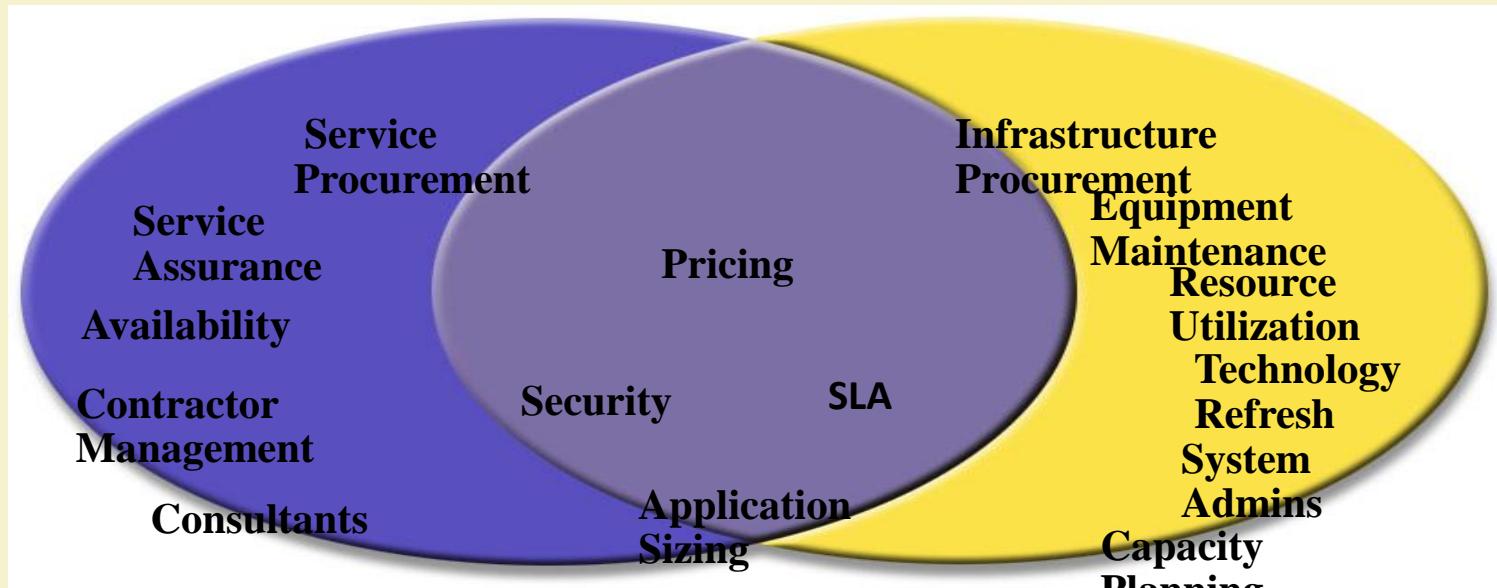
IIT KHARAGPUR



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 ONLINE  
CERTIFICATION COURSES

# Cloud Computing



IIT KHARAGPUR

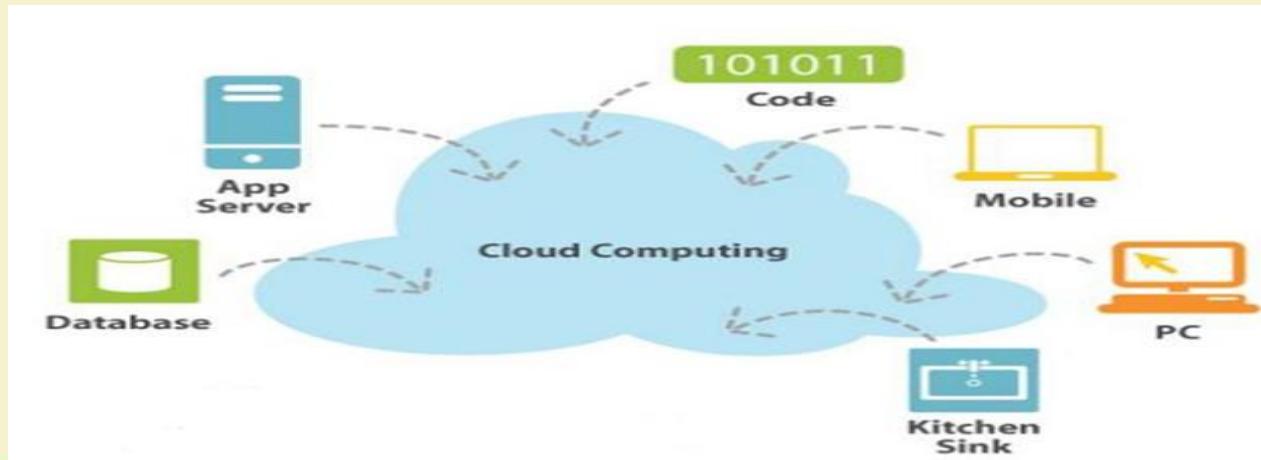


NPTEL  
NPTEL ONLINE  
CERTIFICATION COURSES

# 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!!



IIT KHARAGPUR



NPTEL ONLINE  
CERTIFICATION COURSES



IIT KHARAGPUR



NPTEL ONLINE  
CERTIFICATION COURSES

# Cloud Computing - Overview

Prof. Soumya K Ghosh

Department of Computer Science and Engineering

IIT KHARAGPUR

# Cloud Computing



IIT KHARAGPUR

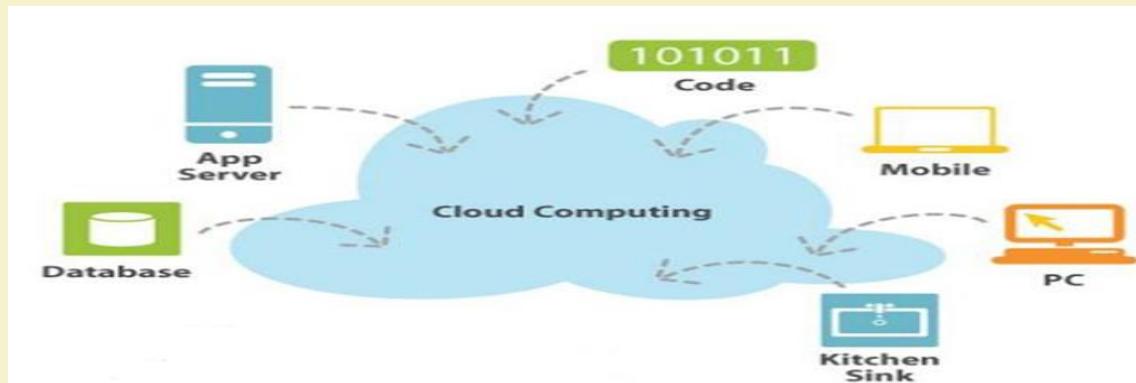


NPTEL  
NPTEL ONLINE  
CERTIFICATION COURSES

# 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)**

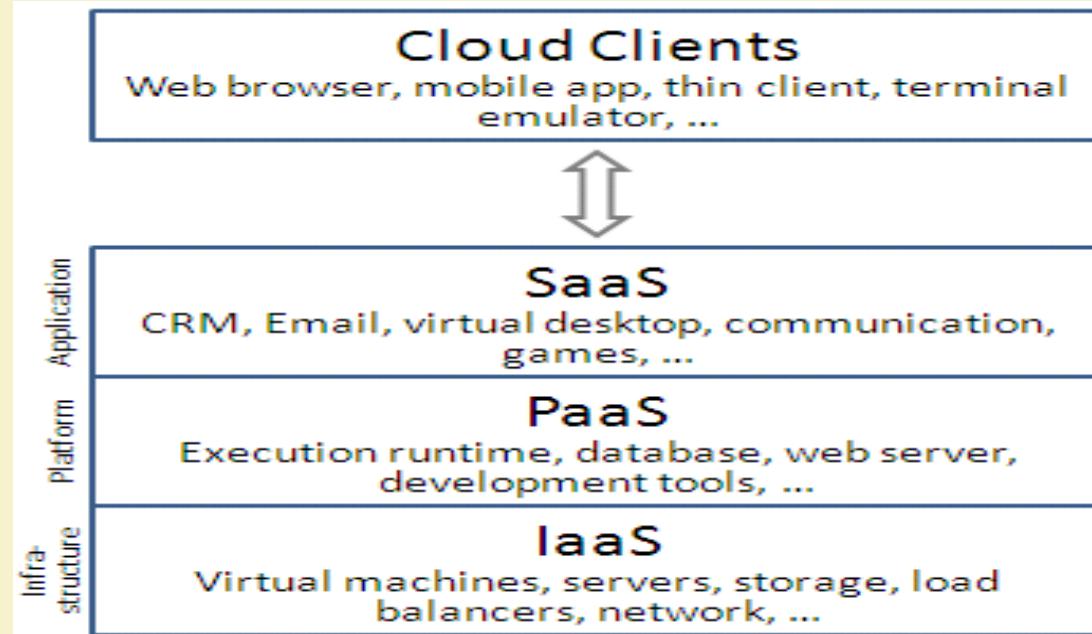
- The capability provided to provision processing, storage, networks, and other fundamental computing resources
- Consumer can deploy and run arbitrary software
- e.g: *Amazon Web Services and Flexi scale.*

# 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.



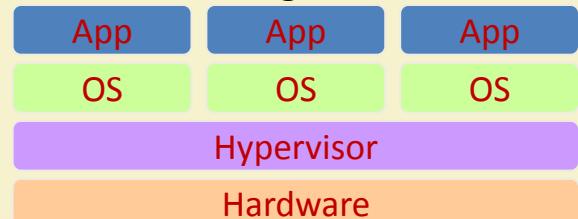
IIT KHARAGPUR



NPTEL  
NPTEL ONLINE  
CERTIFICATION COURSES

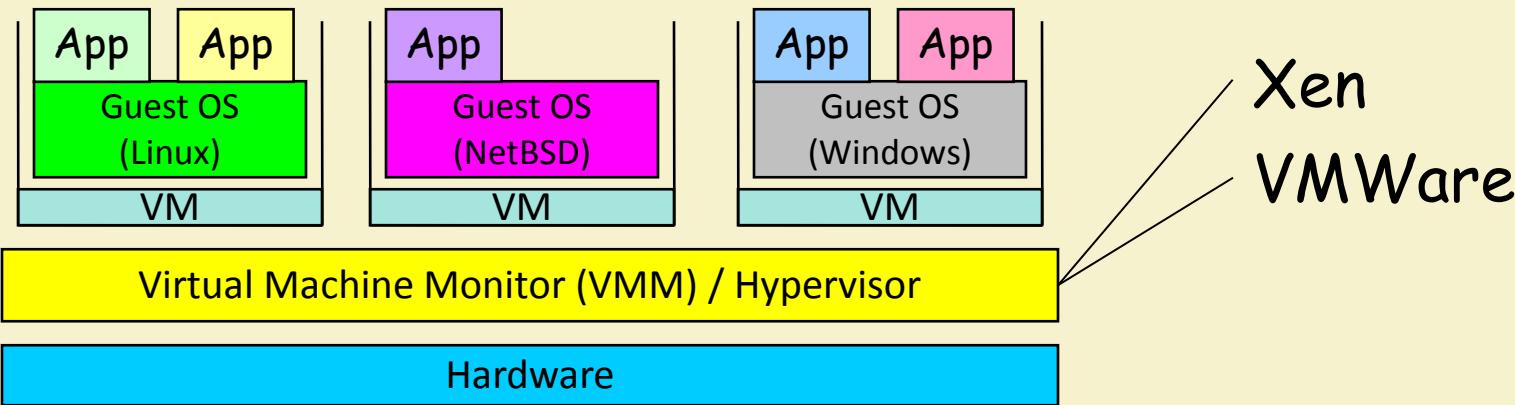
# Cloud and Virtualization

- **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

# 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  
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 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  
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 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.



IIT KHARAGPUR

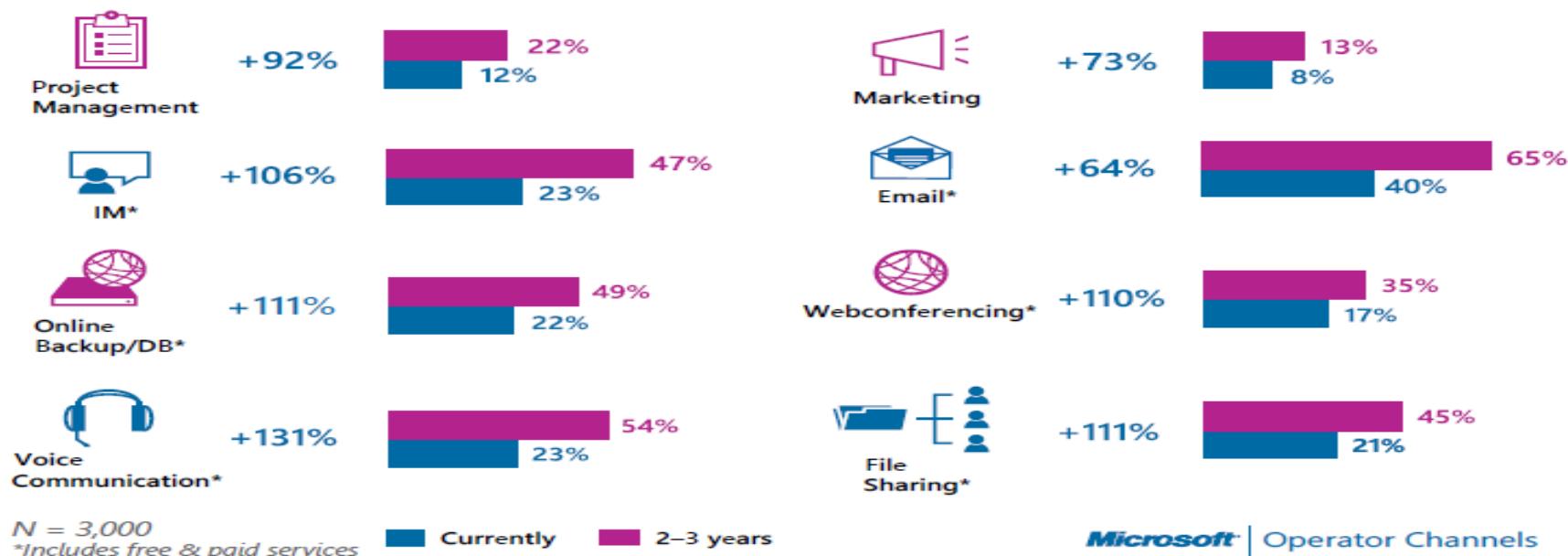


NPTEL  
NPTEL ONLINE  
CERTIFICATION COURSES

# 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>



IIT KHARAGPUR



NPTEL ONLINE  
CERTIFICATION COURSES

# 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



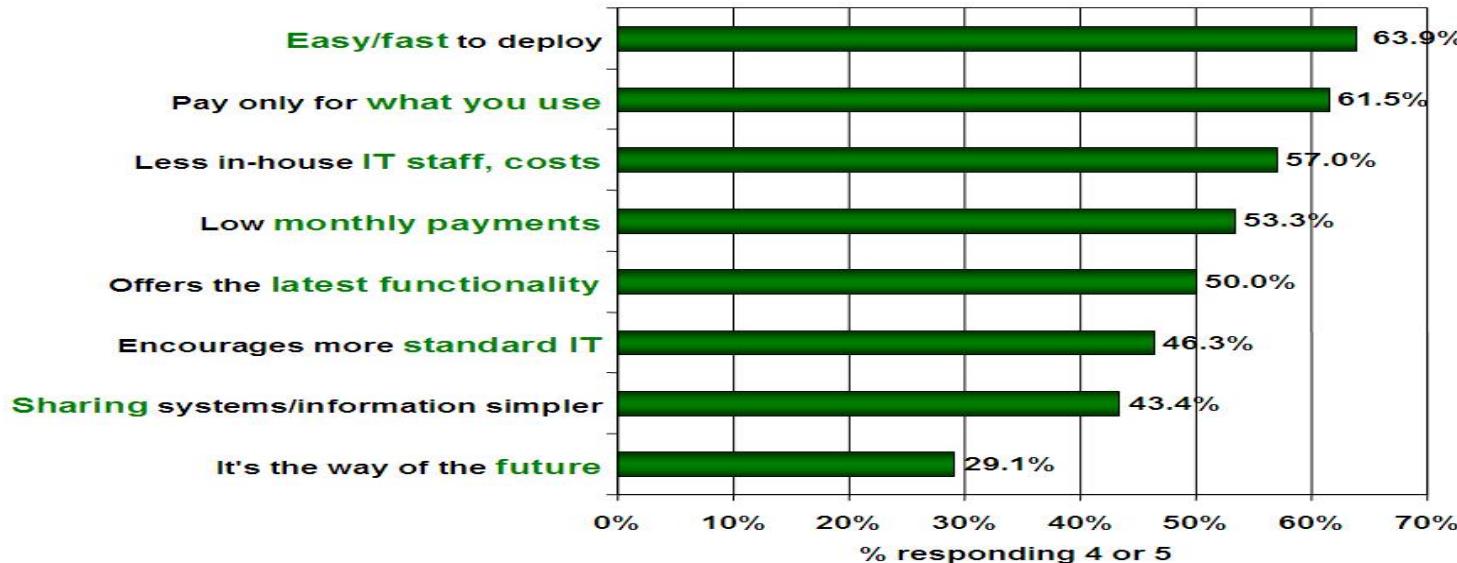
IIT KHARAGPUR



NPTEL 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)

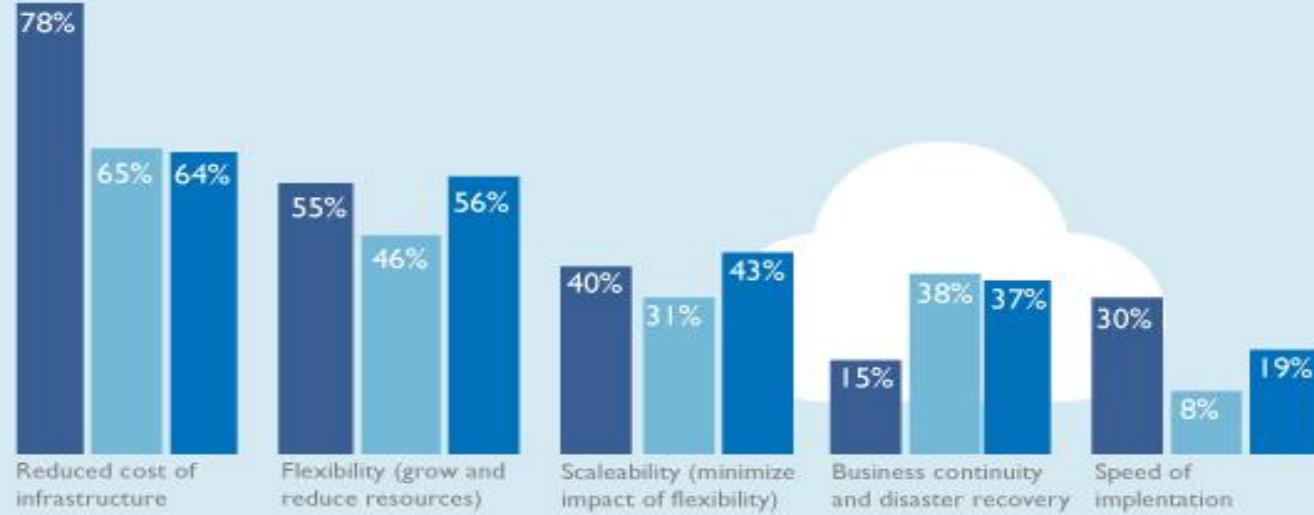


Source: IDC Enterprise Panel, August 2008 n=244

# Factors driving investment in cloud

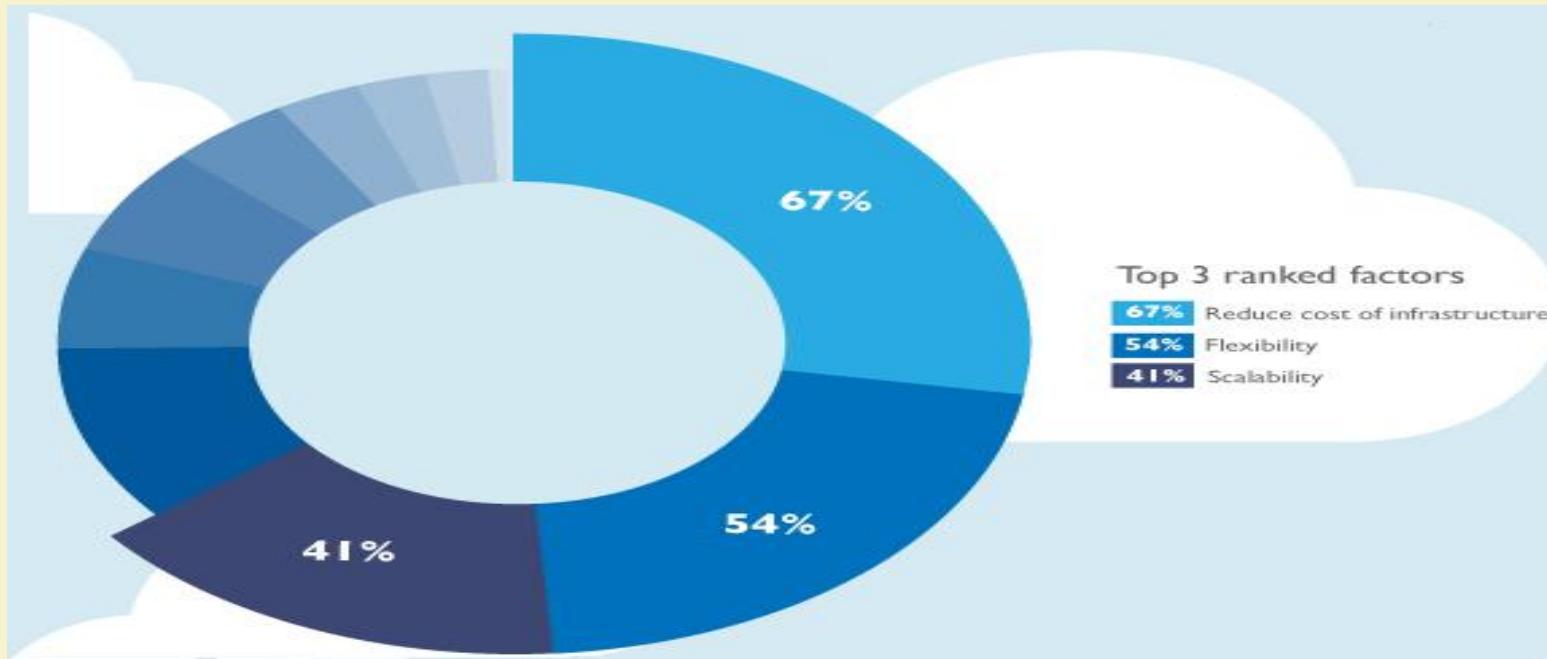
Factors driving investments in cloud per business size

- 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!



IIT KHARAGPUR



NPTEL  
NPTEL ONLINE  
CERTIFICATION COURSES



IIT KHARAGPUR



NPTEL ONLINE  
CERTIFICATION COURSES

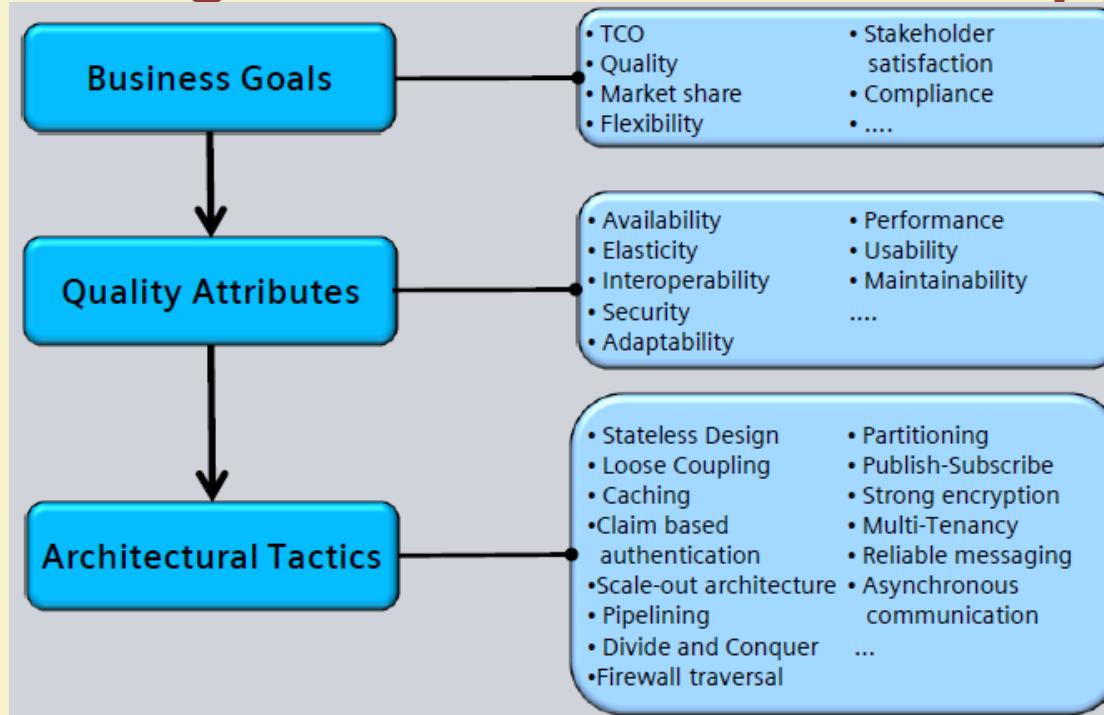
# CLOUD COMPUTING

## CLOUD COMPUTING ARCHITECTURE

PROF. SOUMYA K. GHOSH

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING  
IIT KHARAGPUR

# Context: High Level Architectural Approach



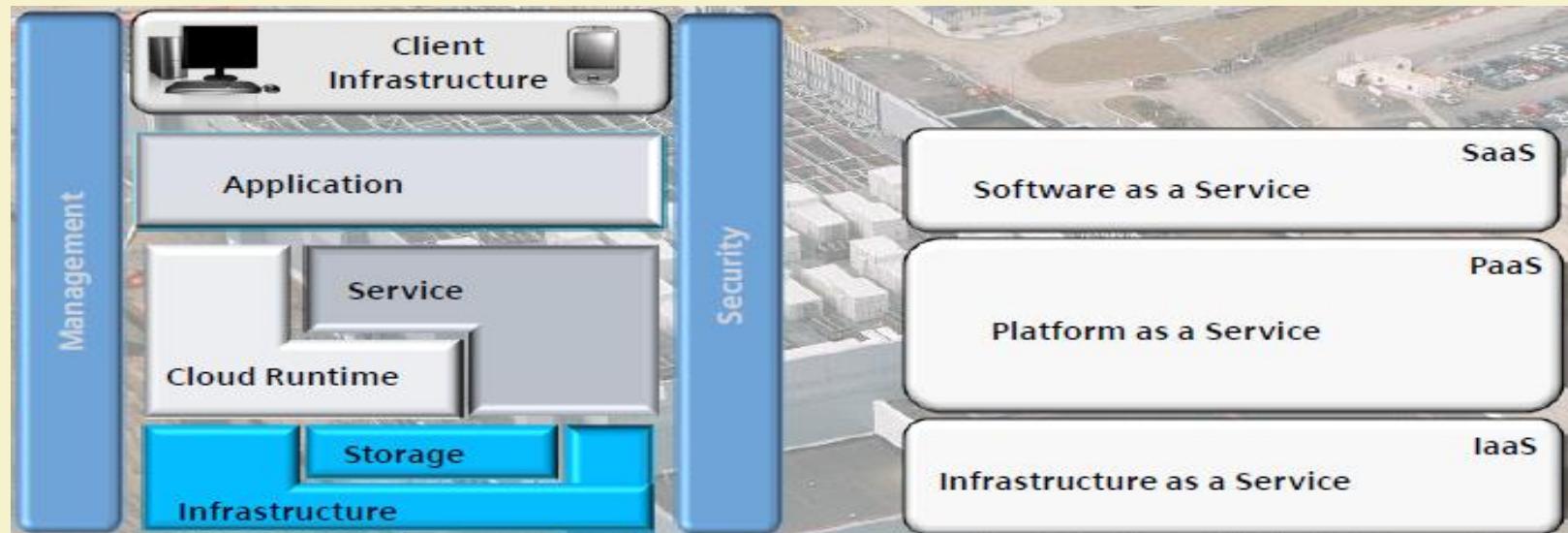
Source: <http://www.sei.cmu.edu/library/assets/presentations/Cloud%20Computing%20Architecture%20-%20Gerald%20Kaefer.pdf>

# 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

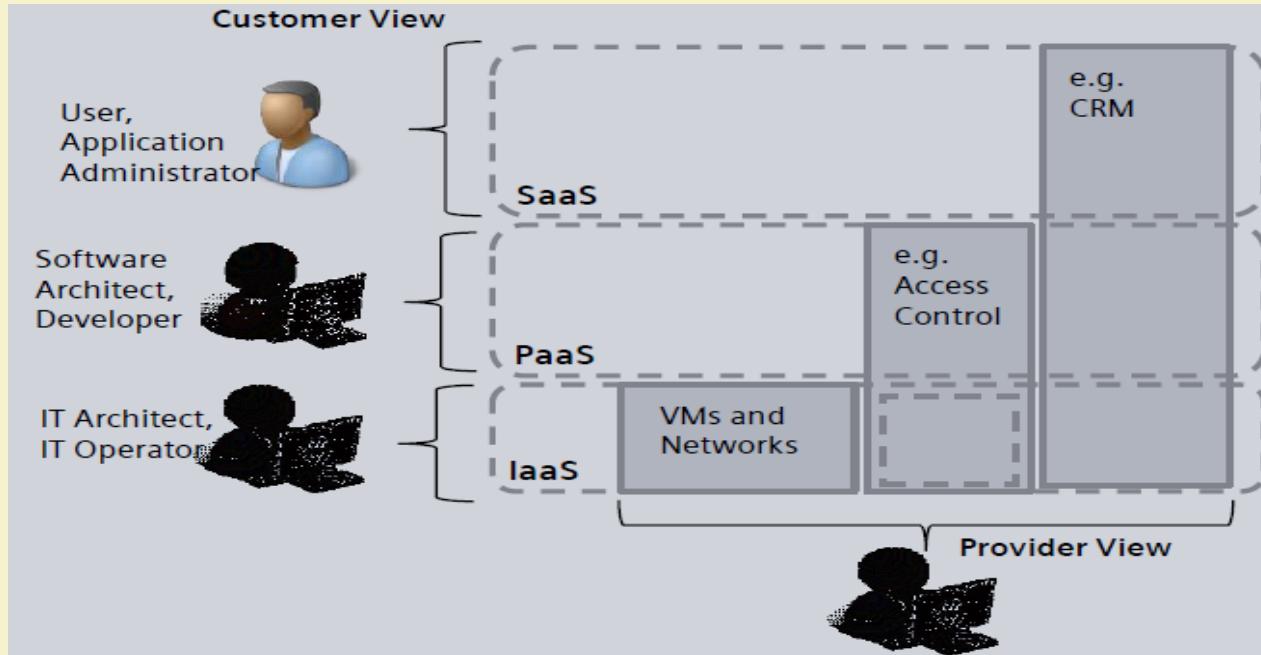
Ref: <http://www.sei.cmu.edu/library/assets/presentations/Cloud%20Computing%20Architecture%20-%20Gerald%20Kaefer.pdf>

# Cloud Computing Architecture - XaaS



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

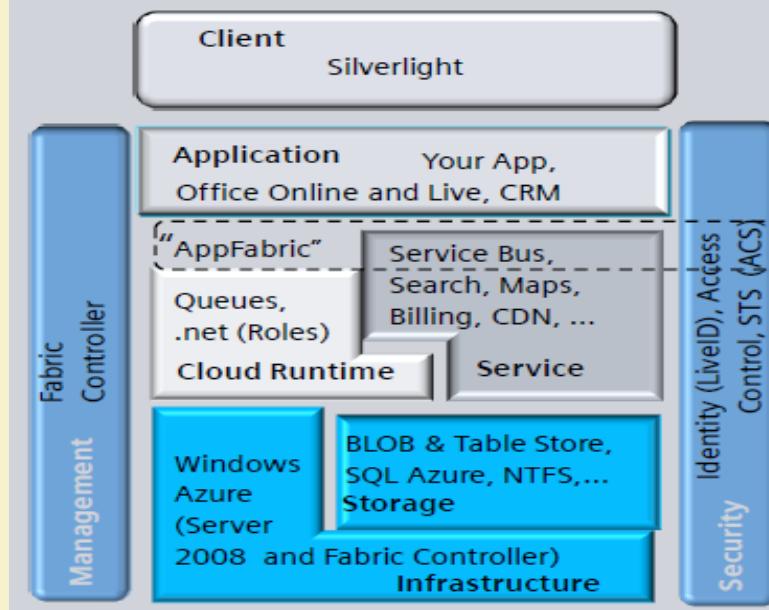
# XaaS Stack views: Customer view vs Provider view



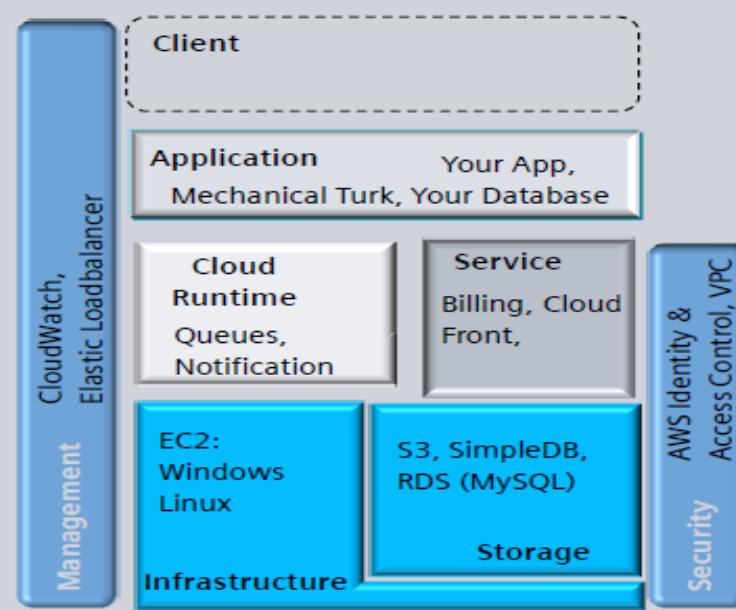
Source: <http://www.sei.cmu.edu/library/assets/presentations/Cloud%20Computing%20Architecture%20-%20Gerald%20Kaefer.pdf>

# Microsoft Azure vs Amazon EC2

e.g. Microsoft Windows Azure Platform



e.g. Amazon Cloud Platform



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

# Architecture for elasticity

## Vertical Scale Up

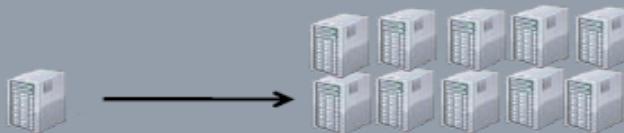
- Add more resources to a single computation unit i.e. Buy a bigger box
- Move a workload to a computation unit with more resources



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

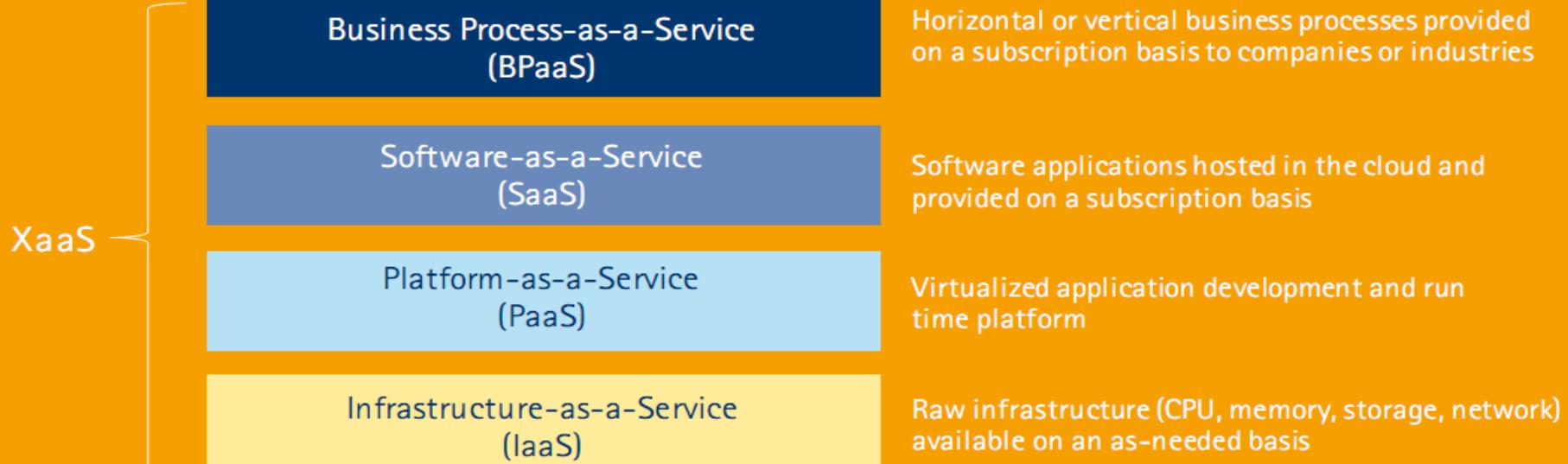


NPTEL  
NPTEL ONLINE  
CERTIFICATION COURSES

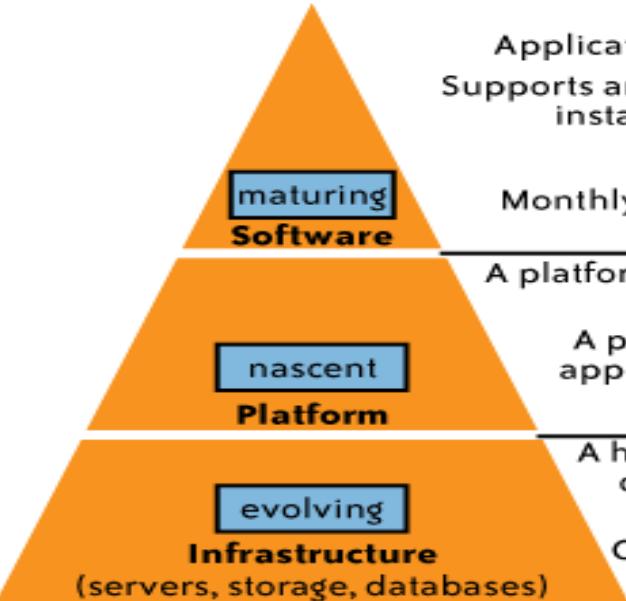
# Service Models (XaaS)

- 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)

	<b>Definition</b>	<b>Examples</b>
	<p>Applications that are enabled for the cloud</p> <p>Supports an architecture that can run multiple instances of itself regardless of location</p> <p>Stateless application architecture</p> <p>Monthly subscription-based pricing model</p>	<p>• Google Docs</p> <p>• MobileMe</p> <p>• Zoho</p>
	<p>A platform that enables developers to write applications that run on the cloud</p> <p>A platform would usually have several application services available for quick deployment</p>	<p>• Microsoft Azure</p> <p>• Google App Engine</p> <p>• Force.com</p>
	<p>A highly scaled redundant and shared computing infrastructure accessible using Internet technologies</p> <p>Consists of servers, storage, security, databases, and other peripherals</p>	<p>• Amazon EC2, S3, etc.</p> <p>• Rackspace Mosso offering</p> <p>• Sun's cloud services</p> <p>• Terremark cloud offering</p>

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



IIT KHARAGPUR



NPTEL ONLINE  
CERTIFICATION COURSES

# 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!

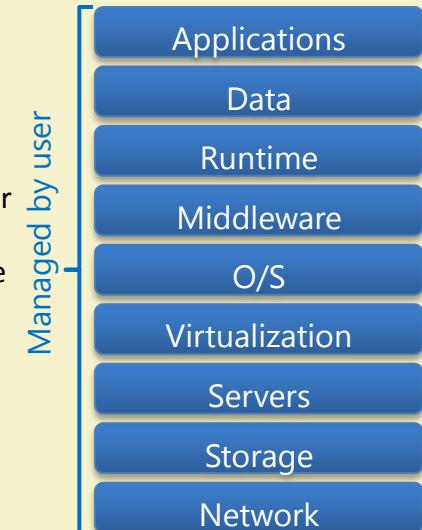


Source: Understanding the Cloud Computing Stack: PaaS, SaaS, IaaS © Diversity Limited, 2011

# Classical Service Model

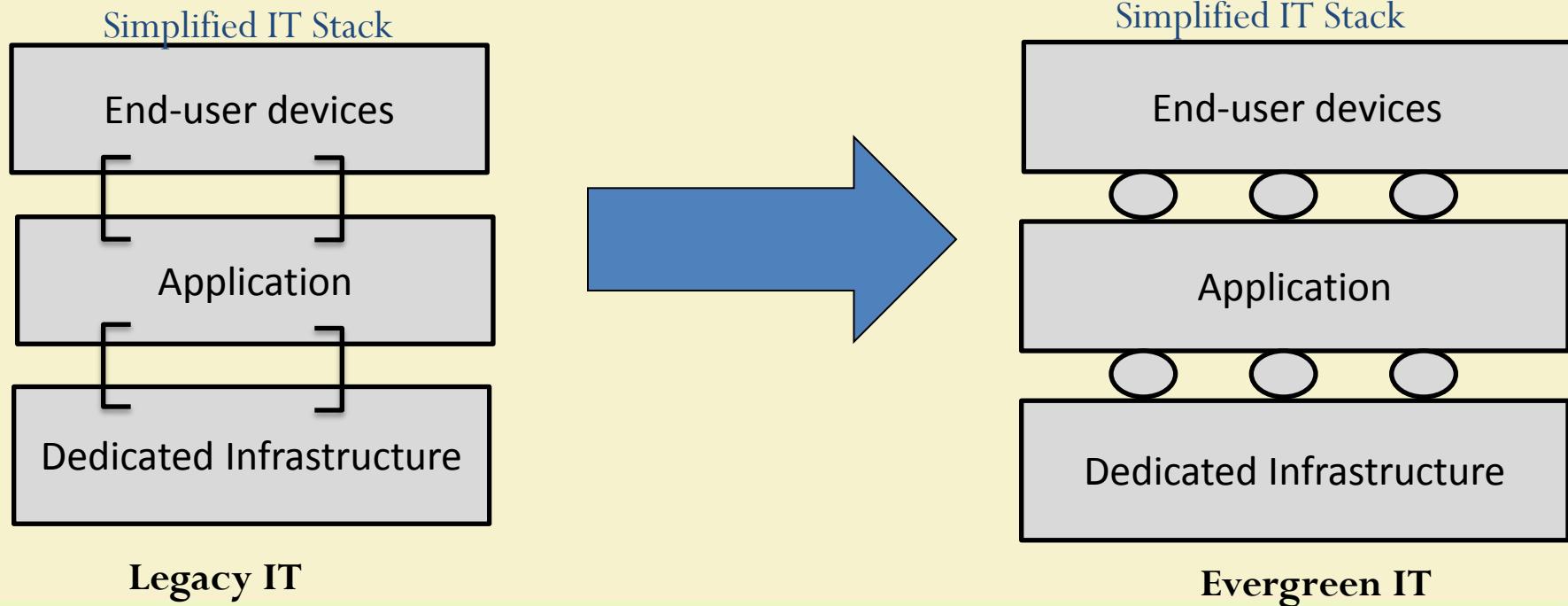
- All the Layers(H/W, Operating System, Development Tools, Applications) Managed by the Users
- Initial IT budget and resources.
- Users bears the costs of the hardware, maintenance and technology.
- Each system is designed and funded for a specific business activity: custom build-to-order
- Systems 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”

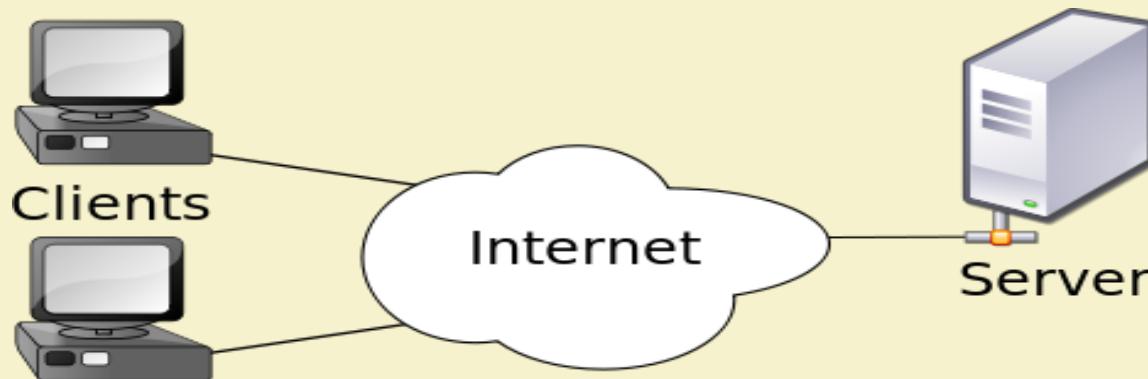
# Key impact of cloud computing for IT function: From Legacy IT to Evergreen IT



# 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

# Thank You!



IIT KHARAGPUR



NPTEL ONLINE  
CERTIFICATION COURSES



IIT KHARAGPUR



NPTEL ONLINE  
CERTIFICATION COURSES

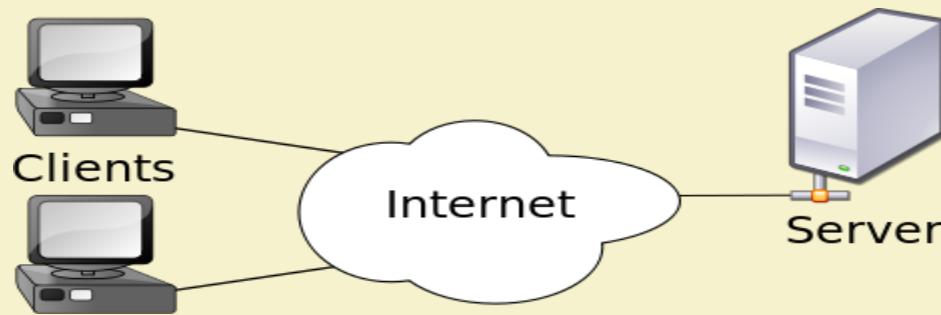
# CLOUD COMPUTING

## 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

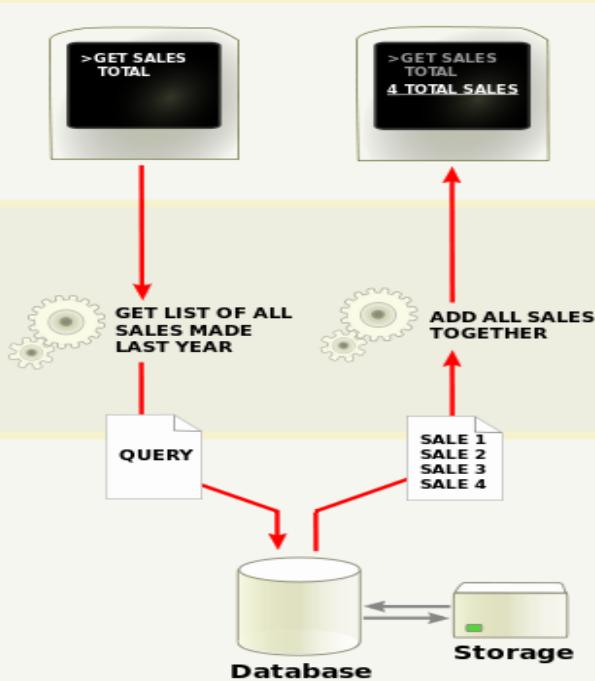
The top-most level of the application is the user interface. The main function of the interface is to translate tasks and results to something the user can understand.

## 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**

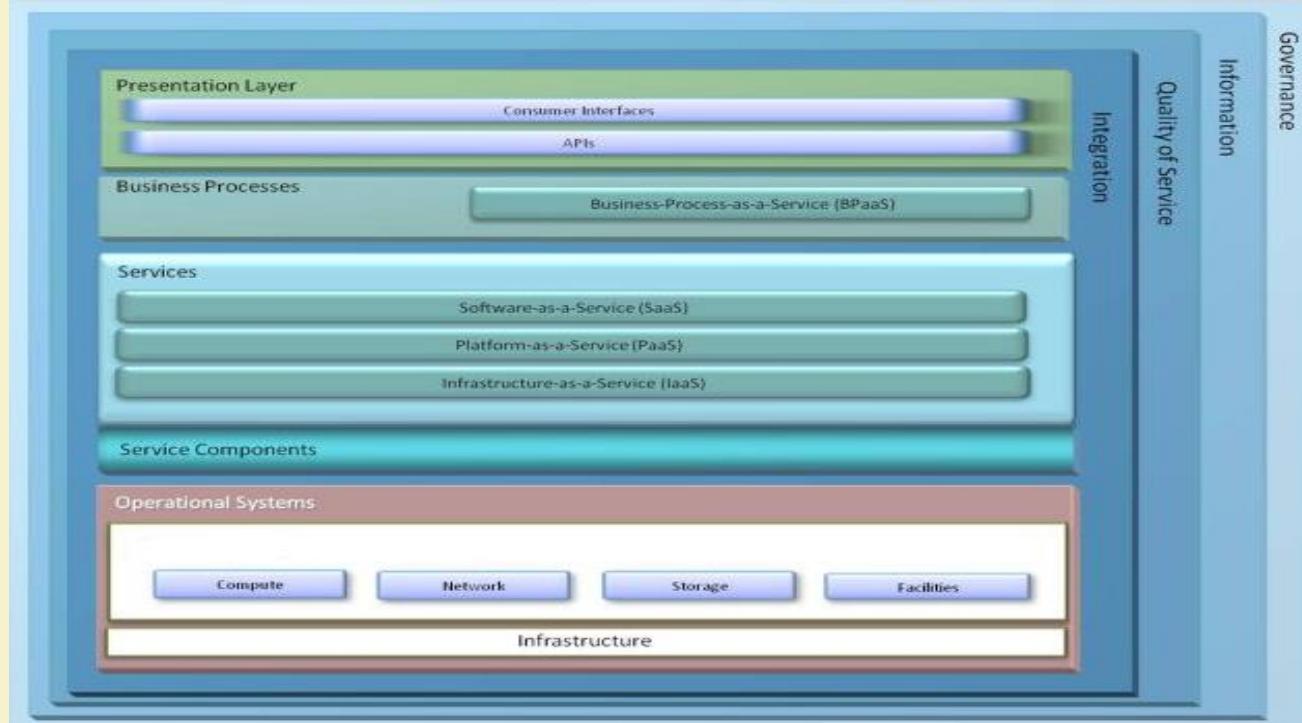
## **Client server model**

- 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

# Cloud Services



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

# Cloud service models

Service Class	Main Access & Management Tool	Service content
 SaaS	Web Browser	<b>Cloud Applications</b> Social networks, Office suites, CRM, Video processing
 PaaS	Cloud Development Environment	<b>Cloud Platform</b> Programming languages, Frameworks, Mashups editors, Structured data
 IaaS	Virtual Infrastructure Manager	<b>Cloud Infrastructure</b> Compute Servers, Data Storage, Firewall, Load Balancer

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

# 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](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](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](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](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](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](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](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](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](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](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>

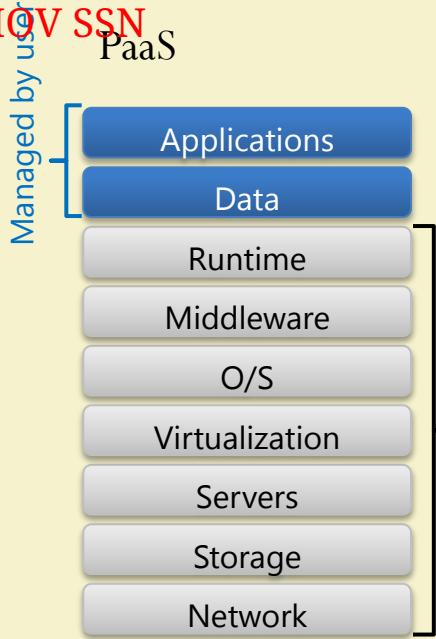
# XaaS

SaaS

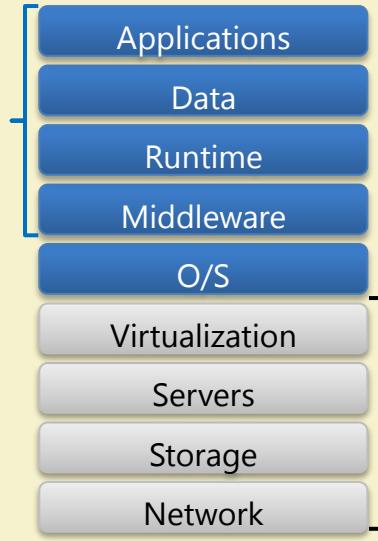


ADR MOV SSN

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](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](https://portal.etsi.org/nfv/nfv_white_paper.pdf)



IIT KHARAGPUR



NPTEL  
NPTEL ONLINE  
CERTIFICATION COURSES

# Thank You!!



IIT KHARAGPUR



NPTEL ONLINE  
CERTIFICATION COURSES

## Cloud Storage

Cloud Storage is a service where data is remotely maintained, managed, and backed up. The service is available to users over a network, which is usually the internet. It allows the user to store files online so that the user can access them from any location via the internet.

### Advantage of cloud storage

- \* Usability and Accessibility
- \*Disaster Recovery(RTO,RPO)
- \*Cost Saving
- \*Collaboration

### Disadvantage

- \* Security
- \*Auditing
- \*Recovery

GFS(google file system)== HDFS(Hadoop distributed File System)Yahoo

### Description of HDFS

- \*Master-slave Architecture
- \*HDFS can created Commodity Hardware as it has Fault Tolerance Capability(Replication,Heart Beat message,Secondary Name Node)
- \*HADOOP support MAP-REDUCE Programming Paradigm
- \*Read Request\_Request send to Master Node,masternode will provide meta data information and then read request is transferred to Datanode
- \*Write Request :All Replica are updated but initially primary replica is update.  
Default Replication:3

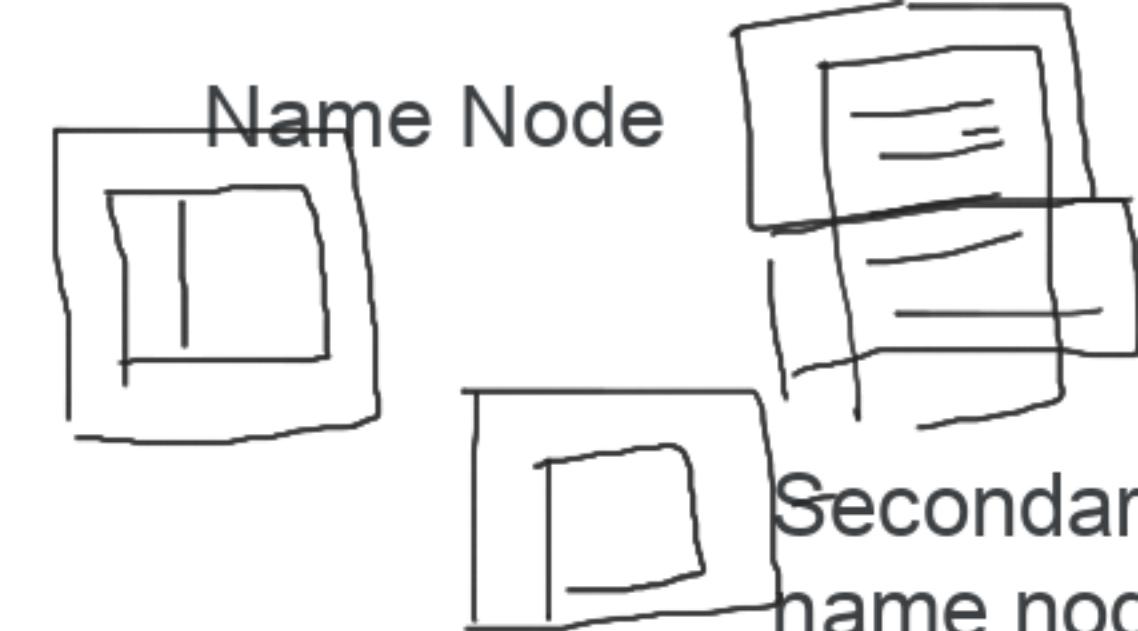
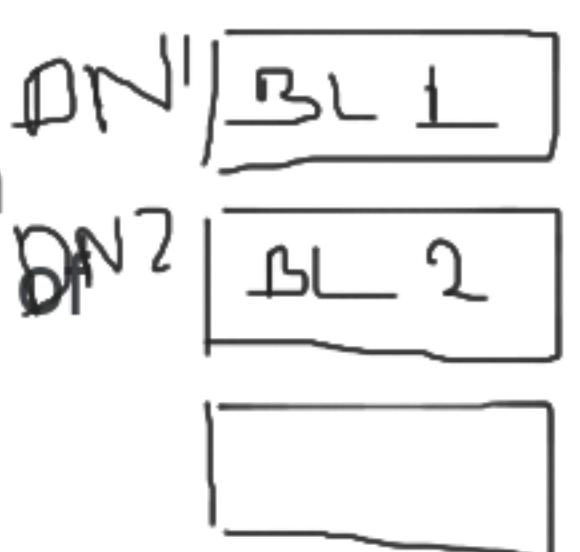
## Features of HDFS(Hadoop Distributed file System)

- \*it can be installed on commodity hardware(group of machine/cluster)
- \*handle structured and unstructured data
- \*Fault Tolerance(replication)(secondary namenode)
- \*Automatic data Recovery and Replication.

High Availability and Throughput

HDFS1 64 MB Block Size

HDFS2 128 MB block size



Name node failure management:

- \*NN store metadata information
- \*File id/Blck no/ replica no/ DN no./ owner/permission/
- \*Metadatafile:fsimage(status of HDFS) and edits (logs)
- \*copied to 2nd NN and checkpoints periodically
- \*remote storage /backup of NN files

a) 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?

1 file 1Block replication 3 blocks

2nd file 65MB/64MB 2 BLOCKS after replication =6blocks

3rd file 127/64 =2 blocks after replication =6blocks

total=3+6+6=15 blocks

## Traditional Computing

- \*Data is sent to computation unit
- \*it work better in small dataset

## Data intensive Computing

- \*
- handle Big data(volume, velocity and varity)
- \*data and process are combine ,so cost and time is reduced
- \* data is split into blocks
- \* Mapreduce paradigm it will provide more scalability and reduce computation time

Data intensive computing uses  
these two types of storage

### File Storage

- \* Hirarchical structure
- \*Less Scalable
- \*Slow Performance
- \* Simple \* sharing
- \*NFS,NAS

### Block Storage

- \*Flat but with no contiguous memory alloaction
- \*SAN
- \*datamanagement is simple and better
- \*better performance as compared to file

### Object Storage

- \*flat stucture(id)
- \*Scalable
- \* fast access capability
- \*unstructured data video, audio,
- \*cost efficient
- \*AWS Simple Storage Service

- \*Software Defined Storage(SDS) (cloud)
- \*Control plane(policy) and device plane
- \*better storage configuration based on requirement

Cloud provides different storage classes

- \*Frequently access
- \*Cold Storage
- \*Archive

# Map Reduce Programming Model

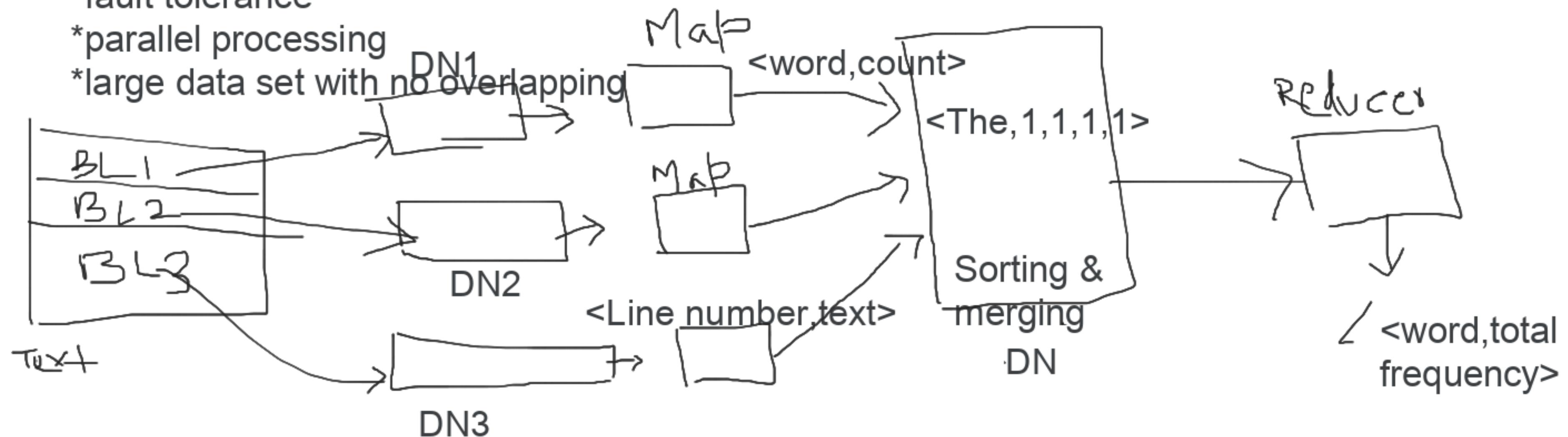
\*google had devlope this paradigm

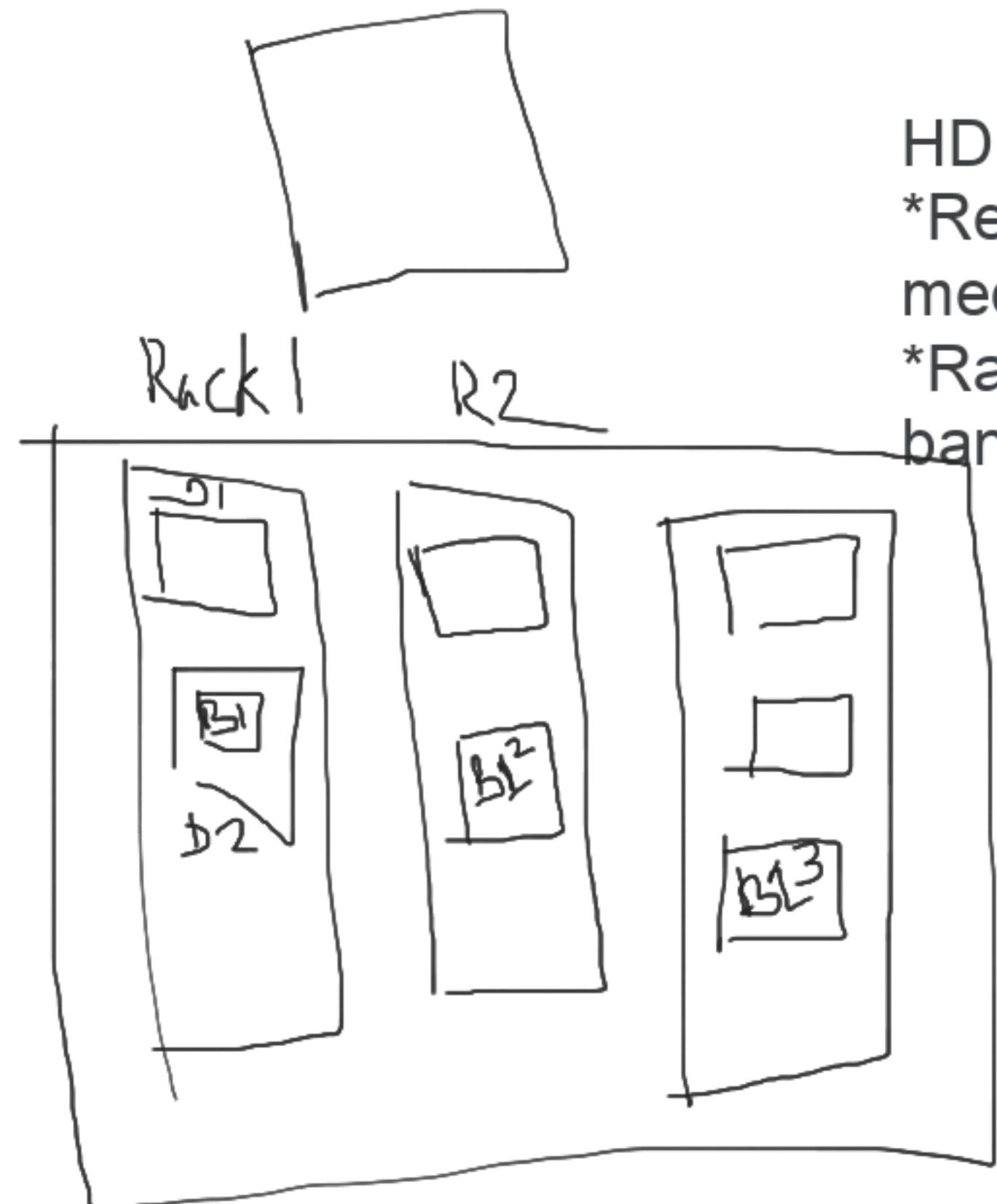
## Features

- \*data aware computing(data location is known while submitting job)
- \*Simplicity(sending process towards data())
- \*scalability
- \*fault tolerance
- \*parallel processing
- \*large data set with no overlapping

## Word Count Problem

- \*count frequency of word  
<key value>



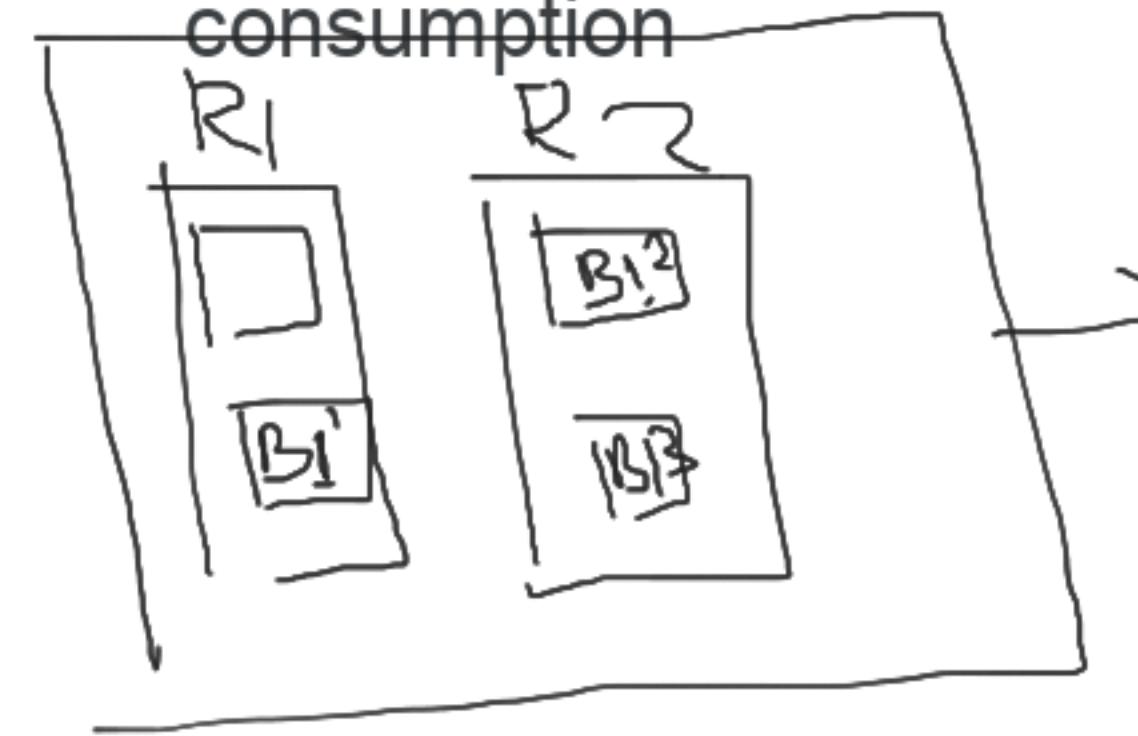


HDFS support fault tolerance :

\*Replication(Redundancy) it is done through pipeline mechanism

\*Rack aware placement(consume more network bandwidth)

Rack aware placement with minimum bandwidth consumption



good balance between replication and bandwidth

## Hadoop

- \* Structured(text.table)/unstructured (audio/vedio) data
- \* no schema/no restriction
- \* Row wise(no constraint)
- \* No ACID property restriction

## Databases

- \* structured
- \* fixed schema
- \* column wise
- \* Supports ACID property

Hadoop main Component

\*HDFS

\*Map Reduce

\*YARN(Yet Another Resource Manager)

Main Class

Mapper

Reducer

Job -Driver class

Jobtracker will run on the name node and Task Tracker  
run on the data node



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 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

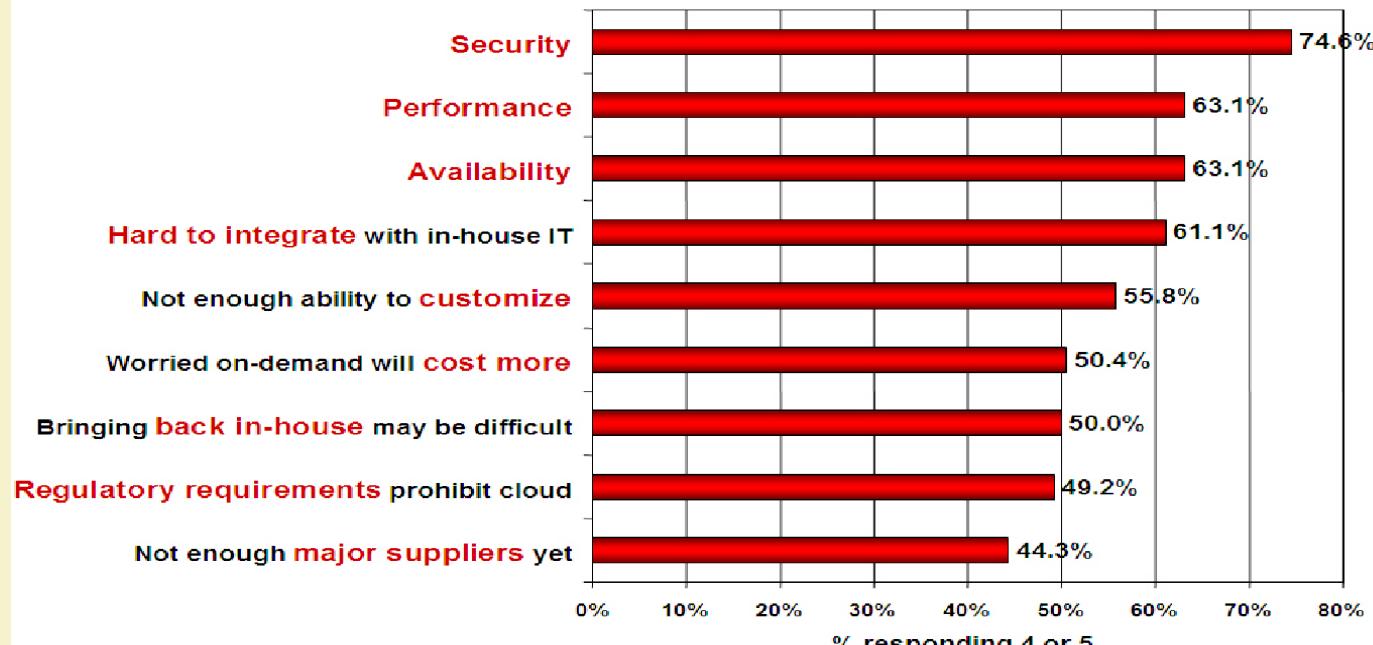


NPTEL  
NPTEL ONLINE  
CERTIFICATION COURSES

# 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

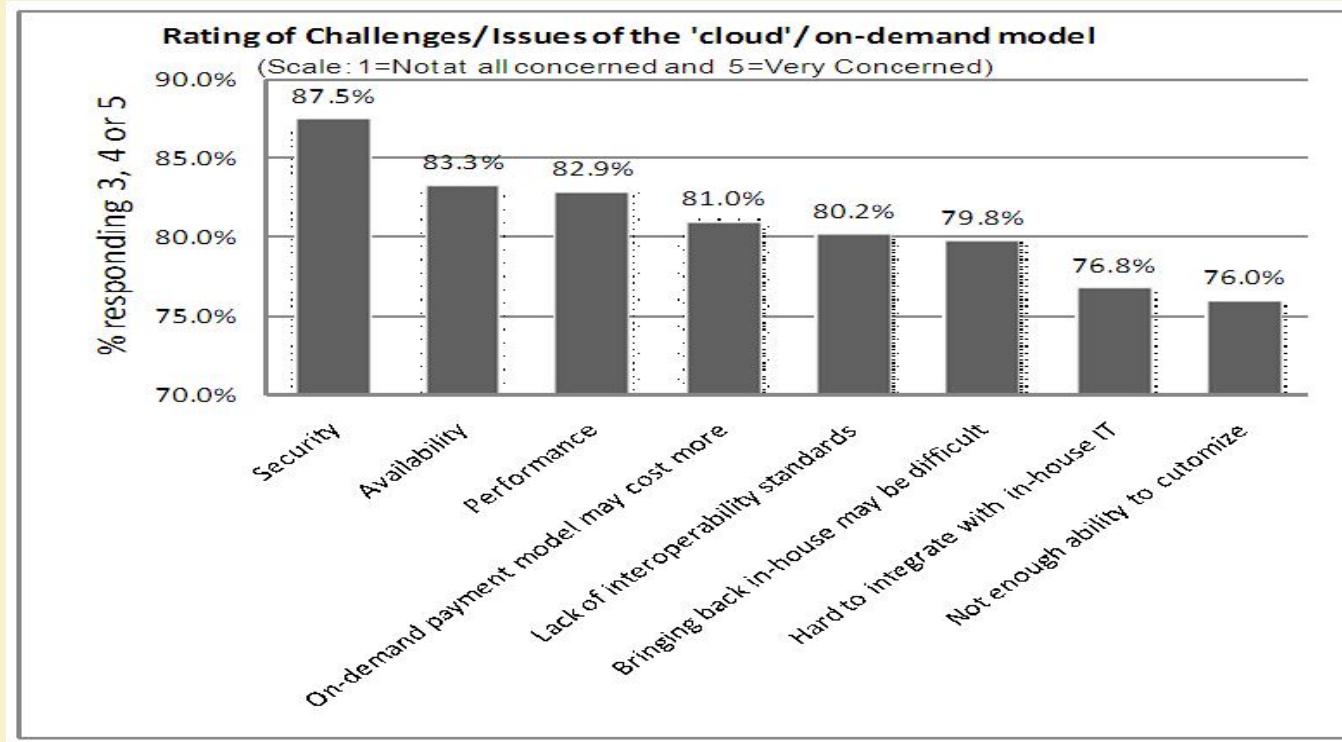


IIT KHARAGPUR



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**



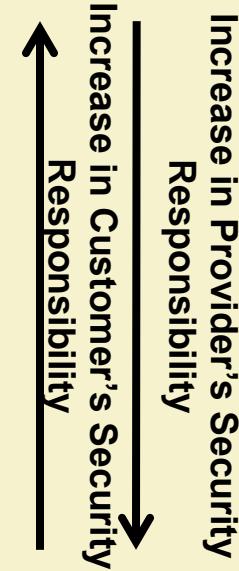
IIT KHARAGPUR



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

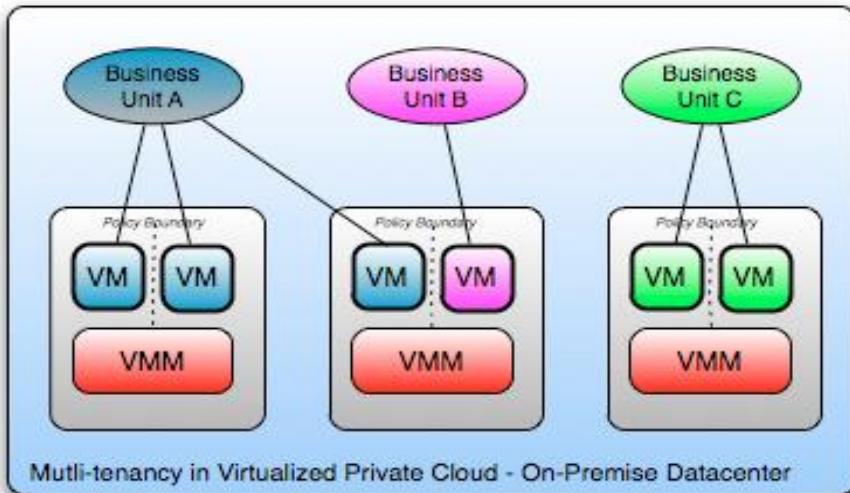


IIT KHARAGPUR

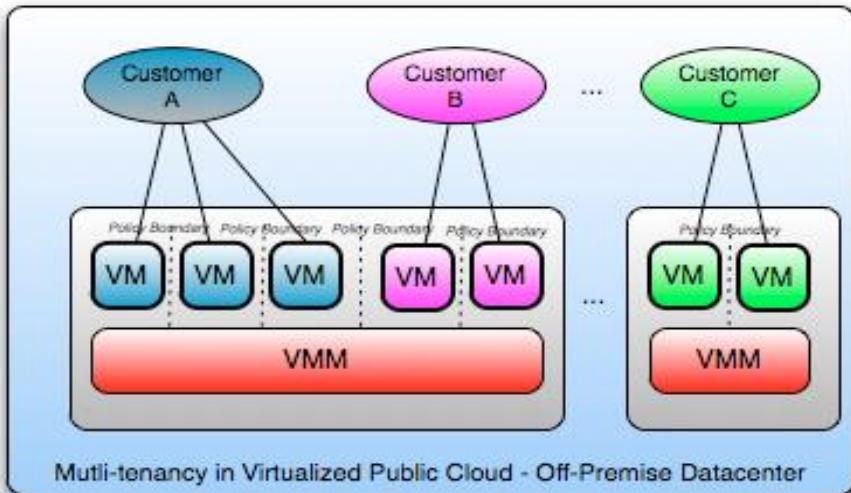


NPTEL  
NPTEL ONLINE  
CERTIFICATION COURSES

# 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 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



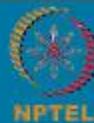
NPTEL  
NPTEL ONLINE  
CERTIFICATION COURSES

# 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  
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 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 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

# 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  
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



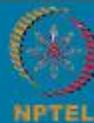
NPTEL ONLINE  
CERTIFICATION COURSES

# 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 ONLINE  
CERTIFICATION COURSES

# Thank You!



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?



IIT KHARAGPUR



NPTEL  
NPTEL ONLINE  
CERTIFICATION COURSES

# 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



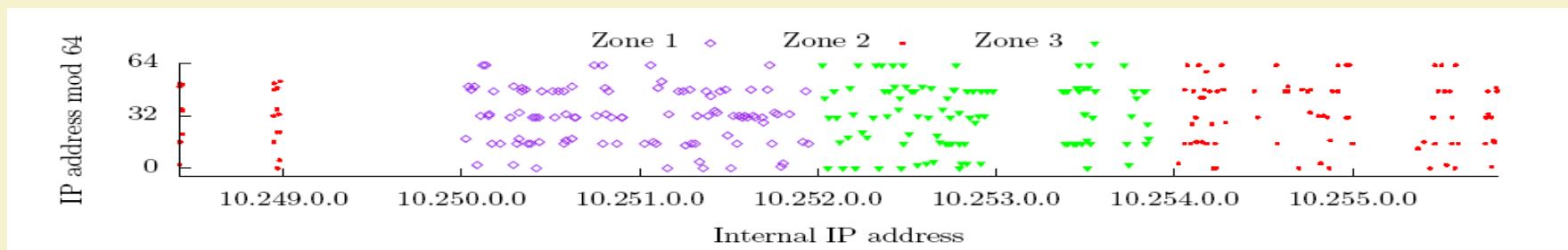
IIT KHARAGPUR



NPTEL  
NPTEL ONLINE  
CERTIFICATION COURSES

# Instance Placement Parameters

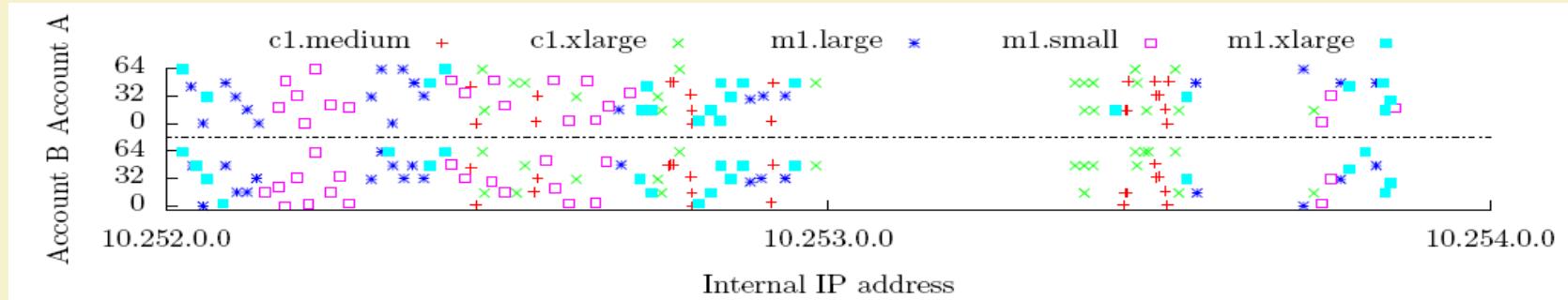
- EC2's internal address space is cleanly partitioned between availability zones
  - Three availability zone; five instance-type/zone
  - 20 instances launched for each of the 15 availability zone-instance type pairs from a particular account (Say, Account A)



- 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

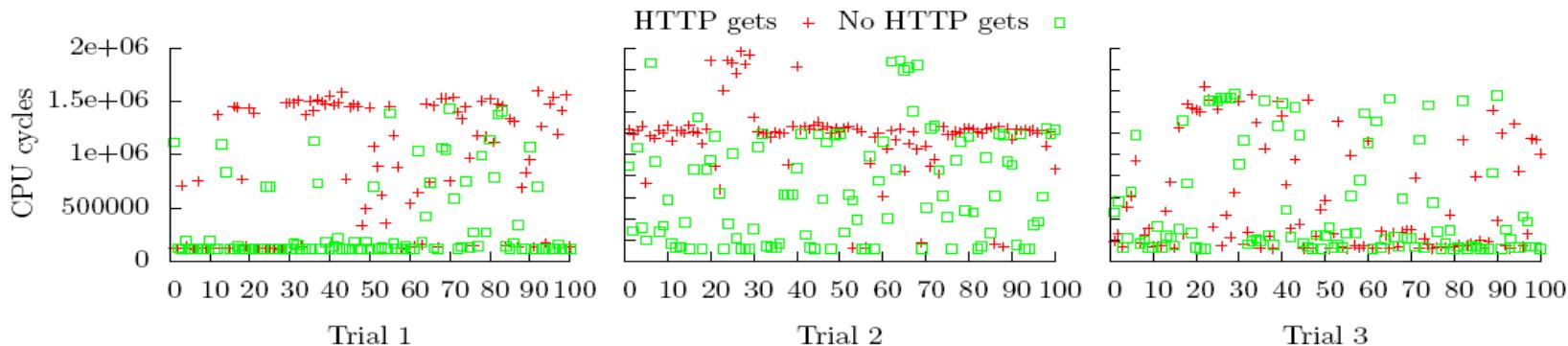
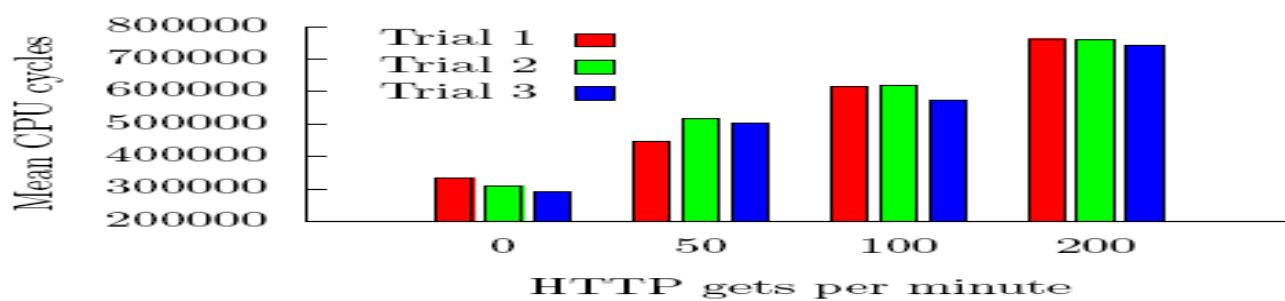


Figure 5: Results of executing 100 Prime+Trigger+Probe cache timing measurements for three pairs of m1.small instances, both when concurrently making HTTP get requests and when not. Instances in Trial 1 and Trial 2 were co-resident on distinct physical machines. Instances in Trial 3 were not co-resident.

- 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  
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!



IIT KHARAGPUR



NPTEL ONLINE  
CERTIFICATION COURSES

CSA-Cloud Security Alliance-

\*Monitoring and controlling authority and Cloud Security issues related to Cloud Platform

Threats to Cloud Security

\*data loss/lose control on the data

\*Account Hijacked

Remedial(Authentication and authorization rule)

\*Shared Technology related issues.(multitenancy)(Hypervisor,VM modification,DDOS,Improper Isolation)

Remedial: proper monitoring and scanning and proper access control.

\*Insecure API

\*Abuse and Nefarious use of Cloud/Malicious Insider :

Remedial :

how CSP is accessing your resources,  
local and global authorization rule/policies must be framed properly

## Remedial action

- \*Secure Methodology to access cloud Resources
- \*Strong Authentication and access control rules
- \*Vulnerability scanning
- \*Log analysis
- \*Monitoring and auditing
- \*SLA
- \*Proper Isolation/compartmentalization
- \*data Categorization and Protection
- \*Monitoring
  - \*proper compartmentalization
  - \*proper installation
  - \*strong authentication and access control
  - \*
  - \*SLA/policies
  - \*vulnerability scanning

## Identity and Access management(IAM)

It is primary mechanism for \*

\*Authentication

\*Access control on data

\*Authorization control

\*Maintain roles

\*Comply with Regulation

\*

useful for

\*Automation

\*Portability

\*Build trust

## IDaaS

Identity refers to set of attributes associated with something and make it recognizable. All objects may have same attributes, but their identity cannot be the same. This unique identity is assigned through unique identification attribute.

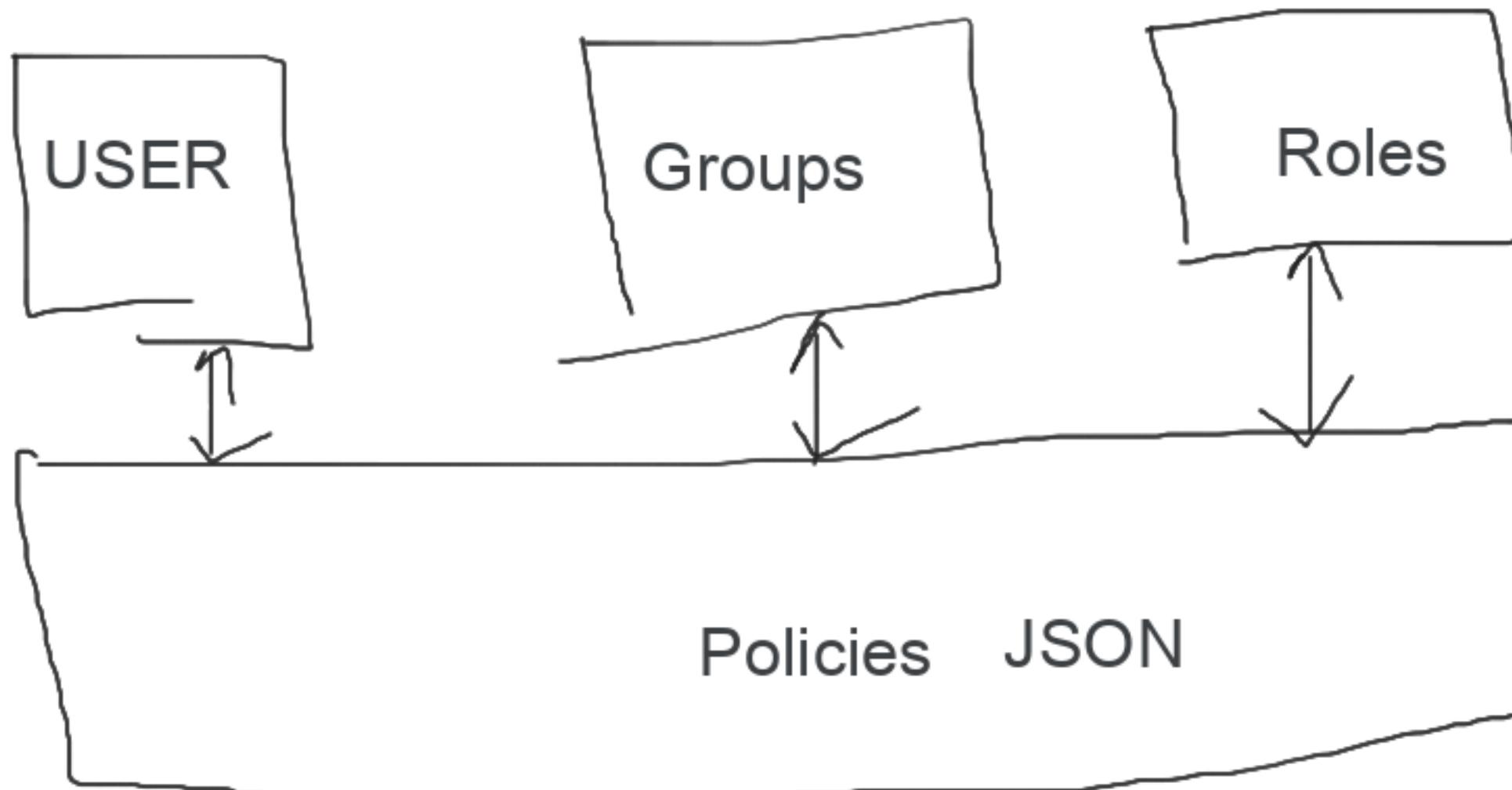
:

Registration  
Authentication  
Directory Service  
Federated Services  
Identity Governance  
Profile management  
Policy Enforcement



SSO(Single Sign On) → Portability  
Organization can integrate their (IAM) repository with other platform

# IAM



Security  
Groups :They control Inbound and outbound Traffic to ur VM

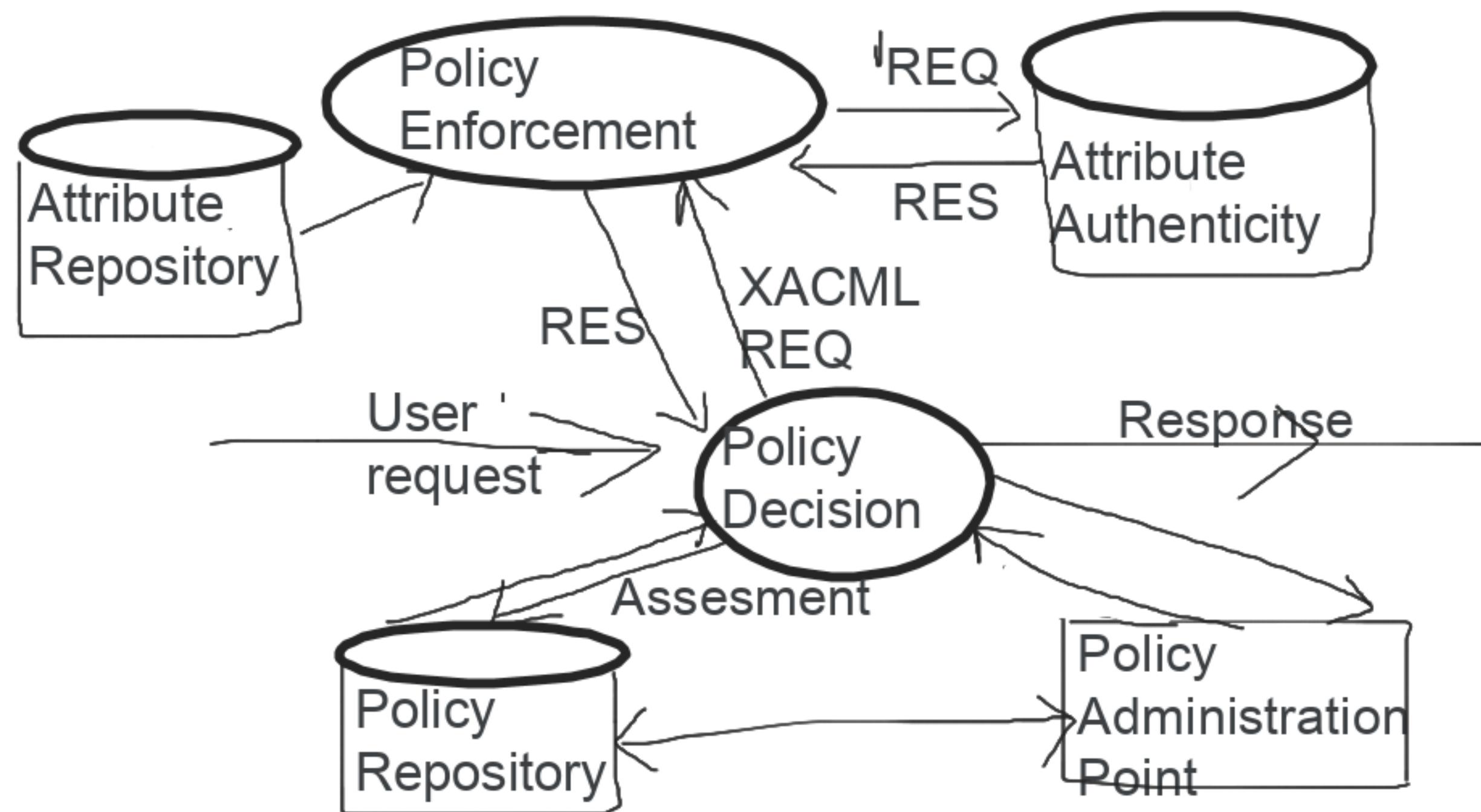
- Features of IAM
- \*Fine Grain Control on resources
  - \*Enhanced the Security
  - \*Multifactor authentication
  - \*Flexible Credential system
  - \*redemine the polocy easily

\*IDP( Identity Provider) Open-Id

## Architecture of Identity Management

\*Versign

XACML: Extensible Access Control Markup language



## Cloud Federation

- \*Collaboration of more than one cloud Platform
- \*To Shared the workload spikes
- \*To share resources and also to share revenue gain
- \*Interoperability and Portabiliry in cloud platform
- \*DMTF(Distributed Management Task force) organization to standarization cloud interopration issues
- \*security issues of Inter cloud/Cloud federation can also be handle through IAM mechanism

# Conventional Computing

vs.

# Cloud Computing

## Conventional

- Manually Provisioned
- Dedicated Hardware
- Fixed Capacity
- Pay for Capacity
- Capital & Operational Expenses

## Cloud

- Self-provisioned
- Shared Hardware
- Elastic Capacity
- Pay for Use
- Operational Expenses

# Cloud Computing Characteristics

## Common Characteristics:

**Massive Scale**

**Resilient Computing**

**Homogeneity**

**Geographic Distribution**

**Virtualization**

**Service Orientation**

**Low Cost Software**

**Advanced Security**

## Essential Characteristics:

**On Demand Self-Service**

**Broad Network Access**

**Scalable and Elastic**

**Shared/Pool Resources**

**Metered by Use**

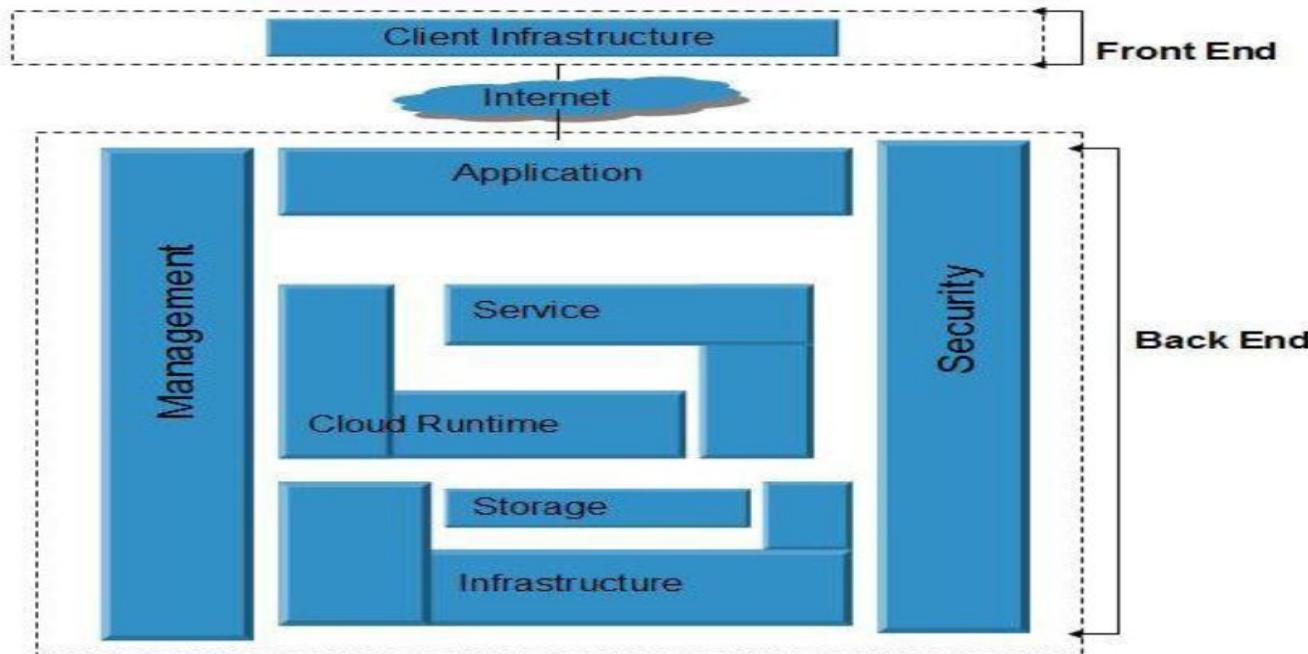
# Advantages and Disadvantages: Cloud Summary

- Advantages
  - Lower computer cost
  - Improved performance
  - Reduced software cost
  - Instant software update
  - Improved doc. Format compatibility
  - Unlimited capacity
  - Increased data reliability
  - Universal doc. Access
  - Version control and availability
  - Easy group collaboration
  - Portability, device independence
- Disadvantages
  - Requires const. internet connection
  - Bad performance with low speed connection
  - Limited features
  - Security and data loss issues

# Elastic Computing

- Elastic computing is the ability to quickly expand or decrease computer processing, memory and storage resources to meet changing demands without worrying about capacity planning and engineering for peak usage. Typically controlled by system monitoring tools, elastic computing matches the amount of resources allocated to the amount of resources actually needed without disrupting operations. With cloud elasticity, a company avoids paying for unused capacity or idle resources and does not have to worry about investing in the purchase or maintenance of additional resources and equipment.
-

# Cloud Computing Reference Architecture



## FRONT END

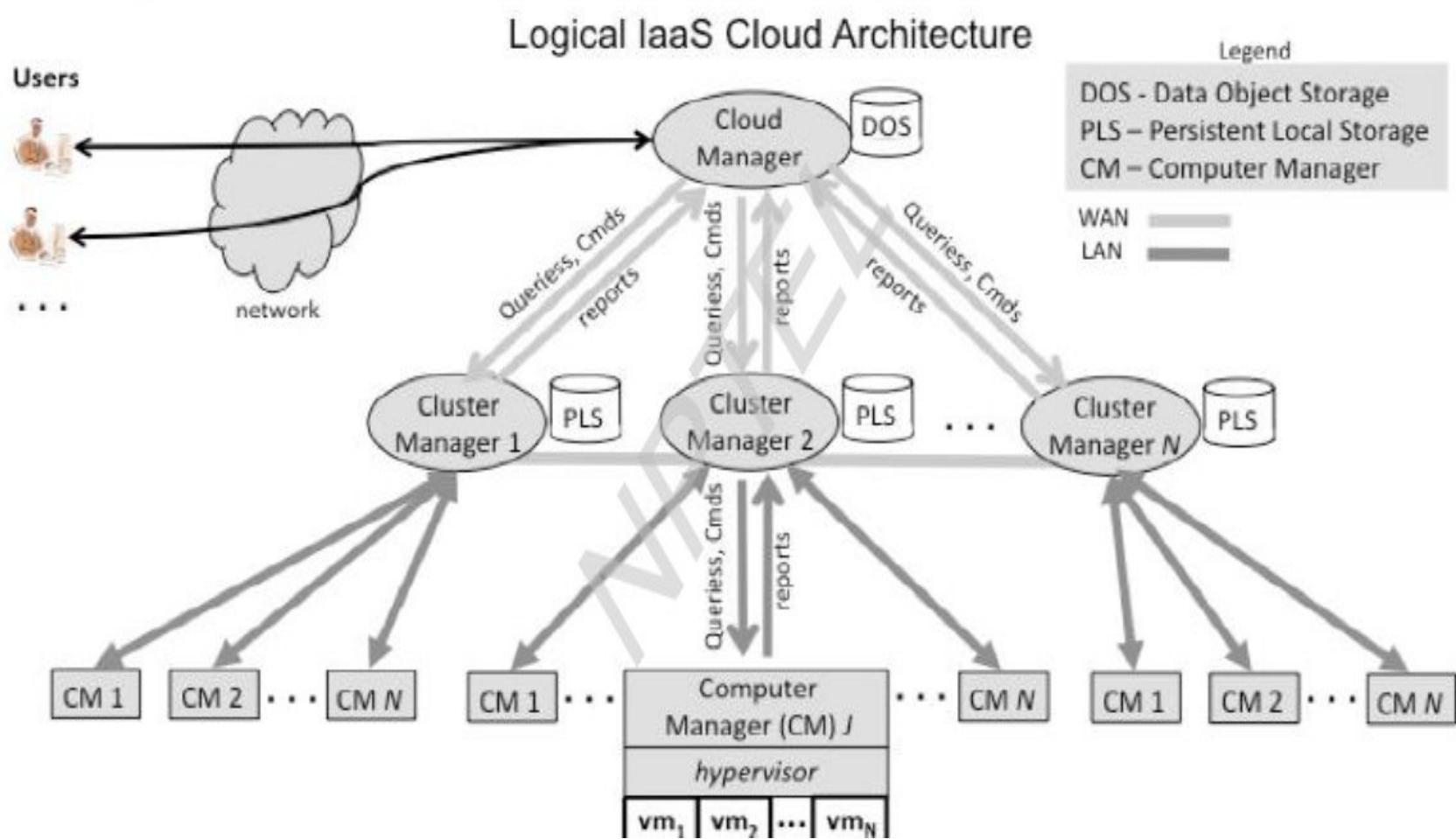
**Front End** refers to the client part of cloud computing system. It consists of interfaces and applications that are required to access the cloud computing platforms, e.g., Web Browser.

## BACK END

**Back End** refers to the cloud itself. It consists of all the resources required to provide cloud computing services. It comprises of huge data storage, virtual machines, security mechanism, services, deployment models, servers, etc.

# IaaS Cloud Architecture

- Logical view of IaaS cloud structure and operation



# 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

# 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

# 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*

# 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

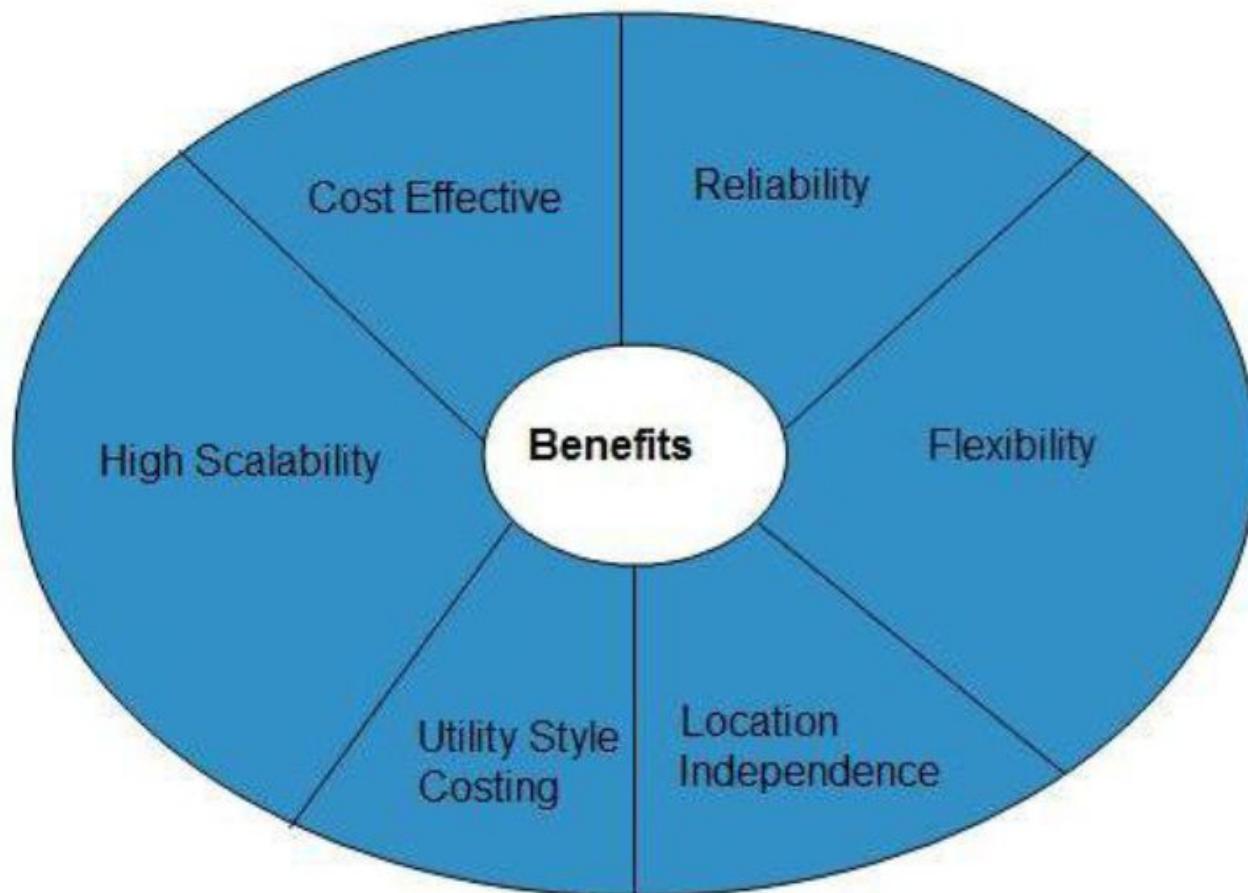
# Operation of the Computer Managers

- 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

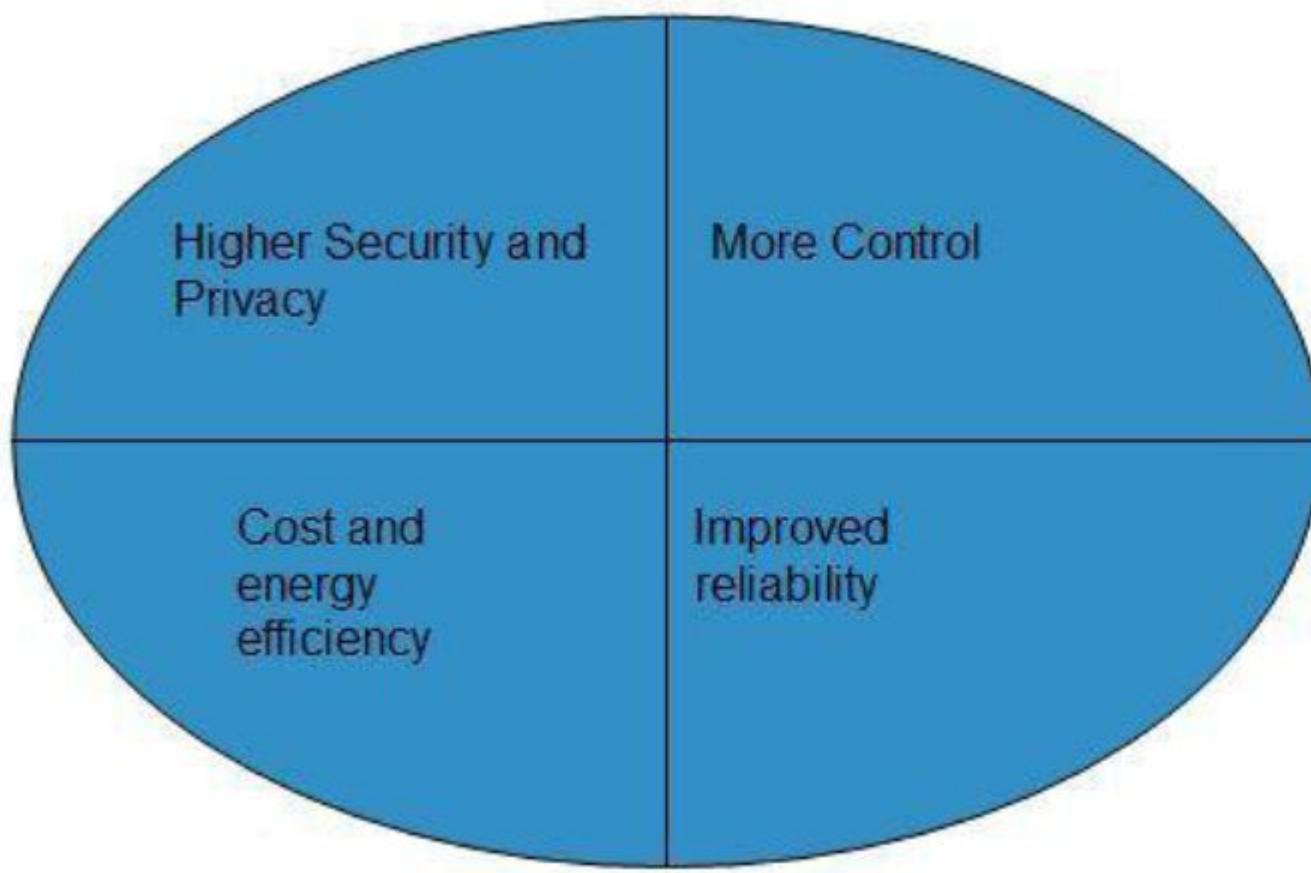
# XaaS

- 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.

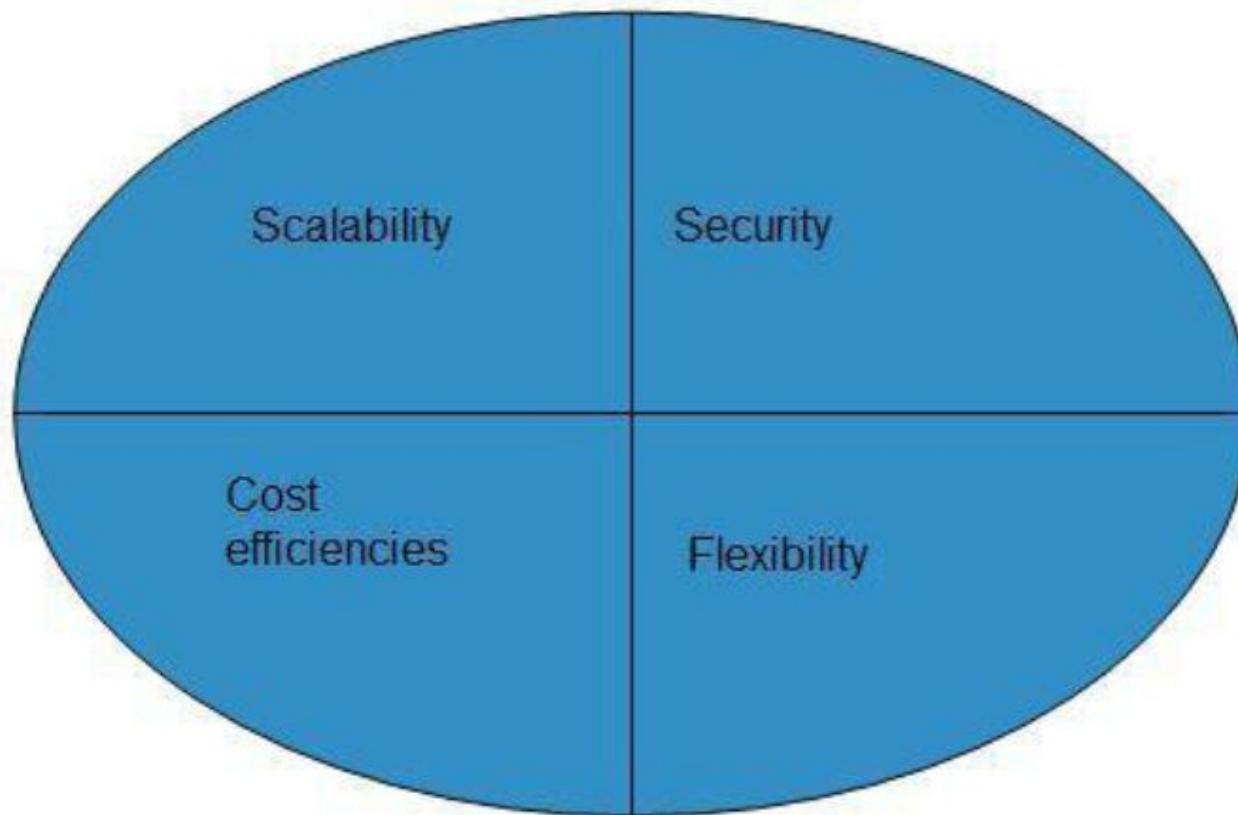
# Benefits of Public Cloud Model



# Benefits of Private Cloud



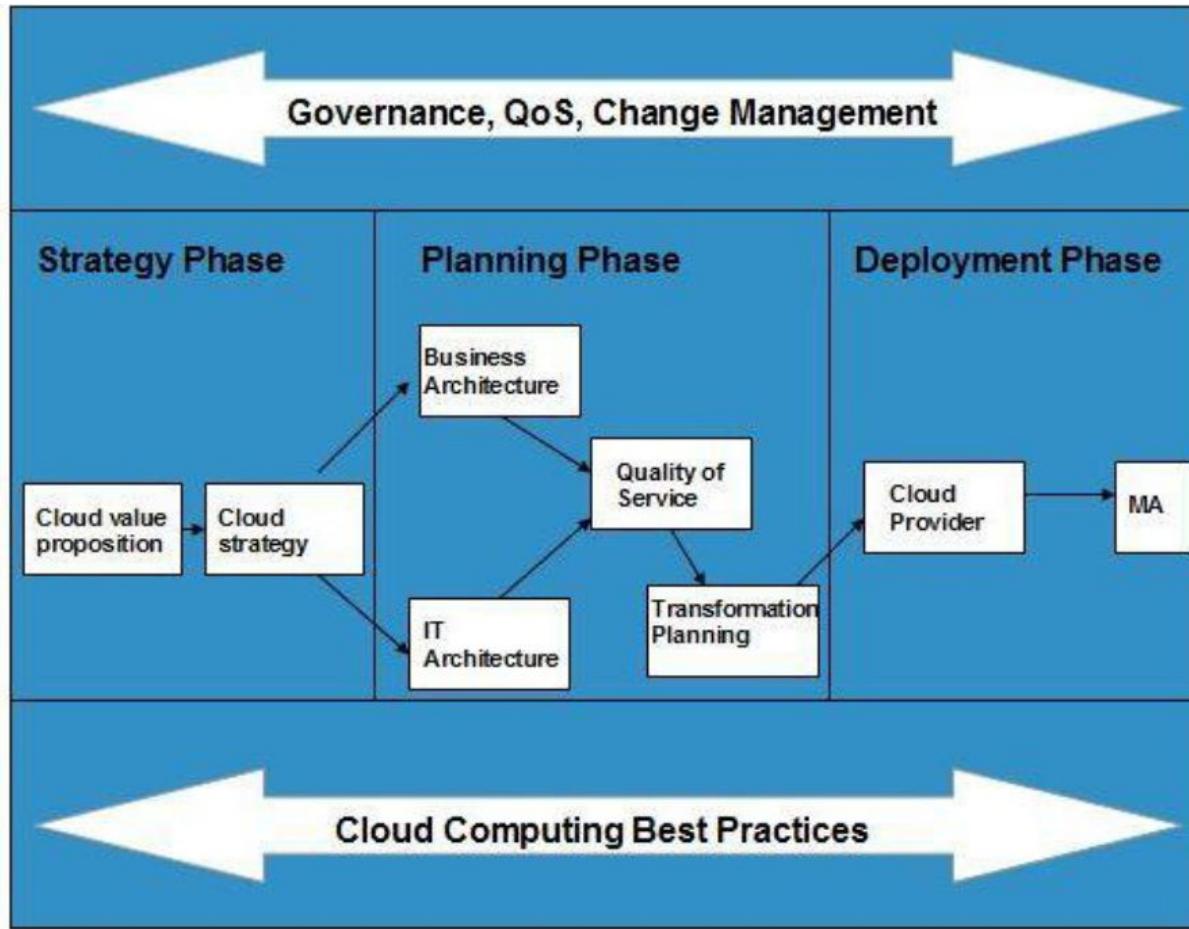
# Benefits of Hybrid cloud Model



# Disadvantage of Following Deployment Model

Public Cloud+	Private Cloud	Hybrid Cloud
LOW SECURITY	RESTRICTED AREA	NETWORKING ISSUES
LESS CUSTOMIZABLE	LIMITED SCALABILITY	SECURITY COMPLIANCE

# Cloud Computing Best Practice



# Cloud Computing Planning

- Following are the issues one must have to think about before opting Cloud Computing for organization:
- Data Security and Privacy Requirement
- Budget Requirements
- Type of cloud - public, private or hybrid
- Data backup requirements
- Training requirements
- Dashboard and reporting requirements
- Client access requirements
- Data export requirements

To meet all of these requirements, it is necessary to have well-compiled planning.

- There are two steps to perform this analysis:
  - Cloud Computing Value Proposition
  - Cloud Computing Strategy Planning
- 
- **CLOUD COMPUTING VALUE PROPOSITION**
  - In this, we analyze the factors influencing the customers when applying cloud computing mode and target the key problems they wish to solve. These key factors are:
    - IT management simplification
    - operation and maintenance cost reduction
    - business mode innovation
    - low cost outsourcing hosting
    - high service quality outsourcing hosting.

# MULTI-TENANCY

- Multi-tenancy is an architectural pattern
- A single instance of the software is run on the service provider's infrastructure
- Multiple tenants access the same instance.
- In contrast to the multi-user model, multi-tenancy requires customizing the single instance according to the multi-faceted requirements of many tenants.

# MULTI-TENANCY

A Multi-tenants application lets customers (tenants) share the **same hardware resources**, by offering them one shared application and database instance ,while allowing them to **configure the application to fit there needs** as if it runs on dedicated environment.

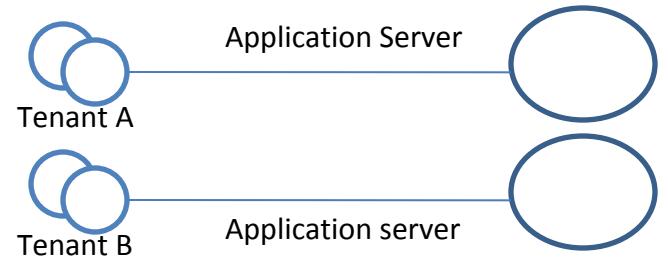
These definition focus on what we believe to be the key aspects of multi tenancy:

1. **The ability of the application to share hardware resources.**
2. **The offering of a high degree of configurability of the software.**
3. **The architectural approach in which the tenants make use of a single application and database instance.**

# Multi-tenants Deployment Modes for Application Server

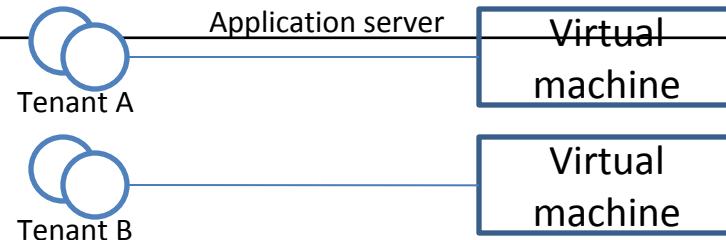
## Fully isolated Application server

Each tenant accesses an application server running on a dedicated servers.



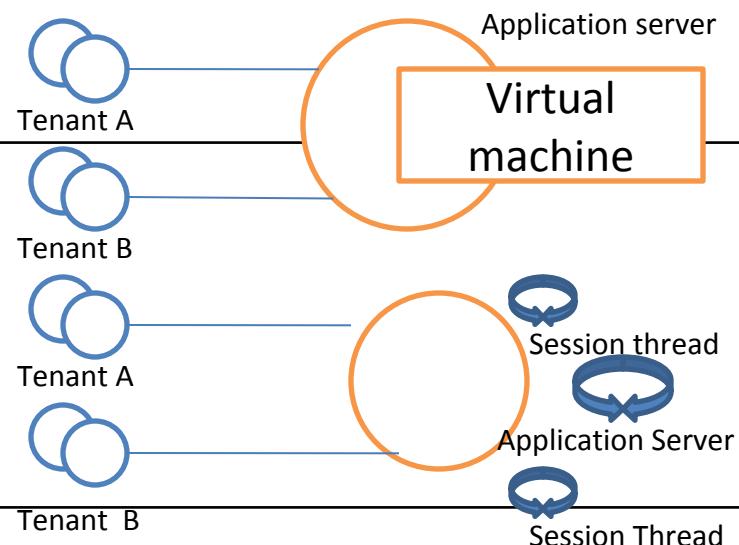
## Virtualized Application Server

Each tenant accesses a dedicated application running on a separate virtual machine.



## Shared Virtual Server

Each tenant accesses a dedicated application server running on a shared virtual machine.



## Shared Application Server

# Multi-tenancy Support

For the most part, multi-tenancy as discussed above appears to be of use primarily in a software as a service model. There are also certain cases where multi-tenancy can be useful within the enterprise as well. We have already seen that supporting multiple entities, such as bank branches, is essentially a multi-tenancy requirement. Similar needs can arise if a workgroup level application needs to be rolled out to many independent teams, who usually do not need to share data. Customizations of the application schema may also be needed in such scenarios, to support variations in business processes. Similar requirements also arise in supporting multiple *legal* entities each of which could be operating in different regulatory environments.

# Multi-tenancy using single schema

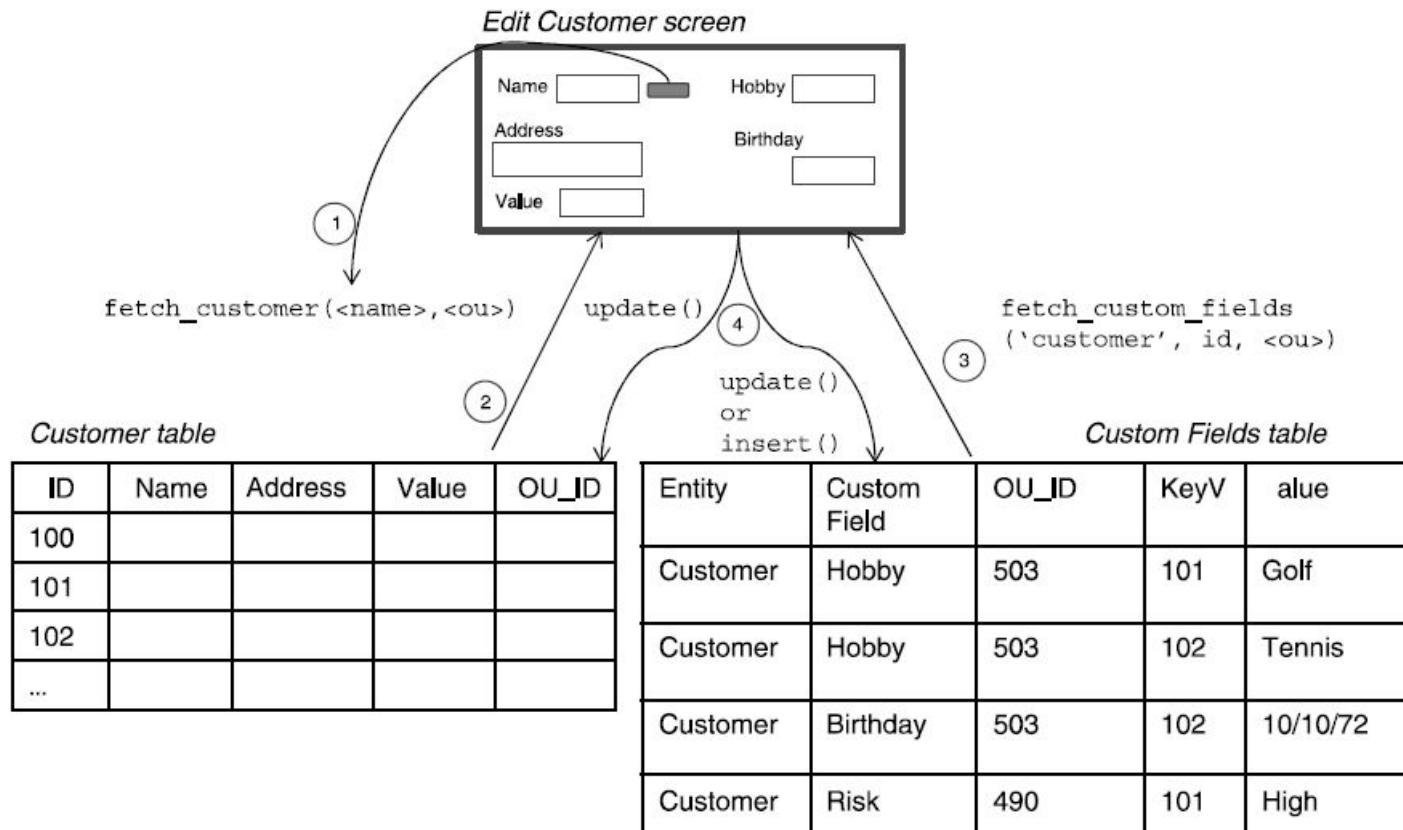


FIGURE 9.2. Multi-tenancy using a single schema

# Multi-tenancy using single schema

The above example is a simple case; more complex requirements also need to be handled, for example where a list of records is to be displayed with the ability to sort and filter on custom fields. It should be clear from this example that the single schema approach to multi-tenancy, while seemingly having the advantage of being able to upgrade the data model in one shot for all customers, has many complicating disadvantages in addition to the fact that major re-engineering of legacy applications is needed to move to this model.

# Multitenancy using Multiple Schema

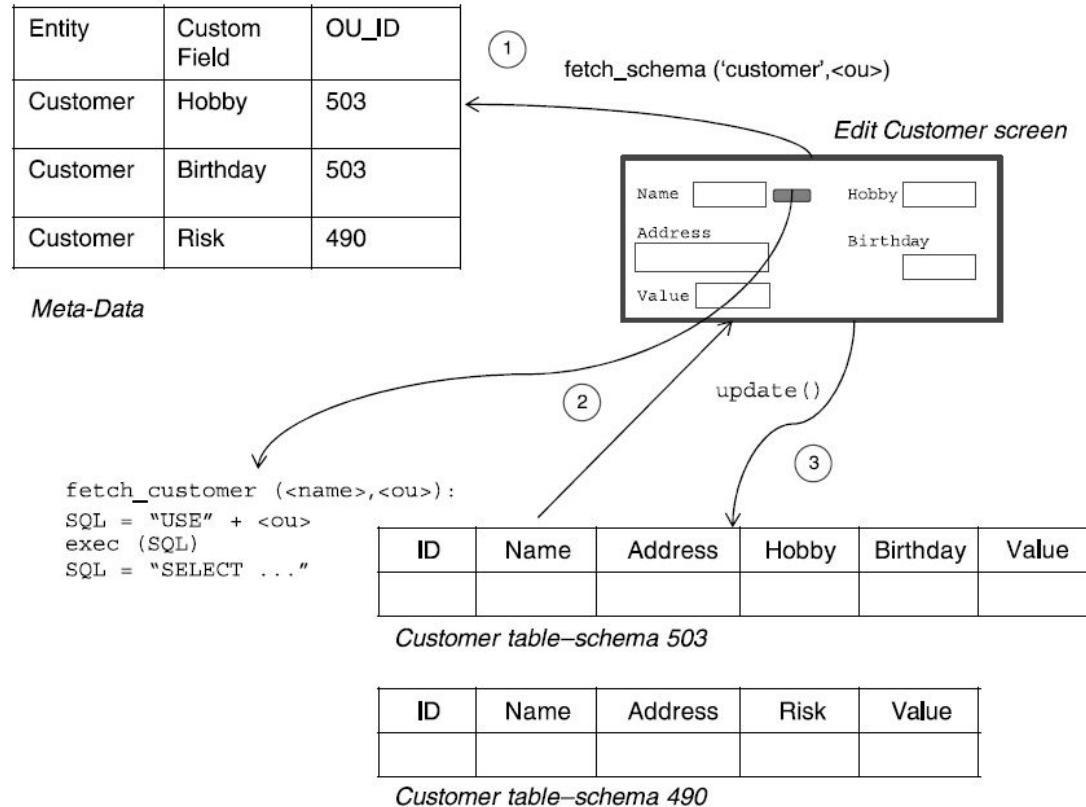


FIGURE 9.3. Multi-tenancy using multiple schemas

# Multi-tenancy using Multiple Schema

In the multiple schema approach a separate database schema is maintained for each customer, so each schema can implement customer-specific customizations directly. Meta-data describing customizations to the core schema is also maintained in a separate table, but unlike the Custom Fields table of Figure 9.2, this is pure meta-data and does not contain field values in individual records. As a result, the application design is simpler, and in case a legacy application needs to be re-engineered for multi-tenancy, it is likely that the modifications will be fewer and easier to accomplish.

# Data Intensive Computing and MapReduce

# Data Intensive Computing

*Data-intensive computing* is concerned with production, manipulation, and analysis of large-scale data in the range of hundreds of megabytes (MB) to petabytes (PB) and beyond [73]. The term *dataset* is commonly used to identify a collection of information elements that is relevant to one or more applications. Datasets are often maintained in *repositories*, which are infrastructures supporting the storage, retrieval, and indexing of large amounts of information. To facilitate the classification and search, relevant bits of information, called *metadata*, are attached to datasets.

Data-intensive computations occur in many application domains. Computational science is one of the most popular ones. People conducting scientific simulations and experiments are often keen to produce, analyze, and process huge volumes of data. Hundreds of gigabytes of data are produced every second by telescopes mapping the sky; the collection of images of the sky easily reaches the scale of petabytes over a year. Bioinformatics applications mine databases that may end up containing terabytes of data. Earthquake simulators process a massive amount of data, which is produced as a result of recording the vibrations of the Earth across the entire globe.

# Motivation

- Process lots of data
  - Google processed about **24 petabytes** of data per day in 2009.
- **A single machine** cannot serve all the data
  - You need a distributed system to store and process **in parallel**
- Parallel programming?
  - **Threading** is hard!
  - How do you facilitate **communication** between nodes?
  - How do you **scale to more machines**?
  - How do you handle machine **failures**?

# MapReduce

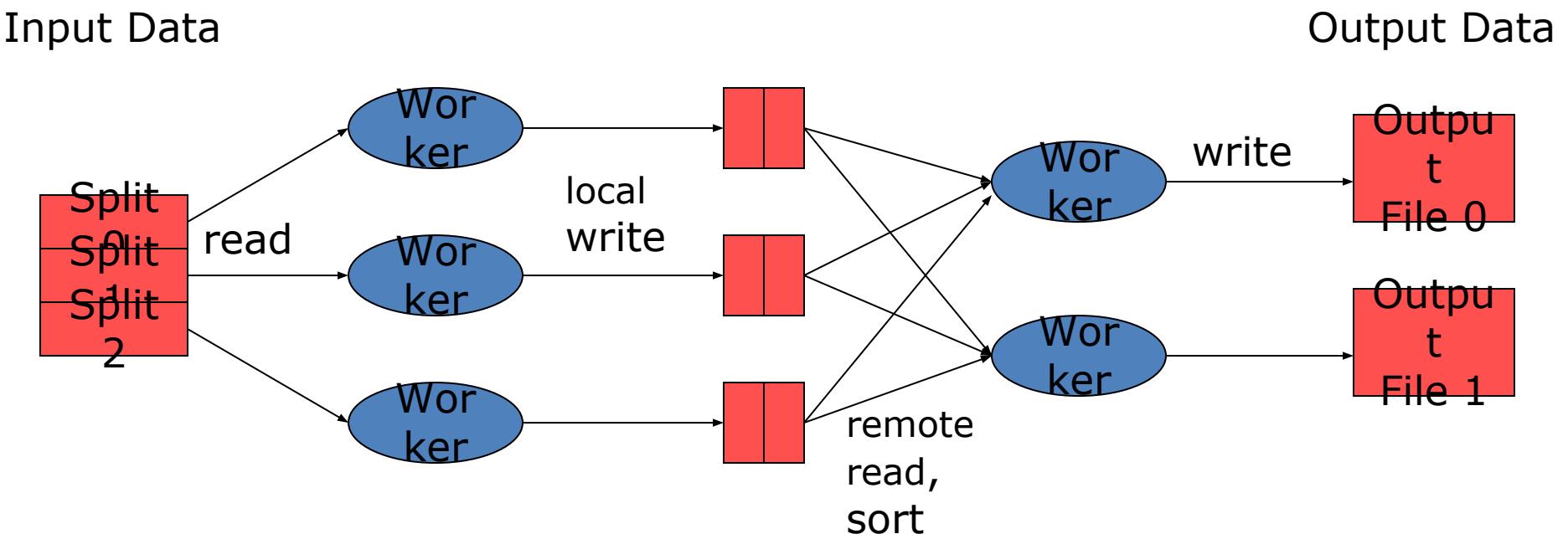
- MapReduce [OSDI'04] provides
  - Automatic parallelization, distribution
  - I/O scheduling
    - Load balancing
    - Network and data transfer optimization
  - Fault tolerance
    - Handling of machine failures
- Need more power: **Scale out, not up!**
  - Large number of **commodity servers** as opposed to some high end specialized servers

**Apache Hadoop:**  
Open source  
implementation of  
MapReduce

# Typical problem solved by MapReduce

- Read a lot of data
- **Map**: extract something you care about from each record
- Shuffle and Sort
- **Reduce**: aggregate, summarize, filter, or transform
- Write the results

# MapReduce workflow



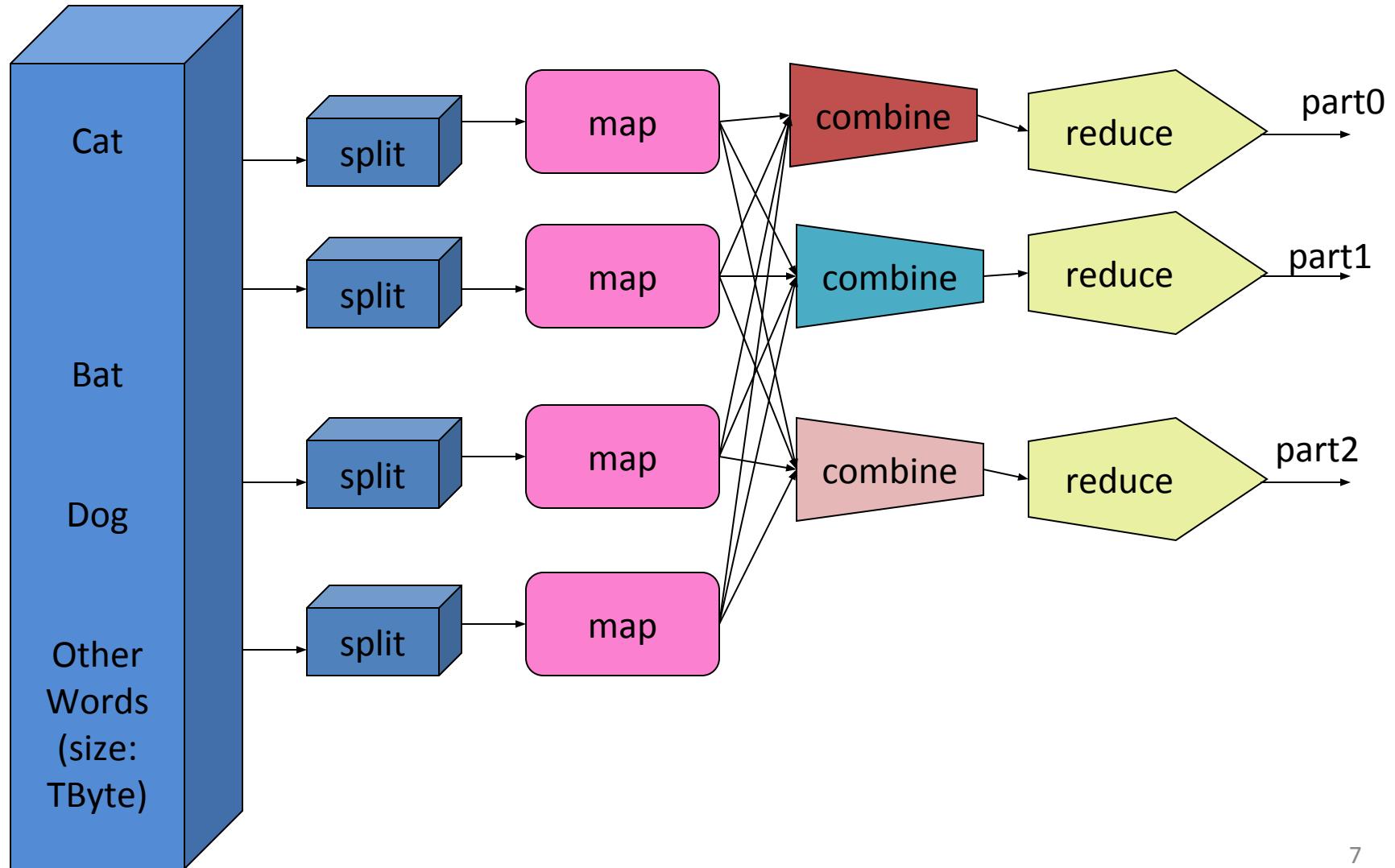
## Map

extract something you care about from each record

## Reduce

aggregate,  
summarize, filter,  
or transform

# MapReduce Example in operating systems command



# MapReduce programming model

- Determine if the problem is parallelizable and solvable using MapReduce (ex: Is the data WORM?, large data set).
- Design and implement solution as Mapper classes and Reducer class.
- Compile the source code with hadoop core.
- Package the code as jar executable.
- Configure the application (job) as to the number of mappers and reducers (tasks), input and output streams
- Load the data (or use it on previously available data)
- Launch the job and monitor.
- Study the result.
- Detailed steps.

# MapReduce Characteristics

- Very large scale data: peta, exa bytes
- Write once and read many data: allows for parallelism without mutexes
- Map and Reduce are the main operations: simple code
- There are other supporting operations such as combine and partition (out of the scope of this talk).
- All the map should be completed before reduce operation starts.
- Map and reduce operations are typically performed by the same physical processor.
- Number of map tasks and reduce tasks are configurable.
- Operations are provisioned near the data.
- Commodity hardware and storage.
- Runtime takes care of splitting and moving data for operations.
- Special distributed file system. Example: Hadoop Distributed File System and Hadoop Runtime.

# Classes of problems “mapreducible”

- Benchmark for comparing: Jim Gray’s challenge on data-intensive computing. Ex: “Sort”
- Google uses it (we think) for wordcount, adwords, pagerank, indexing data.
- Simple algorithms such as grep, text-indexing, reverse indexing
- Bayesian classification: data mining domain
- Facebook uses it for various operations: demographics
- Financial services use it for analytics
- Astronomy: Gaussian analysis for locating extra-terrestrial objects.
- Expected to play a critical role in semantic web and web3.0

# Mappers and Reducers

- Need to handle **more data**? Just add **more Mappers/Reducers**!
- No need to handle **multithreaded code** ☺
  - Mappers and Reducers are typically single threaded and **deterministic**
    - Determinism allows for **restarting** of failed jobs
  - Mappers/Reducers run **entirely independent** of each other
    - In Hadoop, they run in **separate JVMs**

# Example: Word Count

## Input Files

```
Apple Orange Mango  
Orange Grapes Plum
```

```
Apple Plum Mango  
Apple Apple Plum
```

# Mapper

- Reads in input pair  $\langle \text{Key}, \text{Value} \rangle$
- Outputs a pair  $\langle K', V' \rangle$ 
  - Let's count number of each word in user queries (or Tweets/Blogs)
  - The input to the mapper will be  $\langle \text{queryID}, \text{QueryText} \rangle$ :  

```
<Q1, "The teacher went to the store. The store was closed;  
the store opens in the morning. The store opens at 9am."  
>
```
  - The output would be:  

```
<The, 1> <teacher, 1> <went, 1> <to, 1> <the, 1>  
<store, 1> <the, 1> <store, 1> <was, 1> <closed, 1> <the,  
1> <store, 1> <opens, 1> <in, 1> <the, 1> <morning, 1>  
<the 1> <store, 1> <opens, 1> <at, 1> <9am, 1>
```

# Reducer

- Accepts the Mapper output, and aggregates values on the key

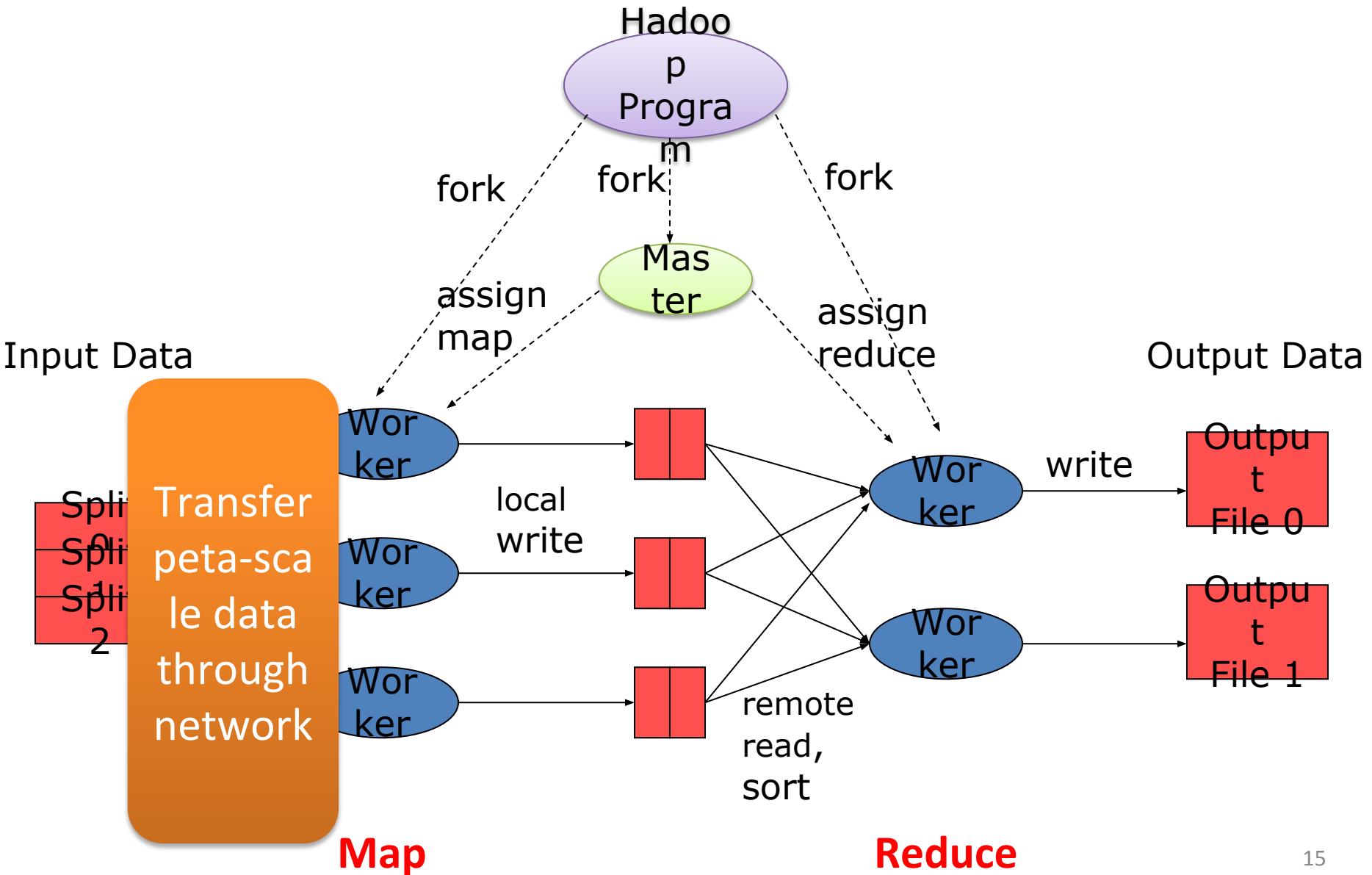
- For our example, the reducer input would be:

```
<The, 1> <teacher, 1> <went, 1> <to, 1> <the, 1> <store, 1>  
<the, 1> <store, 1> <was, 1> <closed, 1> <the, 1> <store, 1>  
<opens,1> <in, 1> <the, 1> <morning, 1> <the 1> <store, 1>  
<opens, 1> <at, 1> <9am, 1>
```

- The output would be:

```
<The, 6> <teacher, 1> <went, 1> <to, 1> <store, 3> <was, 1>  
<closed, 1> <opens, 1> <morning, 1> <at, 1> <9am, 1>
```

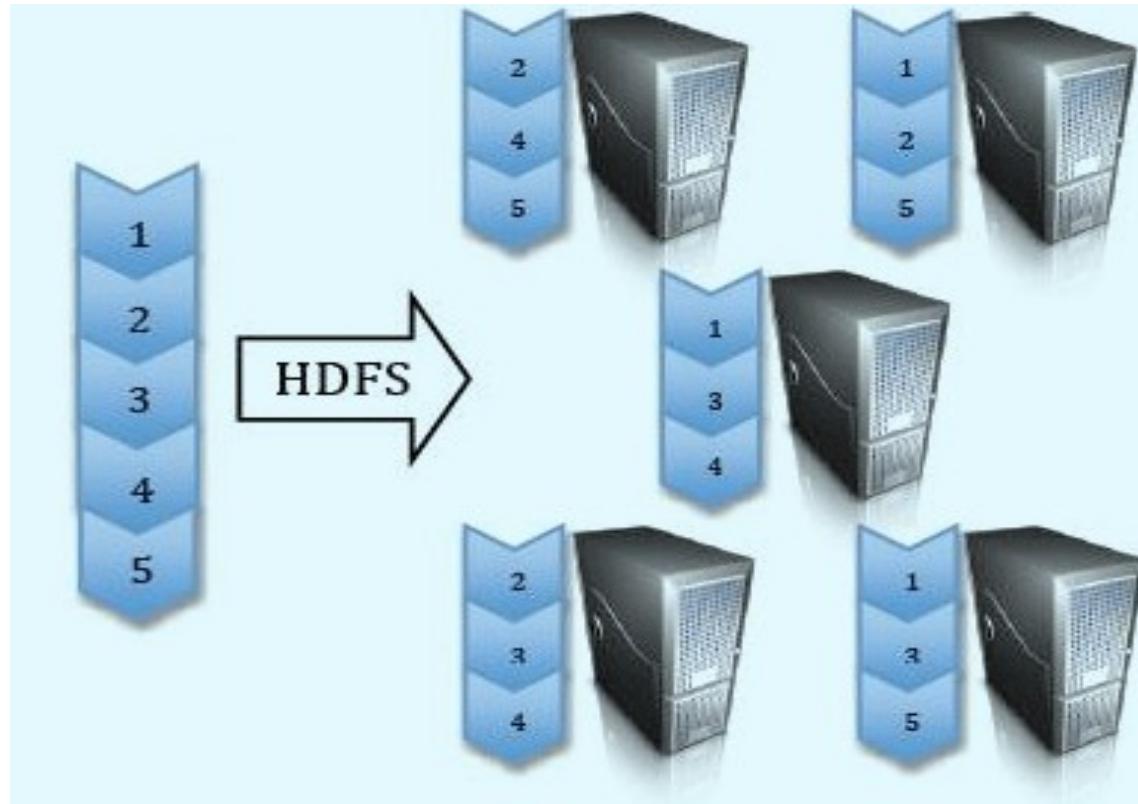
# MapReduce



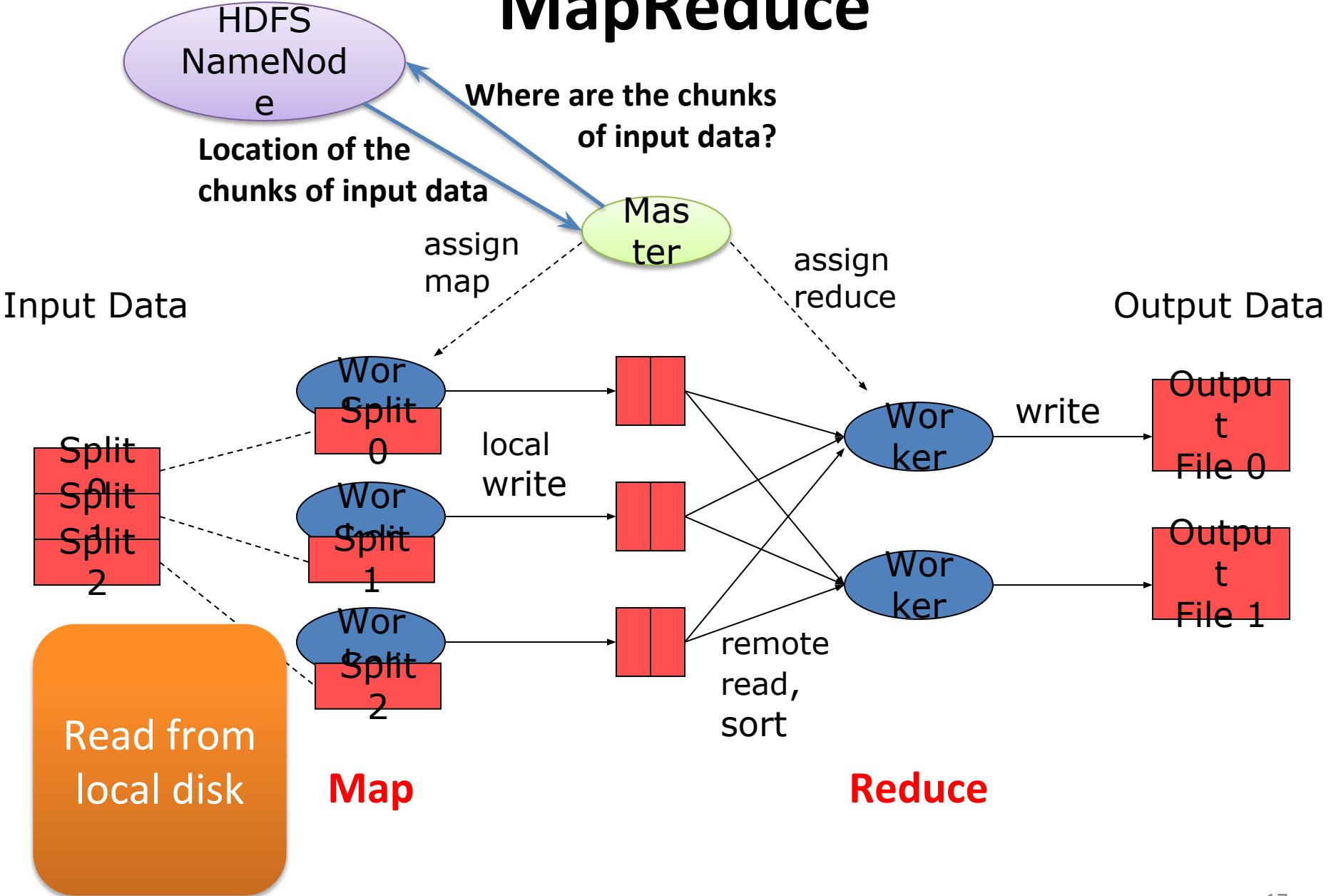
# Google File System (GFS)

# Hadoop Distributed File System (HDFS)

- Split data and store 3 replica on commodity servers



# MapReduce



# Locality Optimization

- **Master scheduling policy:**
  - Asks GFS for locations of replicas of input file blocks
  - Map tasks scheduled so GFS input block replica are on same machine or same rack
- Effect: Thousands of machines **read input at local disk speed**
  - Eliminate network bottleneck!

# Failure in MapReduce

- Failures are norm in commodity hardware
- Worker failure
  - Detect failure via periodic heartbeats
  - Re-execute in-progress map/reduce tasks
- Master failure
  - Single point of failure; Resume from Execution Log
- Robust
  - Google's experience: lost 1600 of 1800 machines once!, but finished fine

# Fault tolerance: Handled via re-execution

- On worker **failure**:
  - Detect failure via periodic heartbeats
  - Re-execute completed and in-progress *map* tasks
  - Task completion committed through master
- Robust: [Google's experience] lost 1600 of 1800 machines, but finished **fine**

# Refinement: Redundant Execution

- **Slow workers** significantly lengthen completion time
  - Other jobs consuming resources on machine
  - Bad disks with soft errors transfer data very slowly
  - Weird things: processor caches disabled (!!)
- **Solution:** spawn backup copies of tasks
  - Whichever one finishes first "wins"

# Refinement: Skipping Bad Records

Map/Reduce functions sometimes fail for particular inputs

- Best solution is to debug & fix, but not always possible
- If master sees **two failures** for the **same record**:
  - Next worker is told to **skip the record**

```

public class WordCount {

    public static class Map extends MapReduceBase implements
        Mapper<LongWritable, Text, Text, IntWritable> {
        private final static IntWritable one = new IntWritable(1);
        private Text word = new Text();

        public void map(LongWritable key, Text value, OutputCollector<Text, IntWritable>
            output, Reporter reporter) throws IOException {
            String line = value.toString();
            StringTokenizer tokenizer = new StringTokenizer(line);
            while (tokenizer.hasMoreTokens()) {
                word.set(tokenizer.nextToken());
                output.collect(word, one);
            }
        }
    }
}

```

## Mapper

```

public static class Reduce extends MapReduceBase implements
    Reducer<Text, IntWritable, Text, IntWritable> {
    public void reduce(Text key, Iterator<IntWritable> values, OutputCollector<Text,
        IntWritable> output, Reporter reporter) throws IOException {
        int sum = 0;
        while (values.hasNext()) { sum += values.next().get(); }
        output.collect(key, new IntWritable(sum));
    }
}

```

## Reducer

```

public static void main(String[] args) throws Exception {
    JobConf conf = new JobConf(WordCount.class);
    conf.setJobName("wordcount");
    conf.setOutputKeyClass(Text.class);
    conf.setOutputValueClass(IntWritable.class);
    conf.setMapperClass(Map.class);
    conf.setCombinerClass(Reduce.class);
    conf.setReducerClass(Reduce.class);
    conf.setInputFormat(TextInputFormat.class);
    conf.setOutputFormat(TextOutputFormat.class);
    FileInputFormat.setInputPaths(conf, new Path(args[0]));
    FileOutputFormat.setOutputPath(conf, new Path(args[1]));

    JobClient.runJob(conf);
}

```

Run this program as  
a MapReduce job



```
public class WordCount {  
    public static class Map extends Mapper<LongWritable, Text, Text, IntWritable> {  
        private final static IntWritable one = new IntWritable(1);  
        private Text word = new Text();  
  
        public void map(LongWritable key, Text value, Context context) throws IOException, InterruptedException {  
            String line = value.toString();  
            StringTokenizer tokenizer = new StringTokenizer(line);  
            while (tokenizer.hasMoreTokens()) {  
                word.set(tokenizer.nextToken());  
                context.write(word, one);  
            }  
        }  
    }  
}
```

## Mapper

```
public static class Reduce extends Reducer<Text, IntWritable, Text, IntWritable> {  
  
    public void reduce(Text key, Iterable<IntWritable> values, Context context)  
        throws IOException, InterruptedException {  
        int sum = 0;  
        for (IntWritable val : values) {  
            sum += val.get();  
        }  
        context.write(key, new IntWritable(sum));  
    }  
}
```

## Reducer

```
public static void main(String[] args) throws Exception {  
    Configuration conf = new Configuration();  
  
    Job job = new Job(conf, "wordcount");  
  
    job.setOutputKeyClass(Text.class);  
    job.setOutputValueClass(IntWritable.class);  
  
    job.setMapperClass(Map.class);  
    job.setReducerClass(Reduce.class);  
  
    job.setInputFormatClass(TextInputFormat.class);  
    job.setOutputFormatClass(TextOutputFormat.class);  
  
    FileInputFormat.addInputPath(job, new Path(args[0]));  
    FileOutputFormat.setOutputPath(job, new Path(args[1]));  
  
    job.waitForCompletion(true);  
}
```

Run this program as  
a MapReduce job

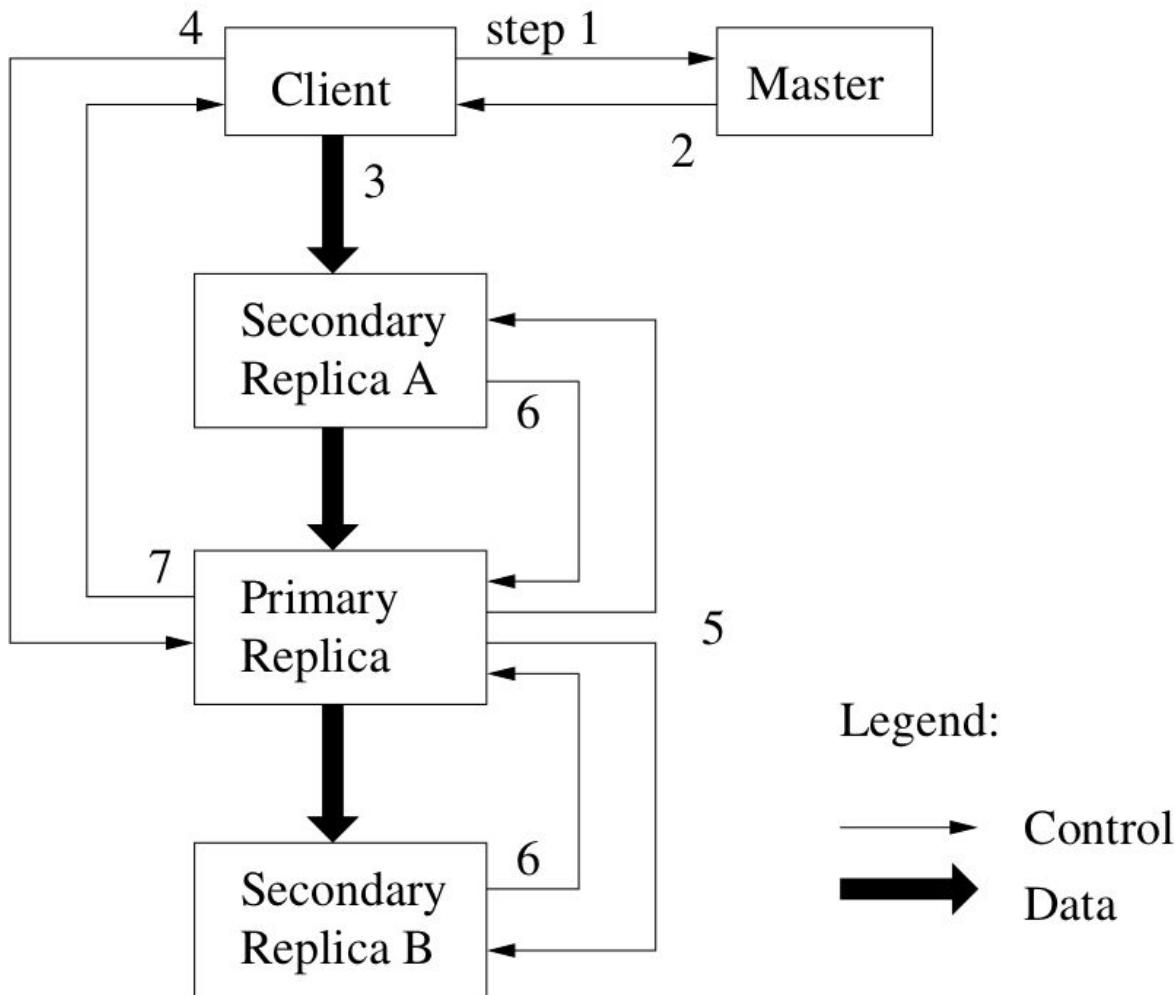
# Summary

- MapReduce
  - Programming paradigm for data-intensive computing
  - Distributed & parallel execution model
  - Simple to program
    - The framework automates many tedious tasks (machine selection, failure handling, etc.)

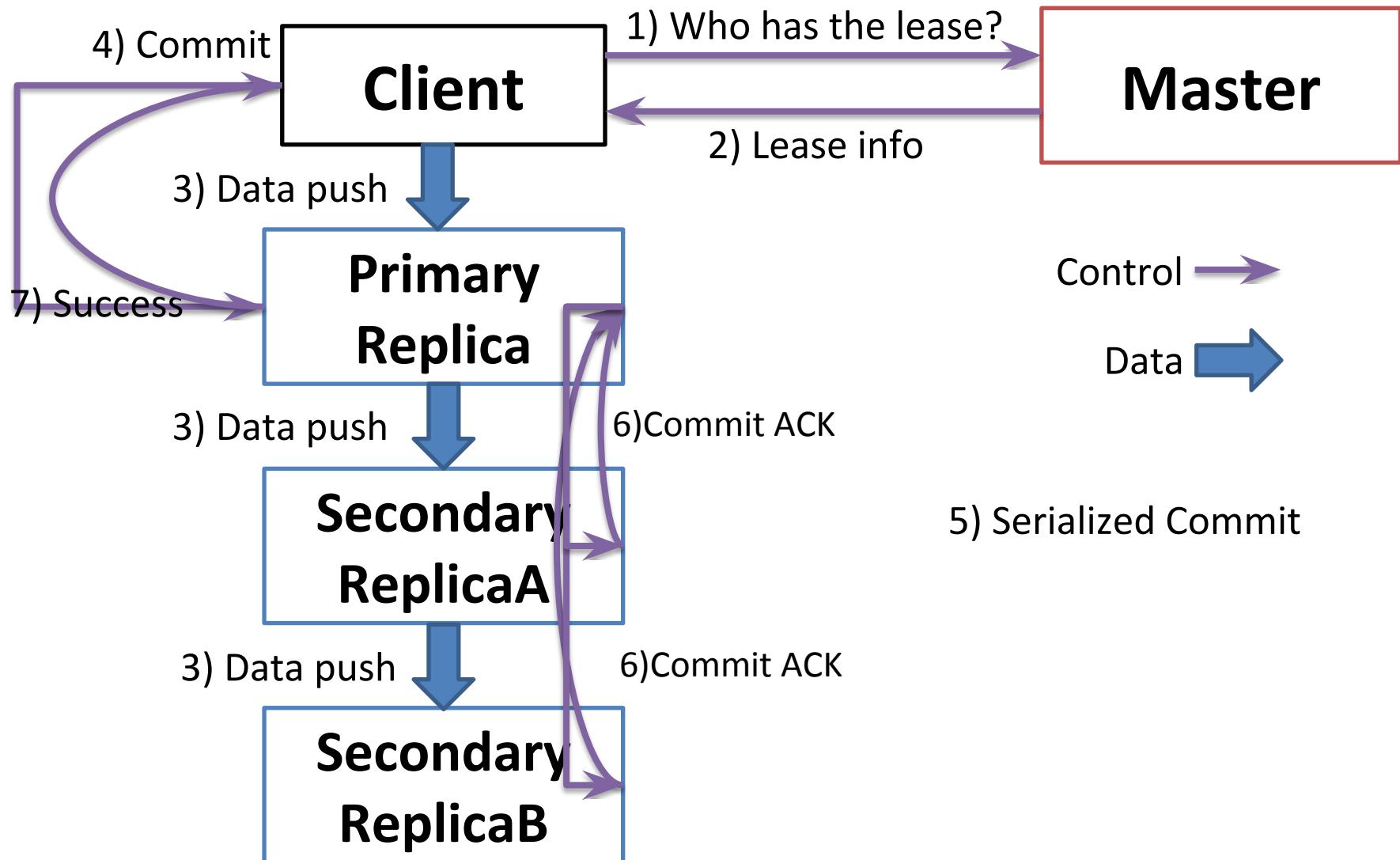
# HDFS: Large Scale Data Storage

- Manipulate large (**Peta Scale**) sets of data
- Large number of machine with **commodity hardware**
- Component failure is the norm
- Goal: **Scalable, high performance, fault tolerant**  
distributed file system

# Write operation



# Write(filename, offset, data)



# RecordAppend (2)

- Record size is limited by chunk size
- When a record does not fit into available space,
  - chunk is padded to end
  - and client retries request.

# Fault tolerance

- Replication
  - High availability for reads
  - User controllable, default 3 (non-RAID)
  - Provides read/seek bandwidth
  - Master is responsible for directing re-replication if a data node dies
- Online checksumming in data nodes
  - Verified on reads

# Replica Management

- Bias towards **topological** spreading
  - Rack, data center
- **Rebalancing**
  - Move chunks around to balance disk fullness
  - Gently fixes imbalances due to:
    - Adding/removing data nodes

# Replica Management (Cloning)

- Chunk replica lost or corrupt
- Goal: minimize app disruption and data loss
  - Approximately in priority order
    - More replica missing-> priority boost
    - Deleted file-> priority decrease
    - Client blocking on a write-> large priority boost
  - Master directs copying of data
- Performance on a production cluster
  - Single failure, full recovery (600GB): 23.2 min
  - Double failure, restored 2x replication: 2min

# Limitations

- Master is a central point of failure
- Master can be a scalability bottleneck
- Latency when opening/statting thousands of files
- Security model is weak

# Google Bigtable

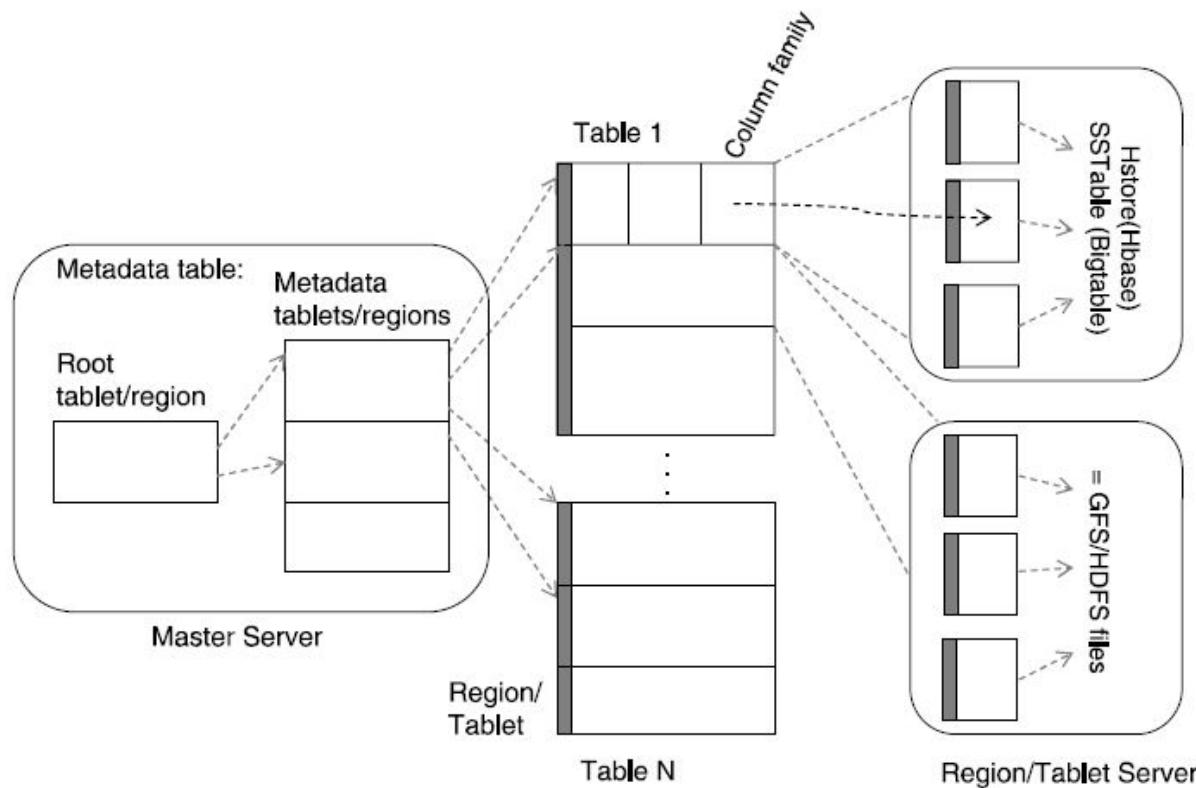


FIGURE 10.5. Google BigTable/Hadoop HDFS

# BigTable

BigTable [9] is a distributed *structured* storage system built on GFS; Hadoop's HBase is a similar open source system that uses HDFS. A BigTable is essentially a sparse, distributed, persistent, multidimensional sorted 'map.'<sup>1</sup> Data in a BigTable is accessed by a row key, column key and a timestamp. Each column can store arbitrary name–value pairs of the form *column-family:label, string*. The set of possible column-families for a table is fixed when it is created whereas columns, i.e. labels within the column family, can be created dynamically at any time. Column families are stored close together in the distributed file system; thus the BigTable model shares elements of column-oriented databases. Further, each Bigtable cell (row, column) can contain multiple versions of the data that are stored in decreasing timestamp order.

# Hbase

BigTable and HBase rely on the underlying distributed file systems GFS and HDFS respectively and therefore also inherit some of the properties of these systems. In particular large parallel reads and inserts are efficiently supported, even simultaneously on the same table, unlike a traditional relational database. In particular, reading all rows for a small number of column families from a large table, such as in aggregation queries, is efficient in a manner similar to column-oriented databases. Random writes translate to data inserts since multiple versions of each cell are maintained, but are less efficient since cell versions are stored in descending order and such inserts require more work than simple file appends. Similarly, the consistency properties of large parallel inserts are stronger than that for parallel random writes, as is pointed out in [26]. Further, writes can even fail if a few replicas are unable to write even if other replicas are successfully updated.

# Amazon Dynamo

We now turn to another distributed data system called Dynamo, which was developed at Amazon and underlies its SimpleDB key-value pair database. Unlike BigTable, Dynamo was designed specifically for supporting a large volume of concurrent updates, each of which could be small in size, rather than bulk reads and appends as in the case of BigTable and GFS.

Dynamo's data model is that of simple key-value pairs, and it is expected that applications read and write such data objects fairly randomly. This model is well suited for many web-based e-commerce applications that all need to support constructs such as a 'shopping cart.'

Dynamo also replicates data for fault tolerance, but uses distributed object versioning and quorum-consistency to enable writes to succeed without waiting for all replicas to be successfully updated, unlike in the case of GFS. Managing conflicts if they arise is relegated to reads which are provided enough information to enable application dependent resolution. Because of these features, Dynamo does not rely on any underlying distributed file system and instead directly manages data storage across distributed nodes.

The architecture of Dynamo is illustrated in Figure 10.6. Objects are key-value pairs with arbitrary arrays of bytes. An MD5 hash of the key is used to generate a 128-bit hash value. The range of this hash function is mapped

# Amazon Dynamo

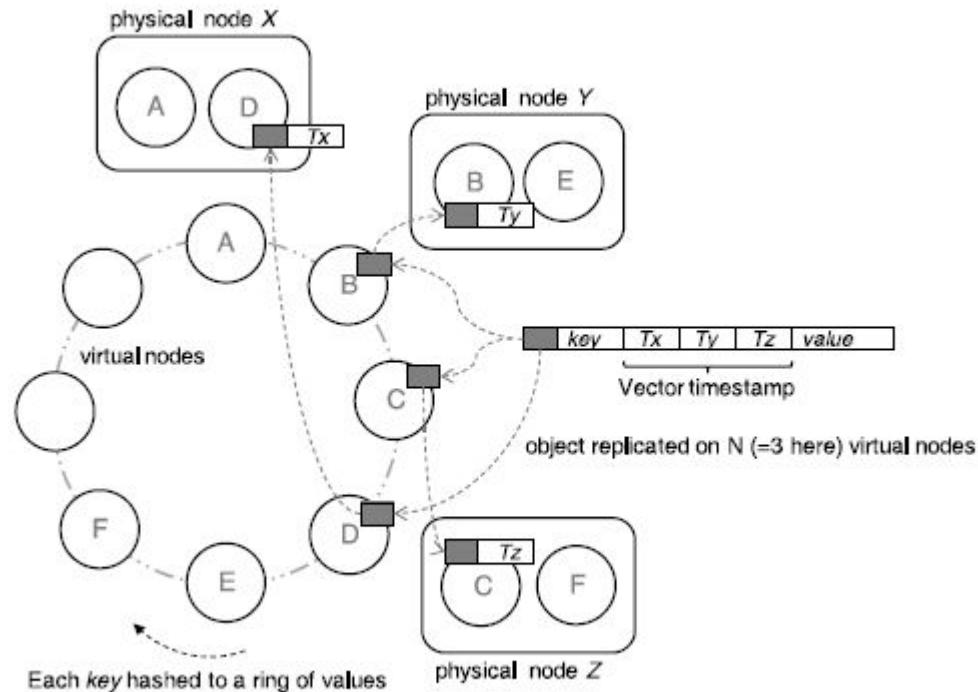


FIGURE 10.6. Amazon Dynamo

# Conclusion

- Inexpensive commodity components can be the basis of a large scale reliable system
- Adjusting the API, e.g. RecordAppend, can enable large distributed apps
- Fault tolerant
- Useful for many similar apps



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

## Static website

nginx 1.5 + modsecurity + openssl + bootstrap 2

## Background workers

Python 3.0 + celery + pyredis + libcurl + ffmpeg + libopencv + nodejs + phantomjs

## User DB

postgresql + pgv8 + v8

## Web frontend

Ruby + Rails + sass + Unicorn

## Analytics DB

hadoop + hive + thrift + OpenJDK

## Queue

Redis + redis-sentinel

## API endpoint

Python 2.7 + Flask + pyredis + celery + psycopg + postgresql-client



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

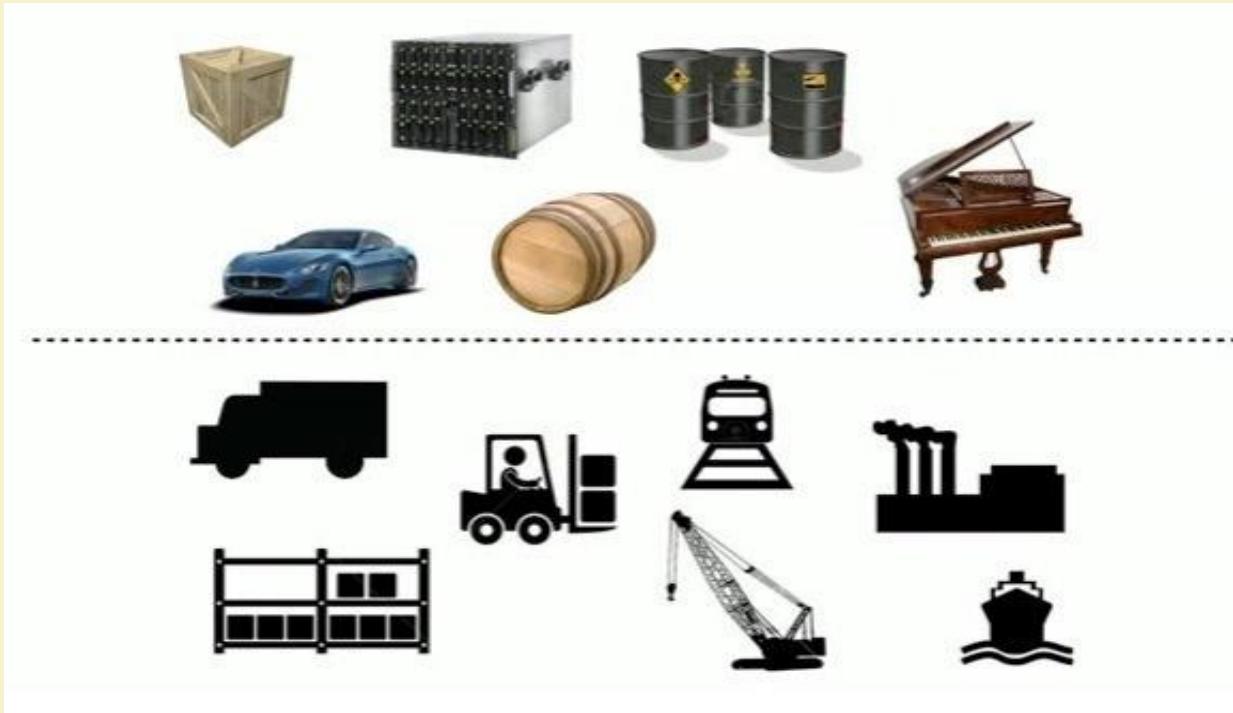


IIT KHARAGPUR



NPTEL  
NPTEL ONLINE  
CERTIFICATION COURSES

# “Shipping”



Ref: Internet/YouTube



IIT KHARAGPUR



NPTEL  
NPTEL ONLINE  
CERTIFICATION COURSES

# “Shipping”



Ref: Internet/YouTube

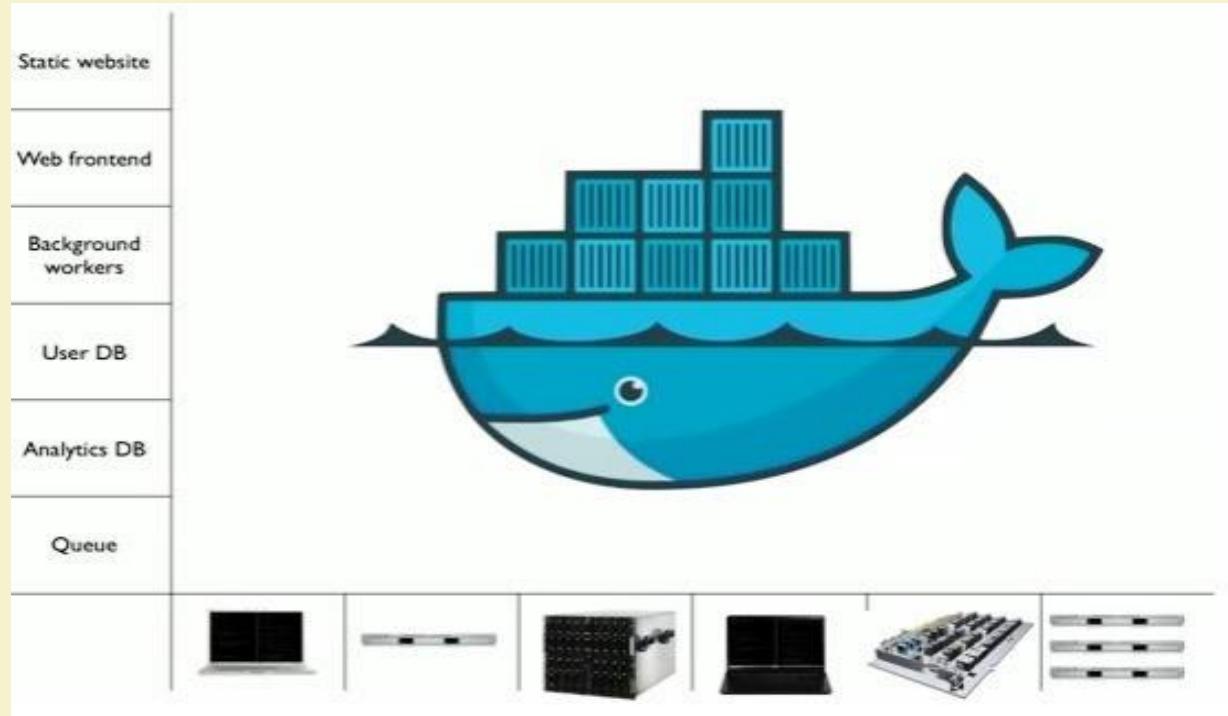


IIT KHARAGPUR



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/>



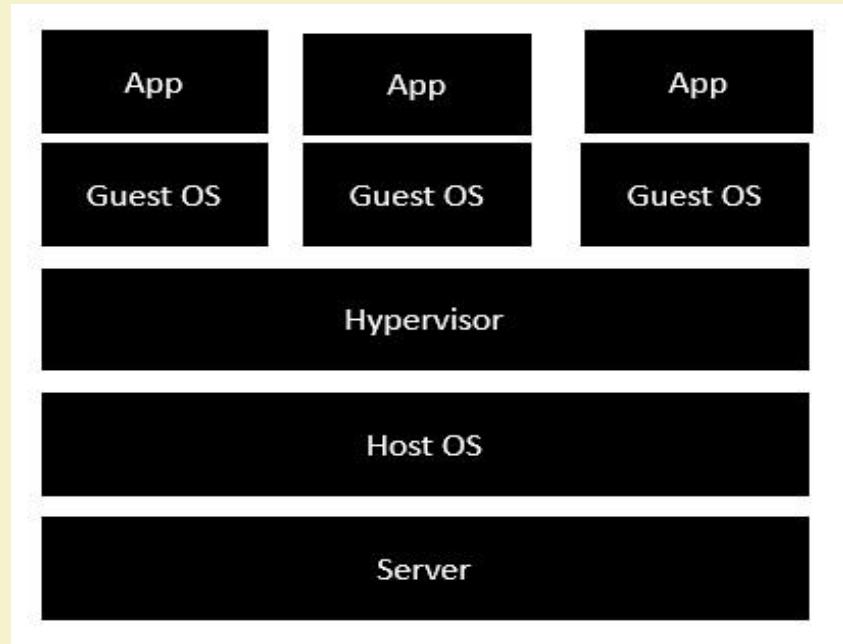
IIT KHARAGPUR



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/>



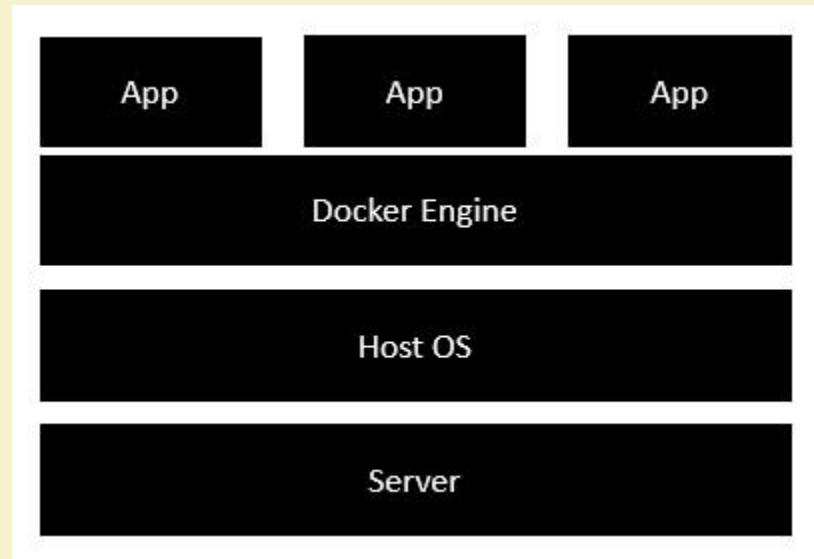
IIT KHARAGPUR



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



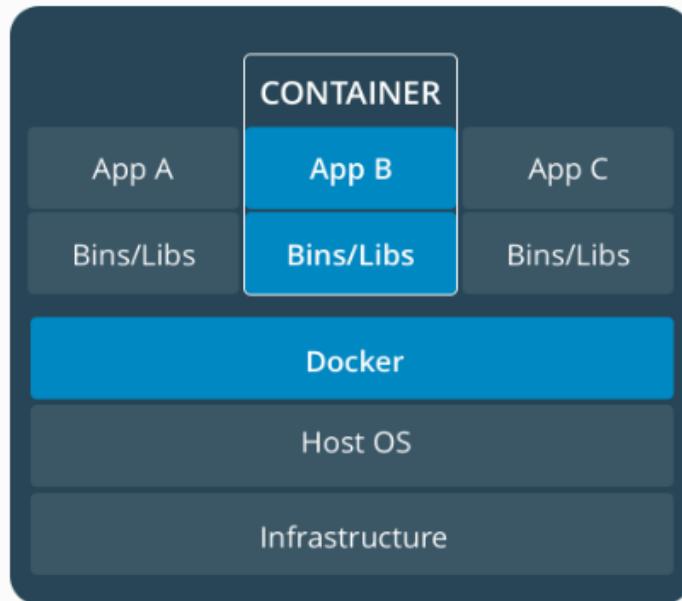
NPTEL  
NPTEL ONLINE  
CERTIFICATION COURSES

# Container (contd...)

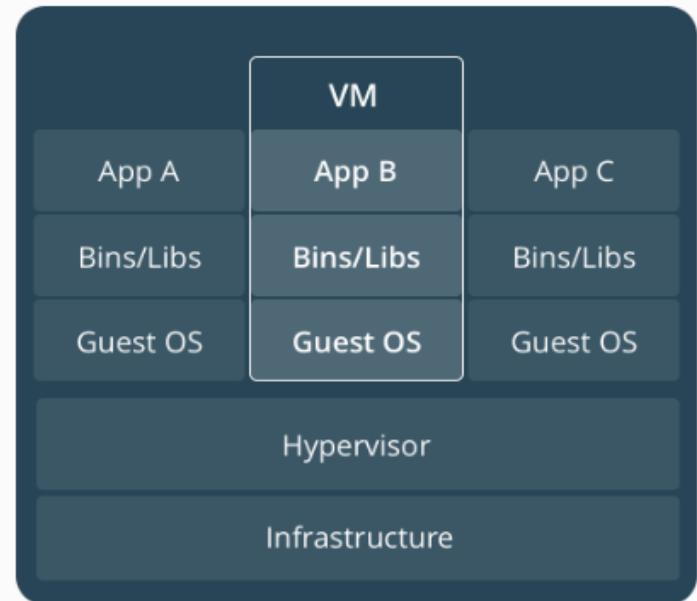
- An ***image*** is a lightweight, stand-alone, executable package that includes everything needed to run a piece of software, including the code, a runtime, libraries, environment variables, and config files.
- A ***container*** is a runtime instance of an image—what the image becomes in memory when 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/>

# 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/>

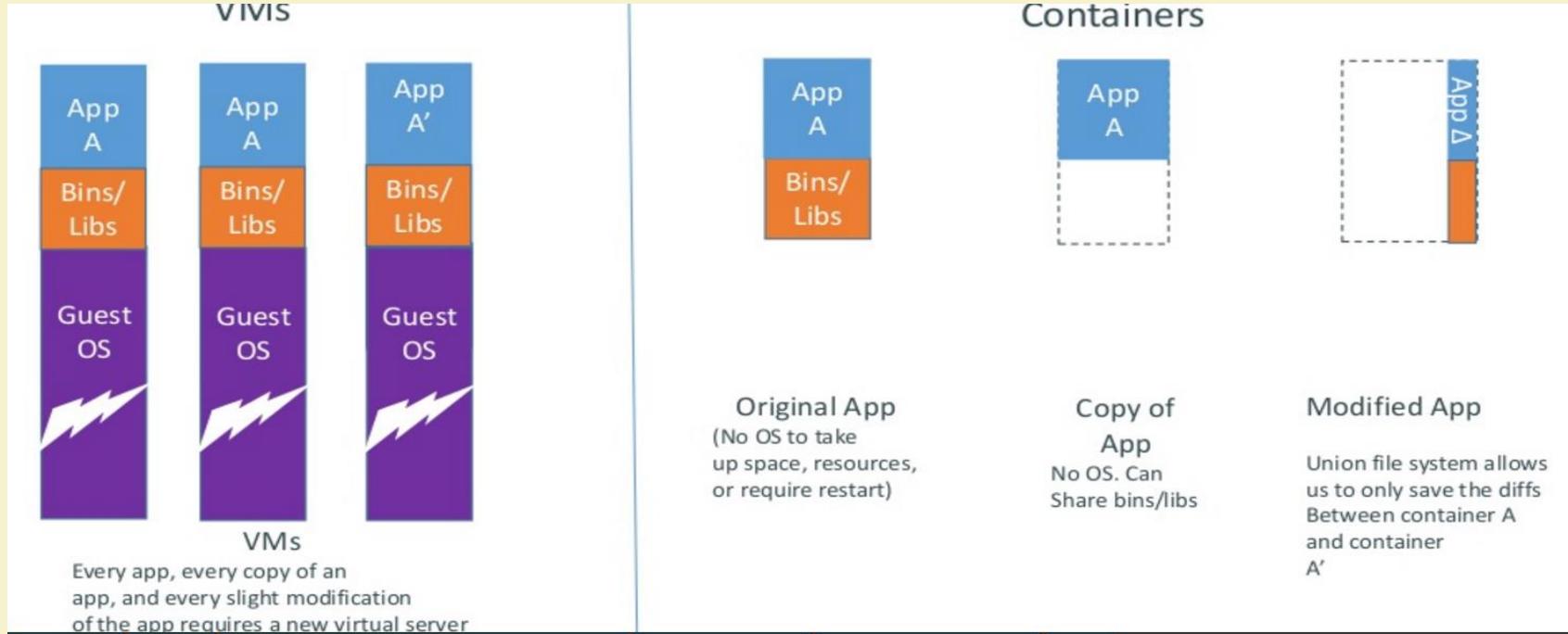


IIT KHARAGPUR



NPTEL  
NPTEL ONLINE  
CERTIFICATION COURSES

# Docker containers are lightweight

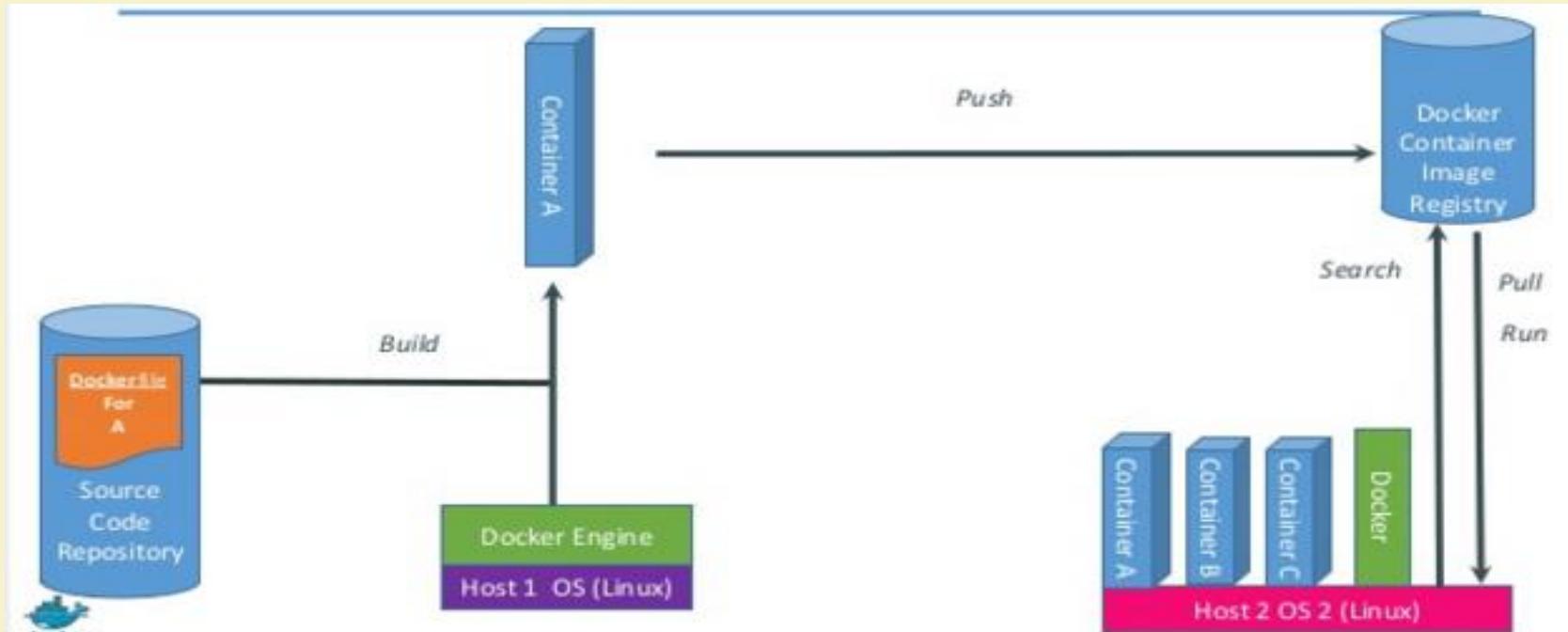


IIT KHARAGPUR



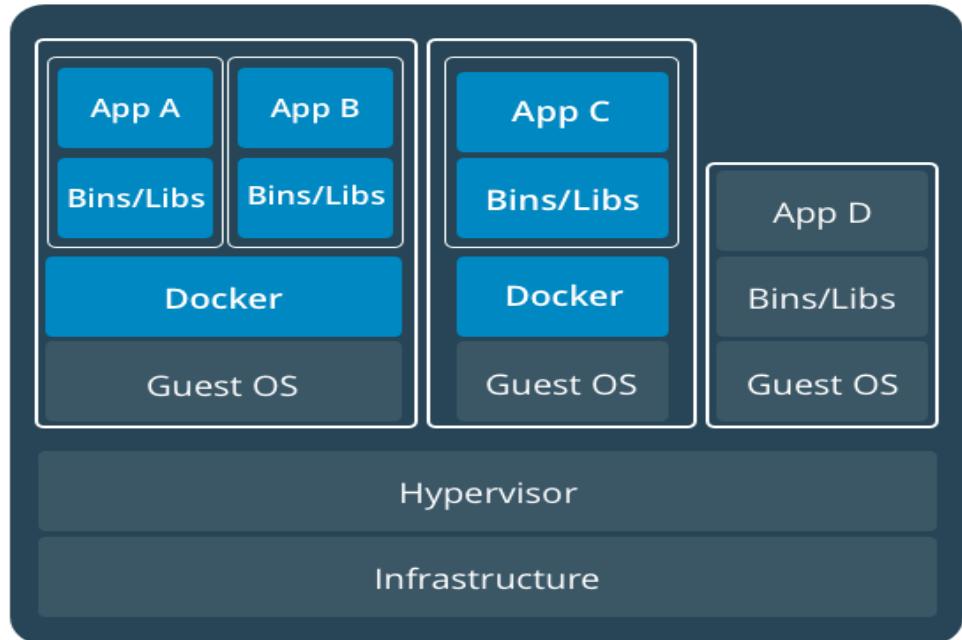
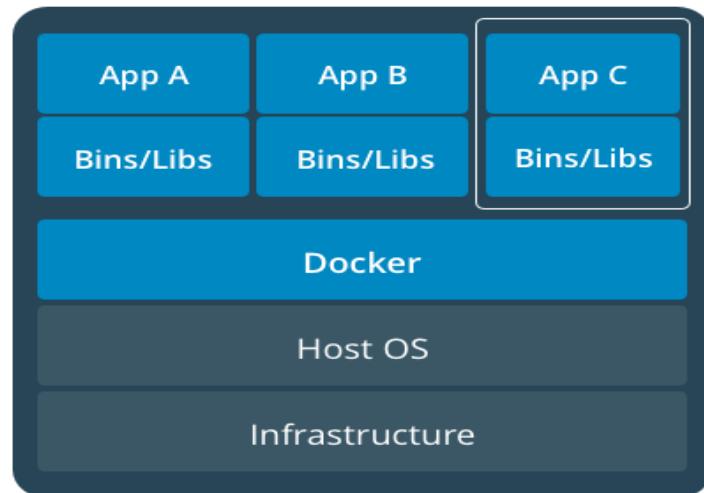
NPTEL  
NPTEL ONLINE  
CERTIFICATION COURSES

# 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!



IIT KHARAGPUR



NPTEL ONLINE  
CERTIFICATION COURSES



IIT KHARAGPUR



NPTEL ONLINE  
CERTIFICATION COURSES

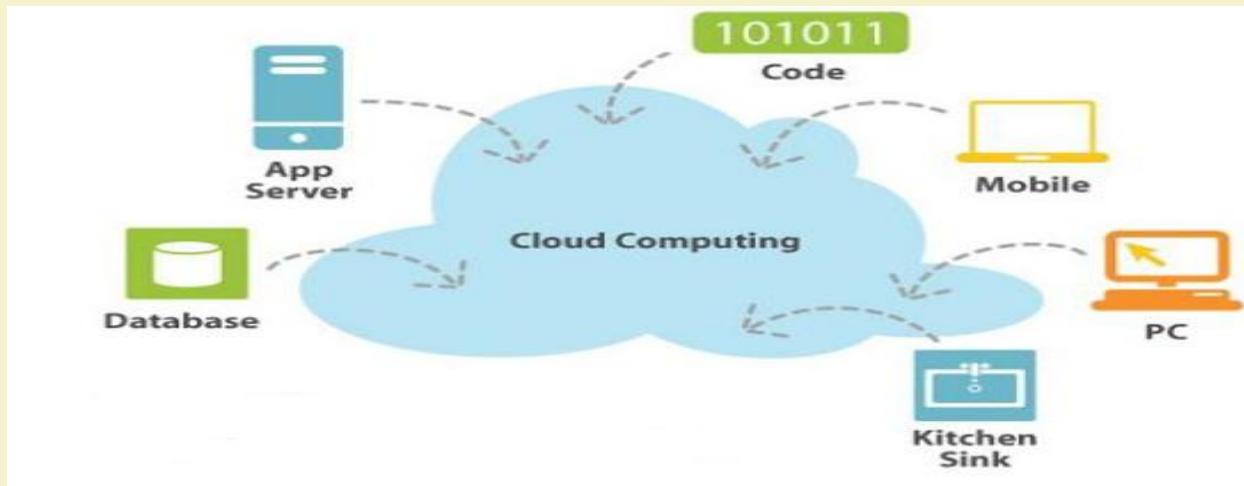
# CLOUD COMPUTING

## 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.



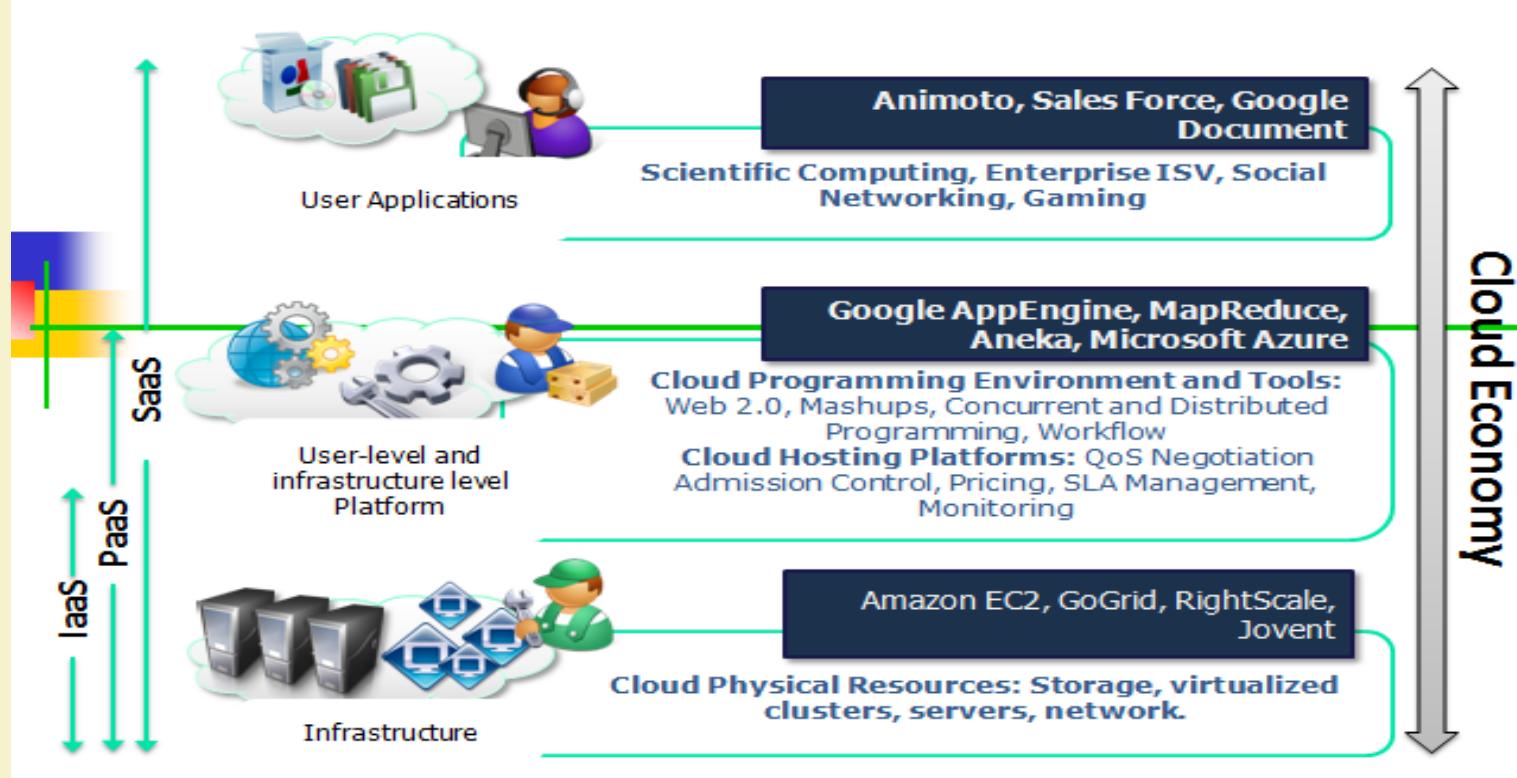
Source: Internet



IIT KHARAGPUR



NPTEL  
NPTEL ONLINE  
CERTIFICATION COURSES



Source: Internet



IIT KHARAGPUR



NPTEL  
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  
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

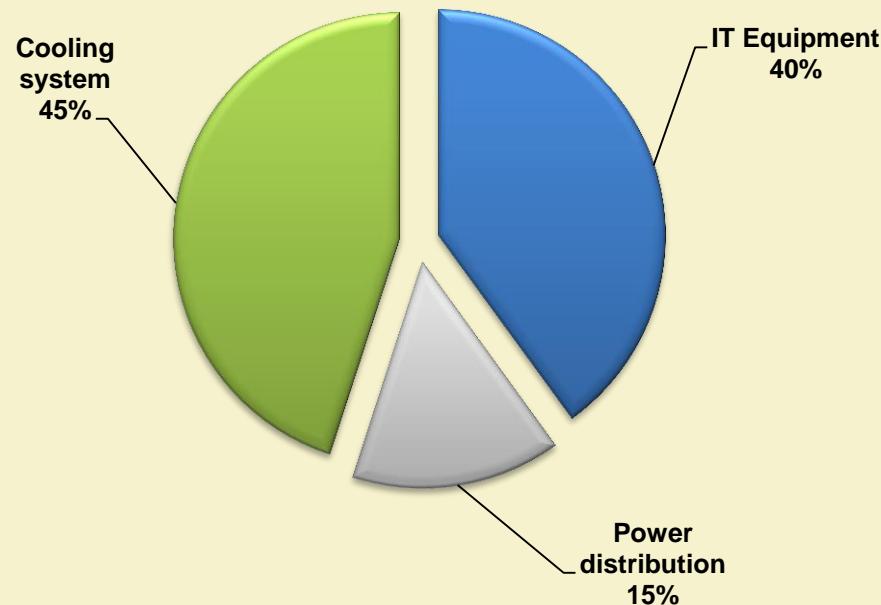


IIT KHARAGPUR



NPTEL  
NPTEL ONLINE  
CERTIFICATION COURSES

# Typical Data Center Energy Consumption

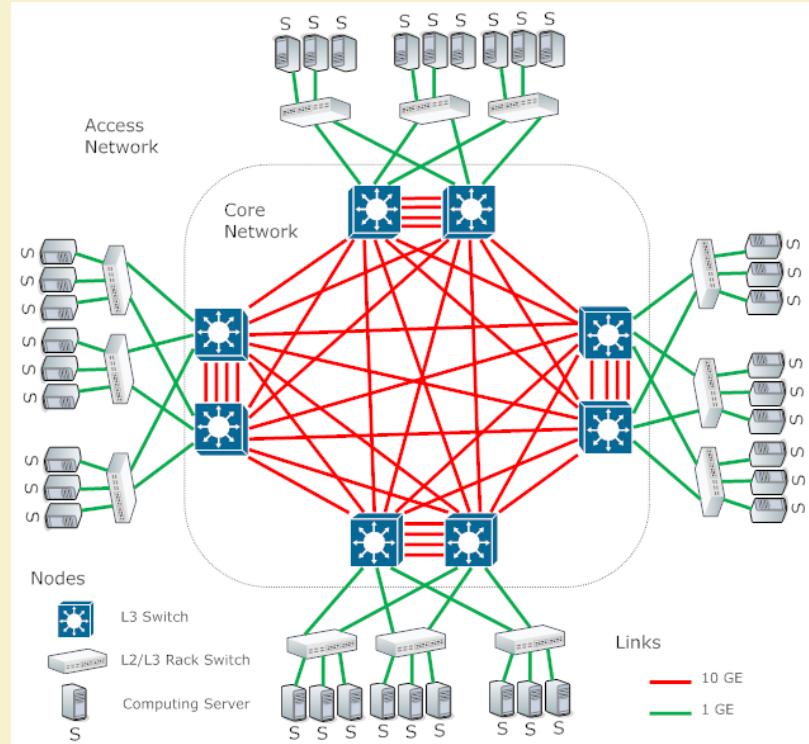


Ref: Dzmitry Kliazovich, University of Luxembourg

# 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

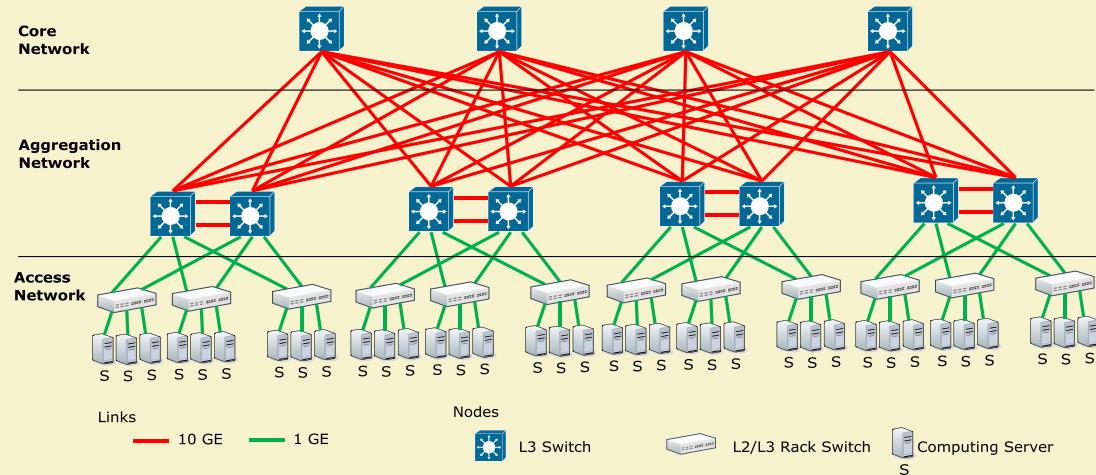


Ref: Dzmitry Kliazovich, University of Luxembourg

# DC Architecture - Present

Three-tier DC architecture

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

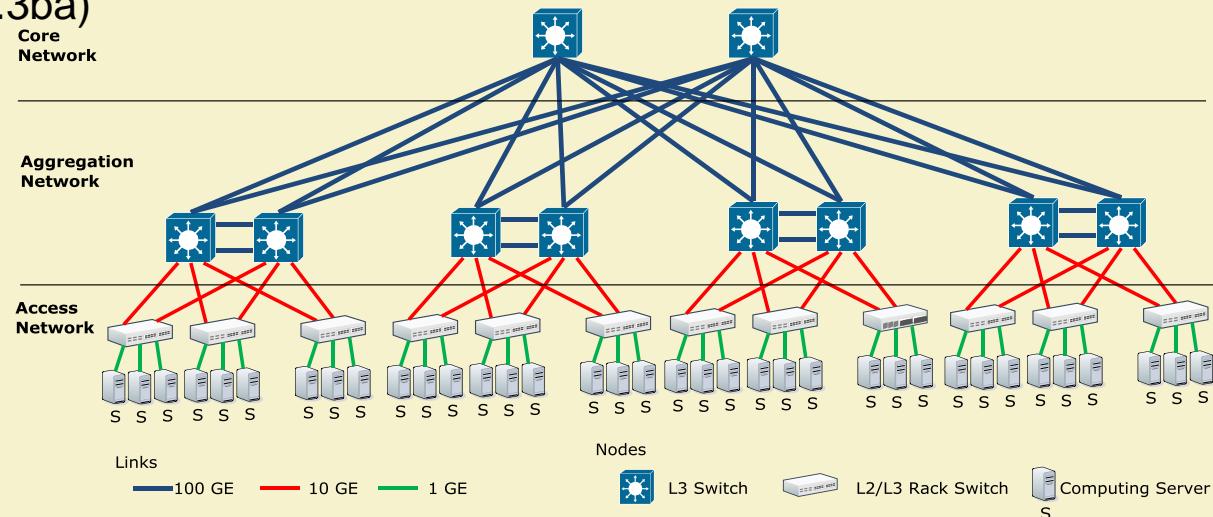


Ref: Dzmitry Kliazovich, University of Luxembourg

# DC Architecture - Present

## Three-tier High-Speed architecture

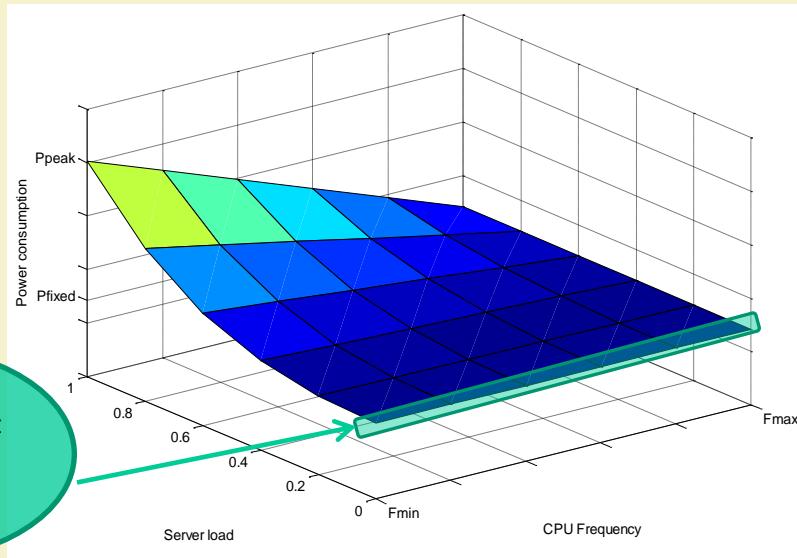
- Increased core network bandwidth
- 2-way ECMP load balancing
- 100 GE standard (IEEE 802.3ba)



Ref: Dzmitry Kliazovich, University of Luxembourg

# DC Server Energy Model

Idle server consumes about 66% of the peak load for all CPU frequencies



The diagram shows a server tower with two callout arrows pointing to its components. One arrow points to the lower half of the server, labeled "memory modules, disks, I/O resources". The other arrow points to the top half, labeled "CPU".

$$P = P_{fixed} + P_f * f^3$$

Ref: Dzmitry Kliazovich, University of Luxembourg

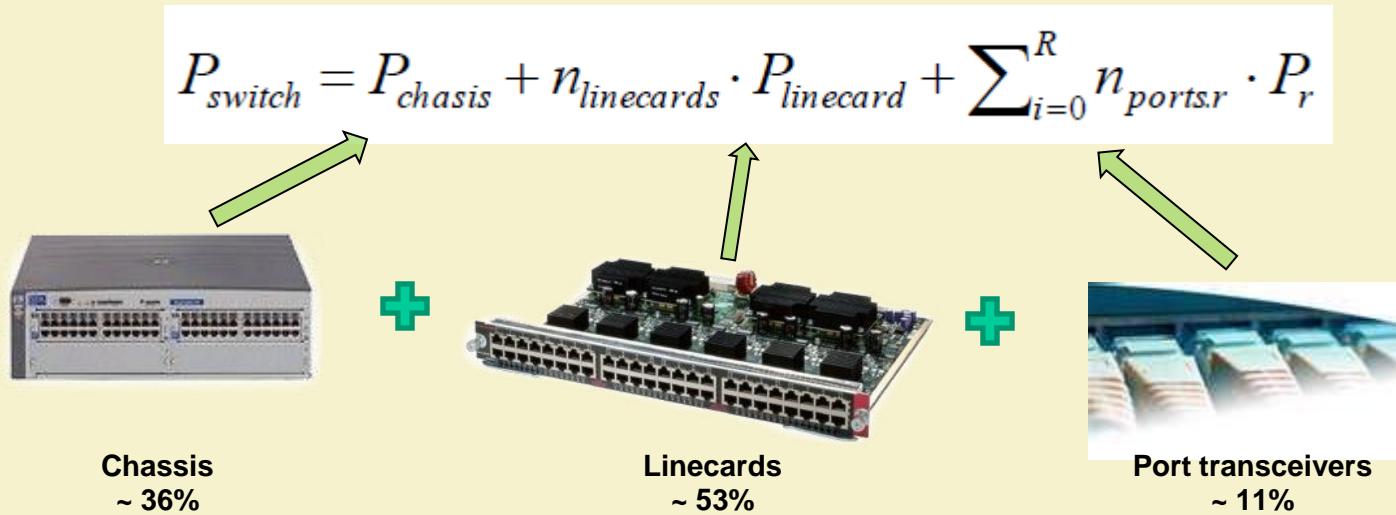


IIT KHARAGPUR



NPTEL  
NPTEL ONLINE  
CERTIFICATION COURSES

# DC Network Switches' Energy Model



Ref: Dzmitry Kliazovich, University of Luxembourg

# 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



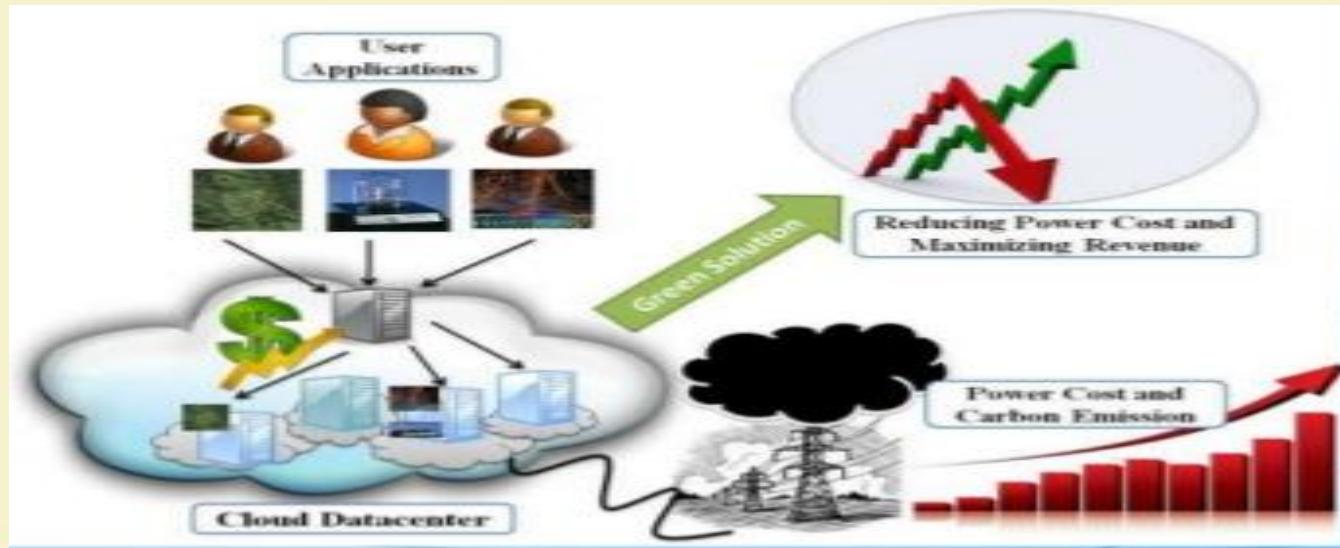
IIT KHARAGPUR



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

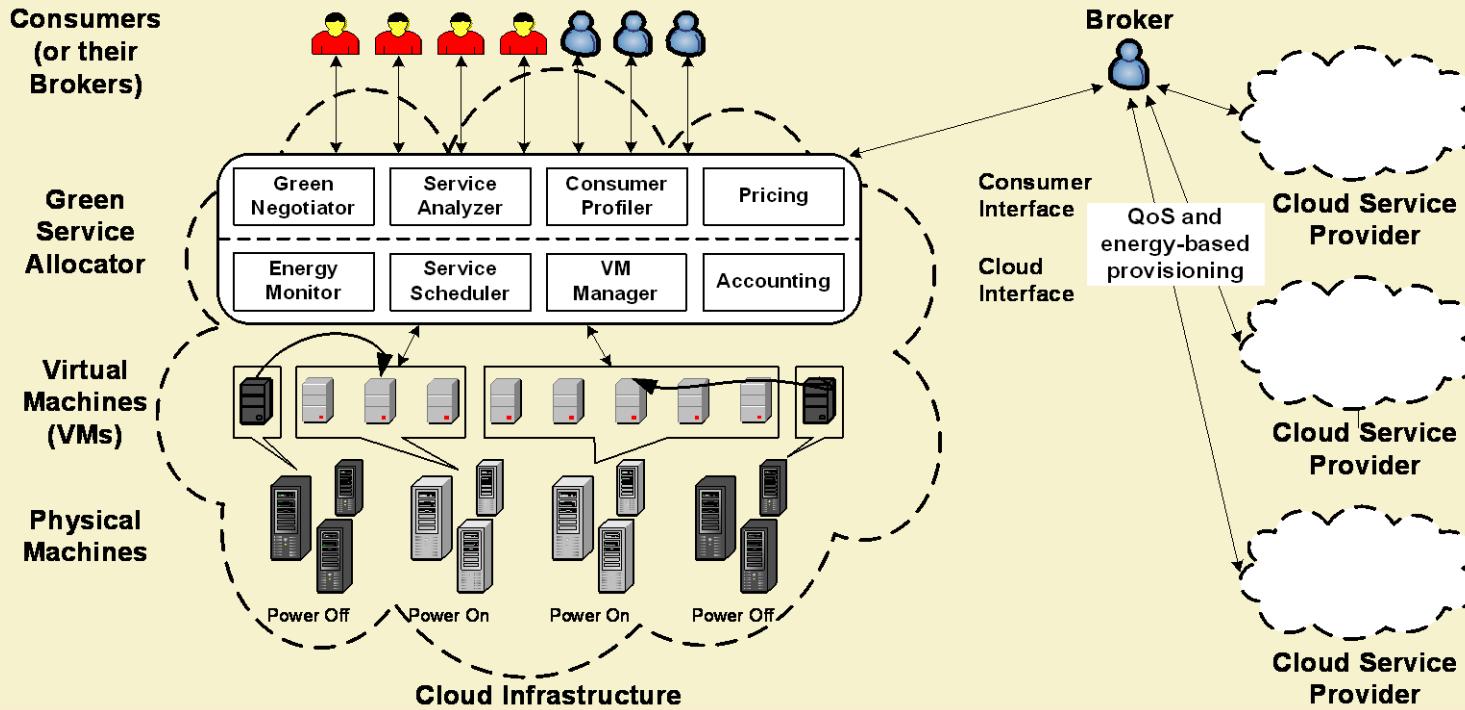


IIT KHARAGPUR



NPTEL ONLINE  
CERTIFICATION COURSES

# A Typical Green Cloud Architecture



Source: Internet



IIT KHARAGPUR



NPTEL  
NPTEL ONLINE  
CERTIFICATION COURSES

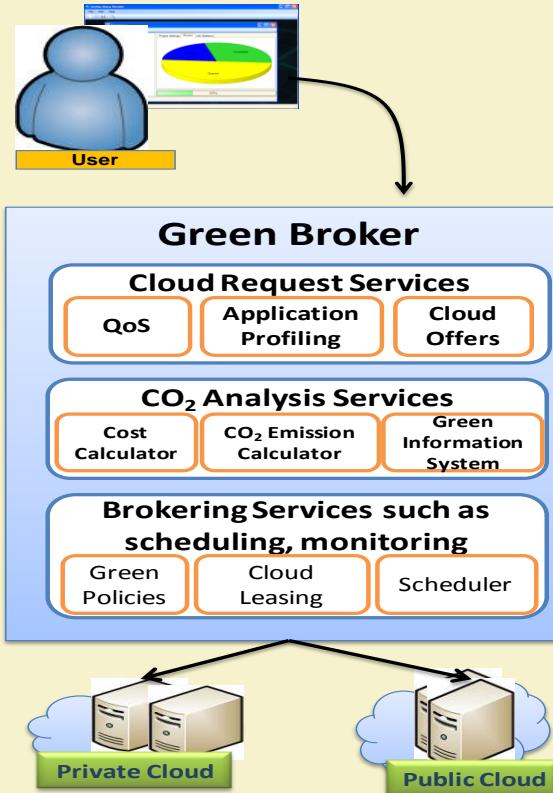
# Green Broker

A typical Cloud broker

- Lease Cloud services
- Schedule applications

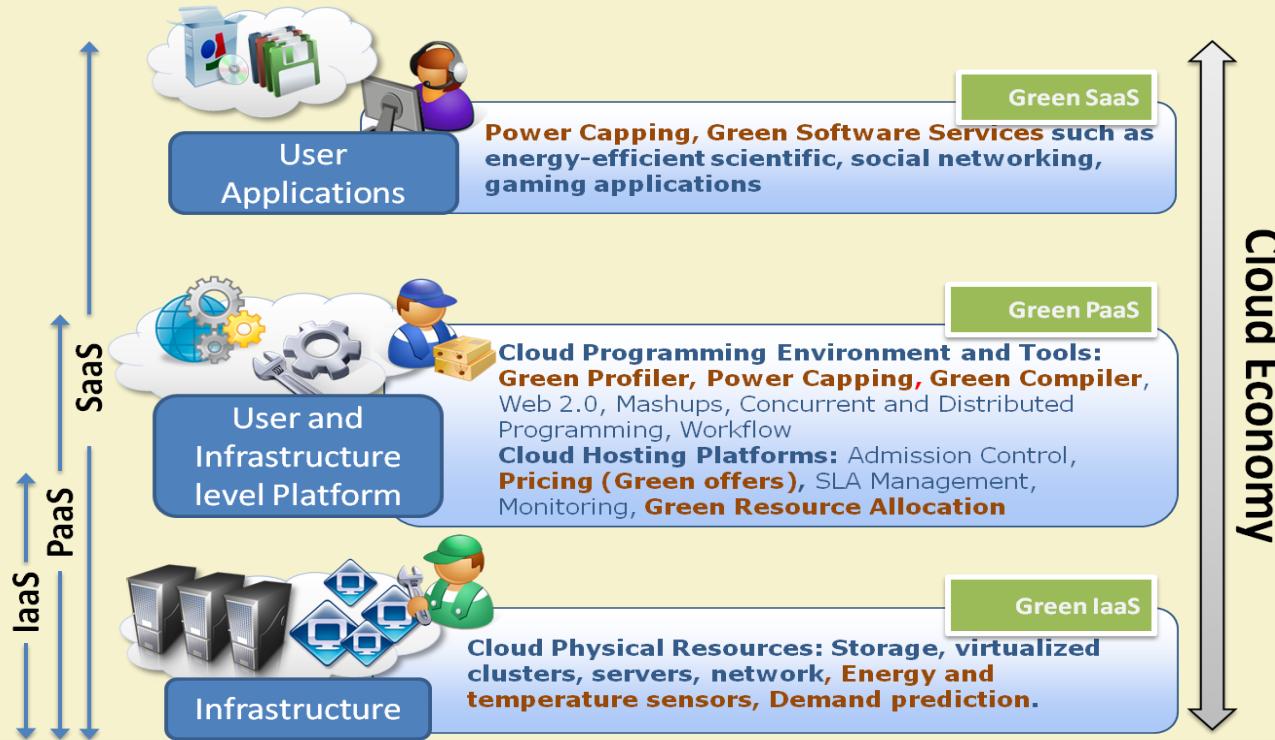
Green Broker

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



Source: Internet

# 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  
NPTEL ONLINE  
CERTIFICATION COURSES

# Thank You!



IIT KHARAGPUR



NPTEL ONLINE  
CERTIFICATION COURSES



IIT KHARAGPUR



NPTEL ONLINE  
CERTIFICATION COURSES

# CLOUD COMPUTING

## Sensor Cloud Computing

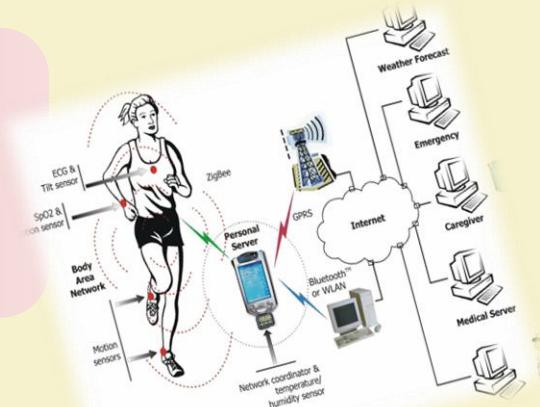
Prof. Soumya K Ghosh

Department of Computer Science and Engineering  
IIT KHARAGPUR

# Motivation



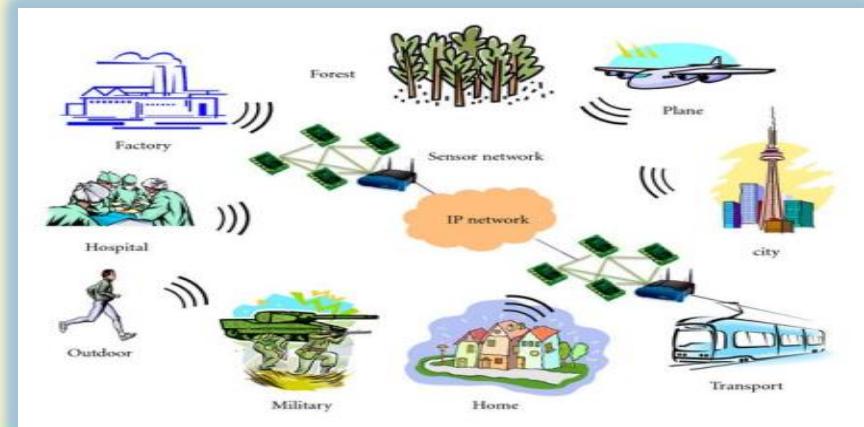
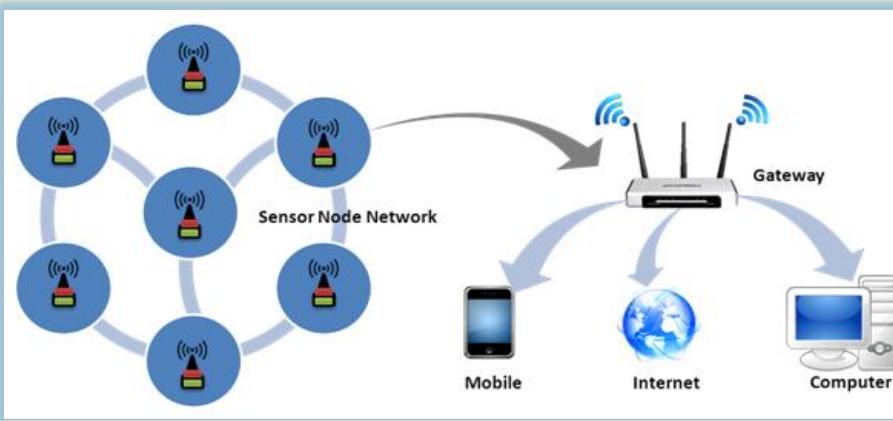
- Increasing adoption of sensing technologies (e.g., RFID, cameras, mobile phones)
- Internet has become a source of real time information (e.g., through blogs, social networks, live forums) for events happening around us



- 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

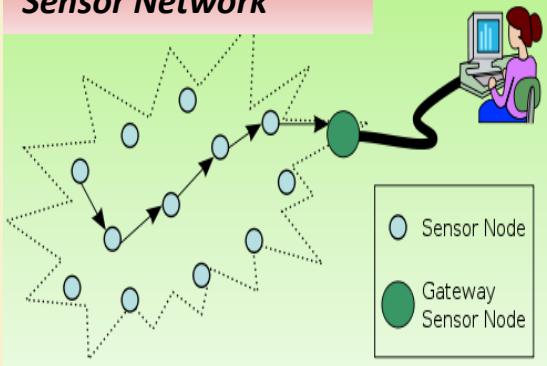
# Limitations of Sensor Networks

- 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.

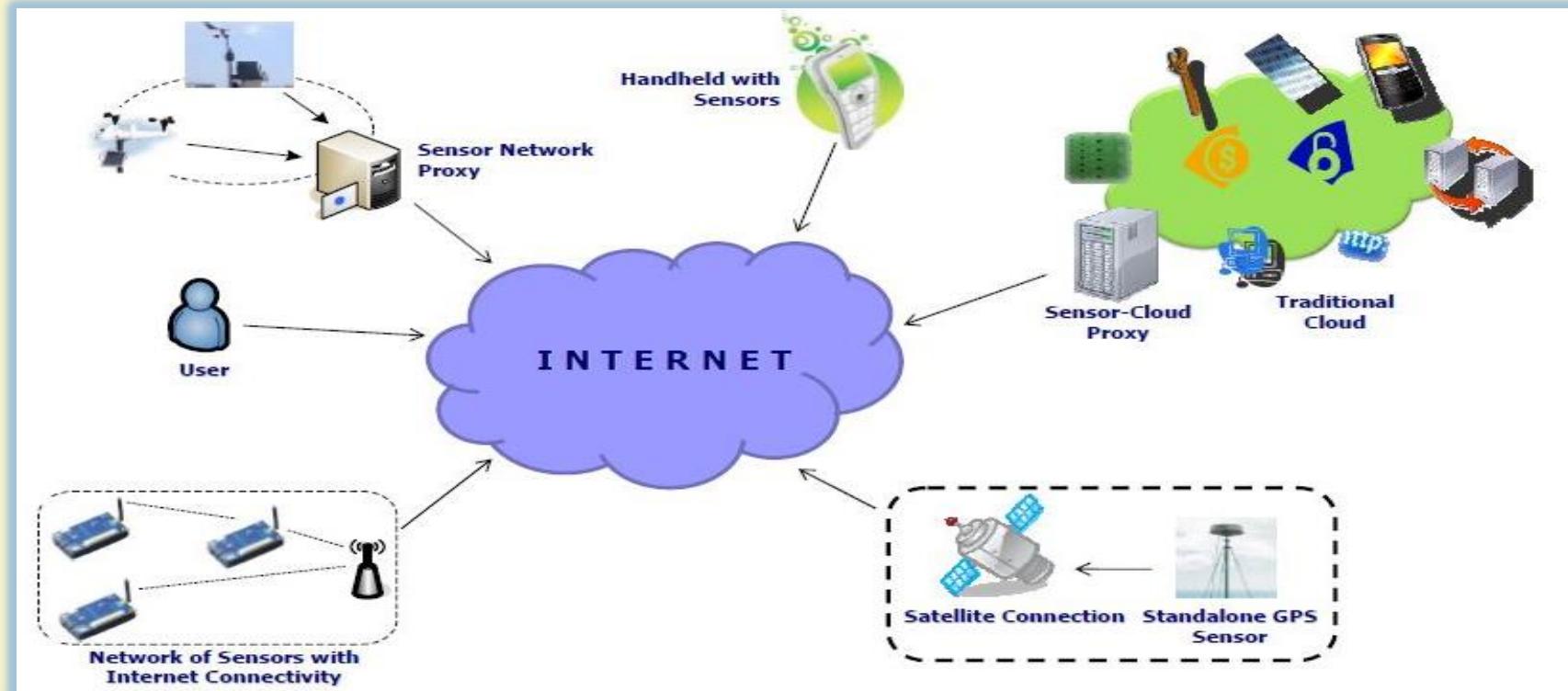


IIT KHARAGPUR



NPTEL ONLINE  
CERTIFICATION COURSES

# Overview of Sensor-Cloud Framework



IIT KHARAGPUR



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 !

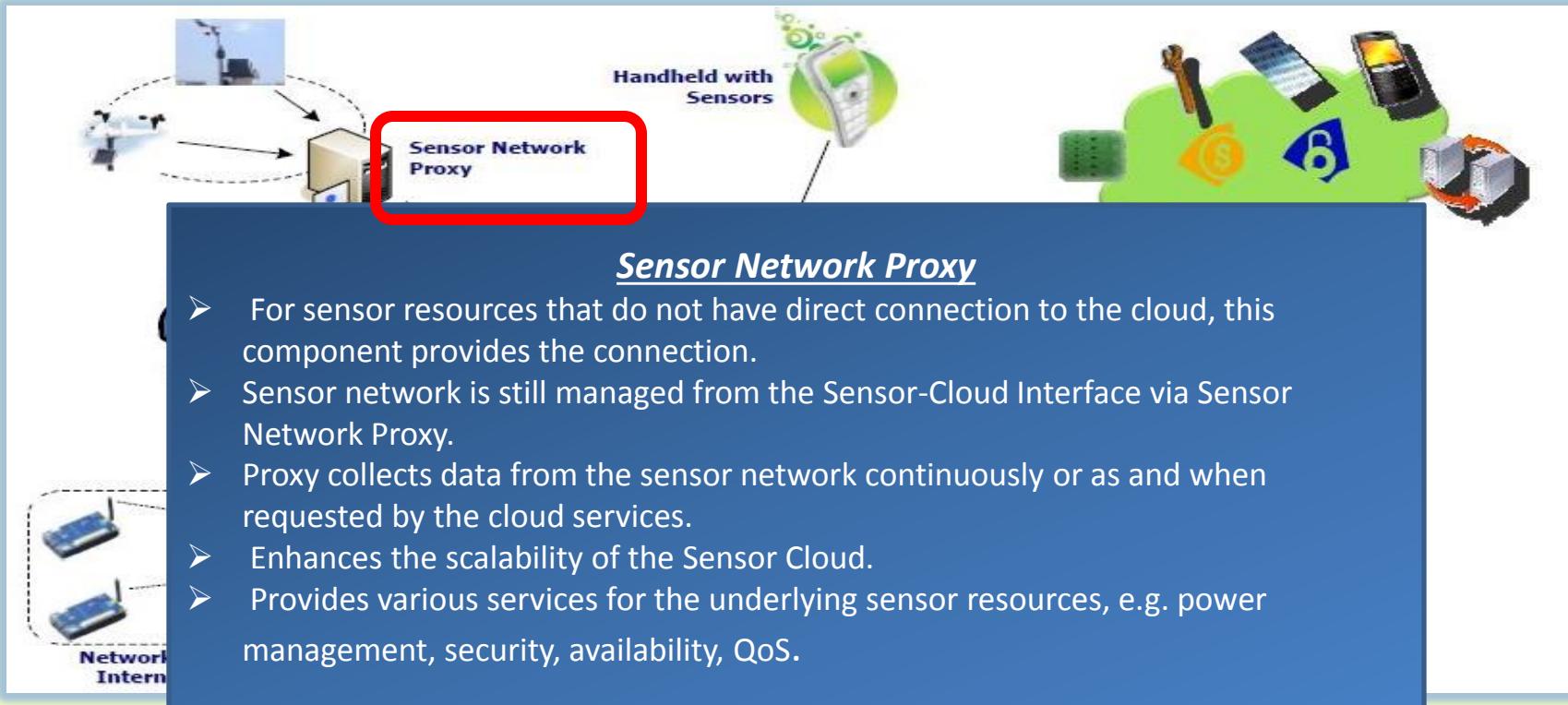


IIT KHARAGPUR



NPTEL  
ONLINE  
CERTIFICATION COURSES

# Overview of Sensor-Cloud Framework



IIT KHARAGPUR

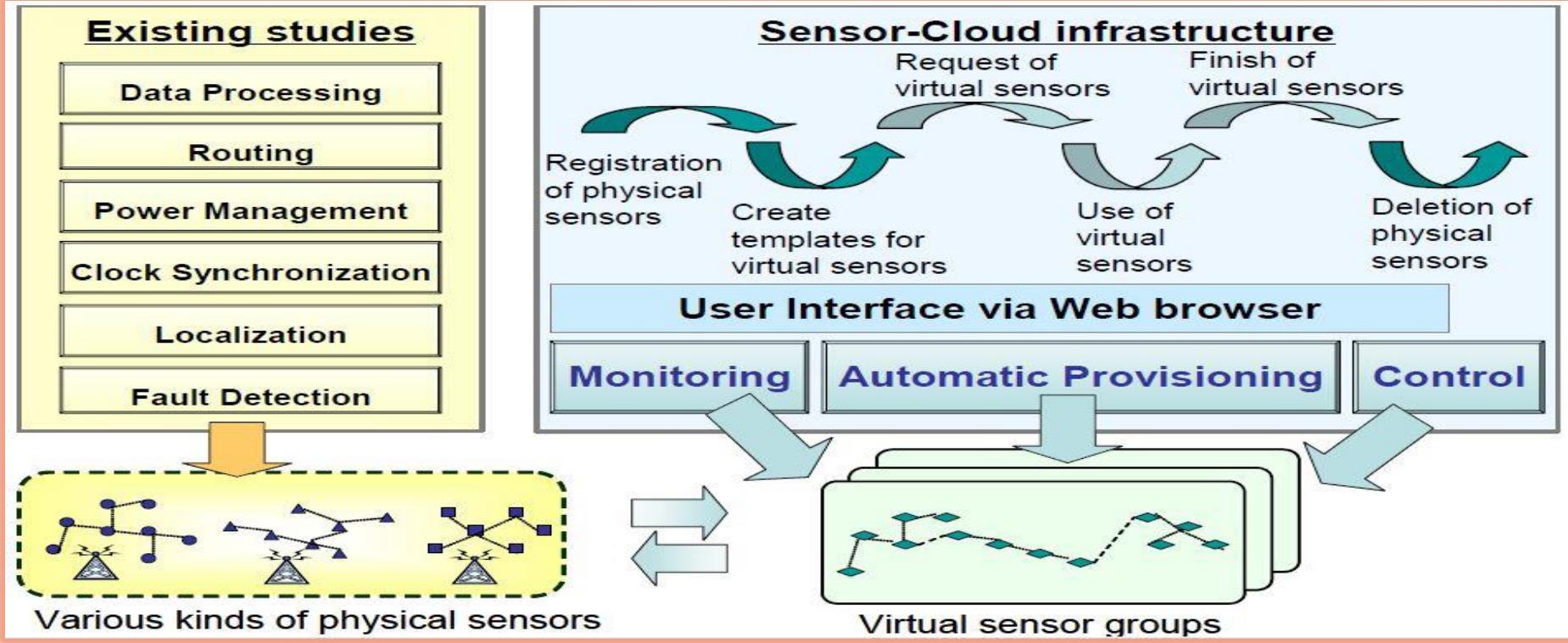


NPTEL ONLINE  
CERTIFICATION COURSES

## Another Use case...

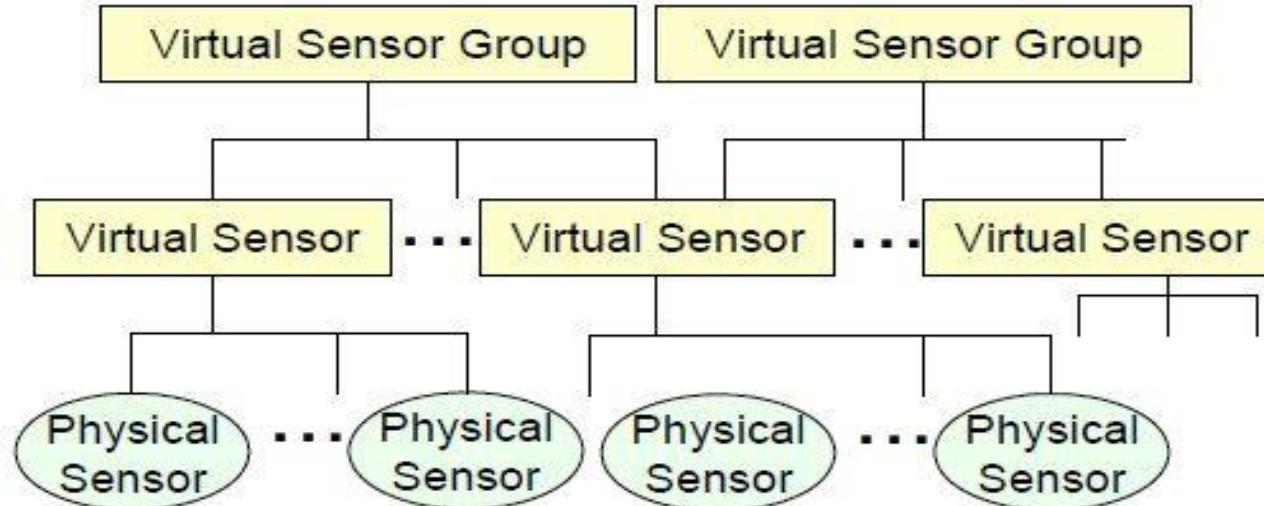
- Traffic flow sensors are widely deployed in large numbers in places/ cities.
- These sensors are mounted on traffic lights and provide real-time traffic flow data.
- Drivers can use this data to better plan their trips.
- In addition, if the traffic flow sensors are augmented with low-cost humidity 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

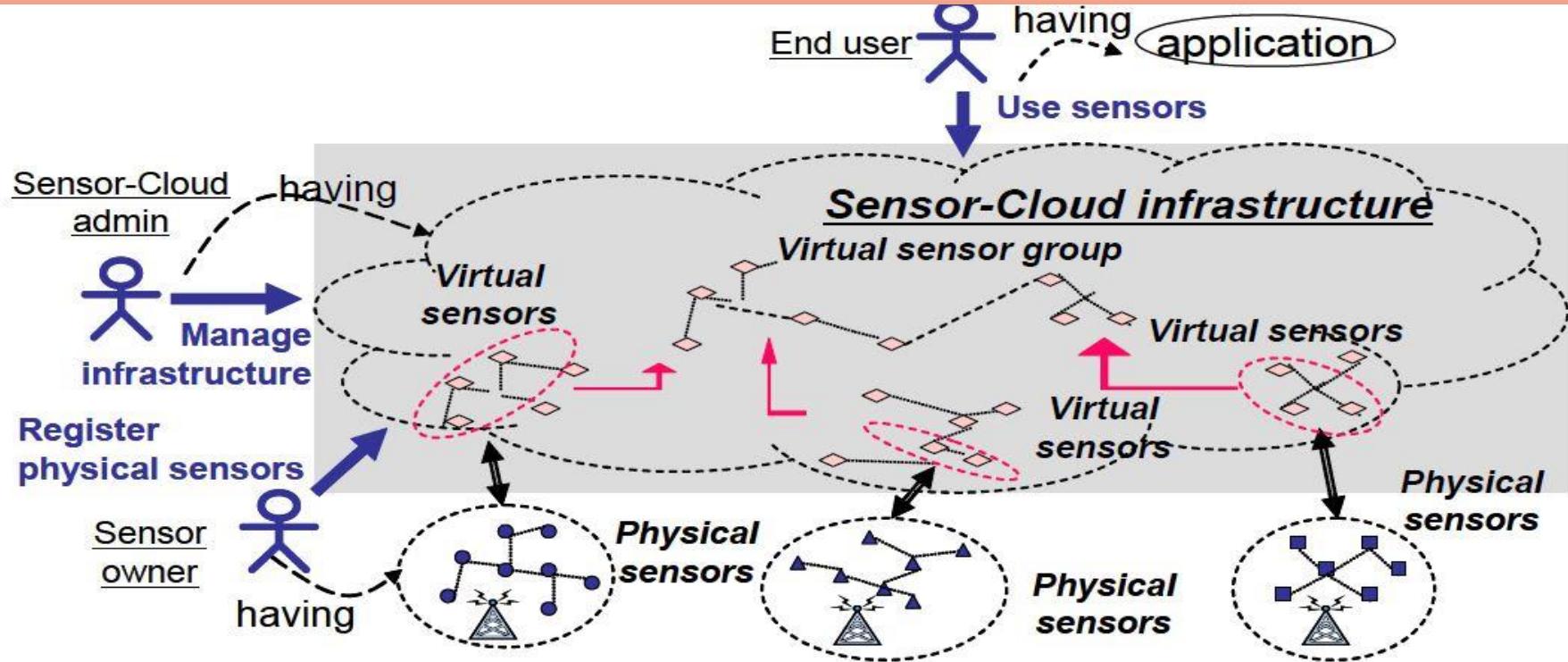


# Virtual Sensors?

- A virtual sensor is an emulation of a physical sensor that obtains its data from under
- Virtual sensors are location tr
- In wireless sensor networks, time and computing
- To overcome the limitation of the current sensor
- The virtual sensors contain metadata about the physical sensors and the user currently holding that virtual sensor.



# Relationship among Actors and Sensor Cloud Infrastructure



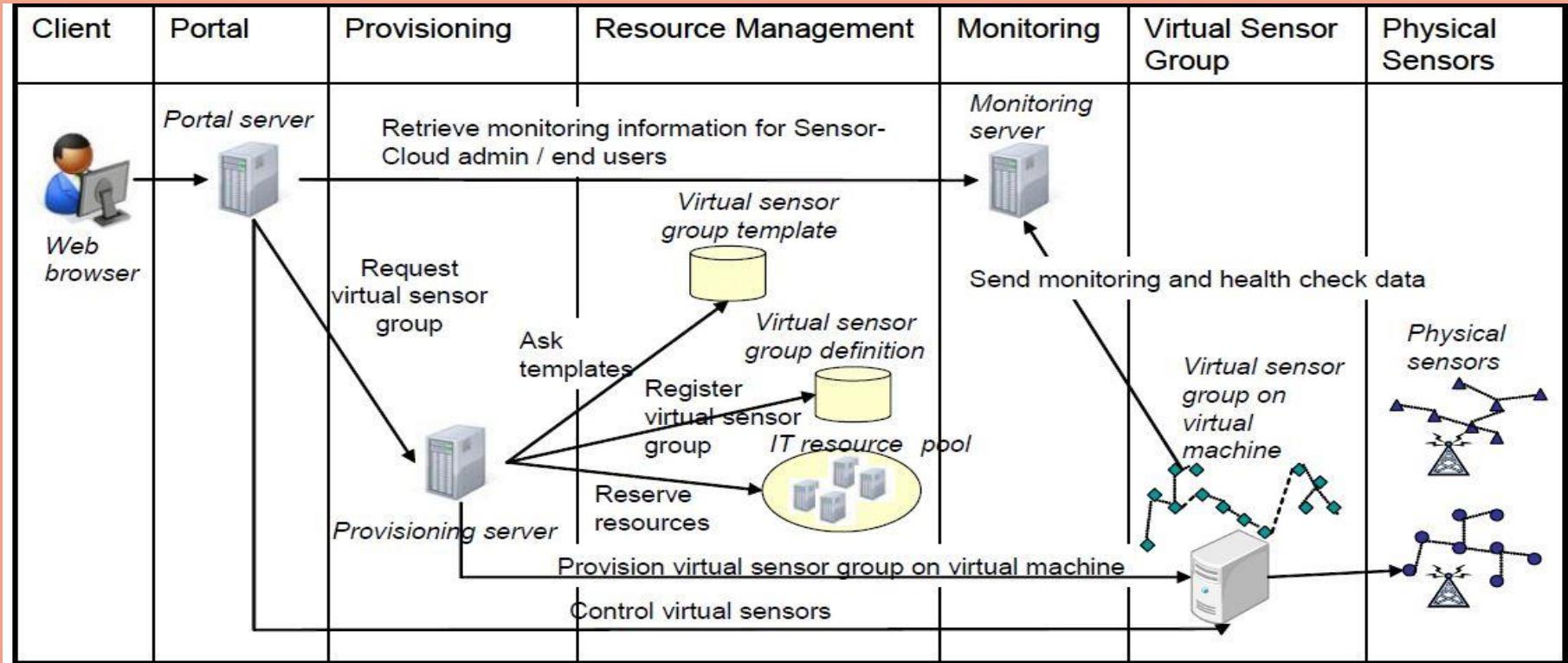
IIT KHARAGPUR



NPTEL  
ONLINE  
CERTIFICATION COURSES

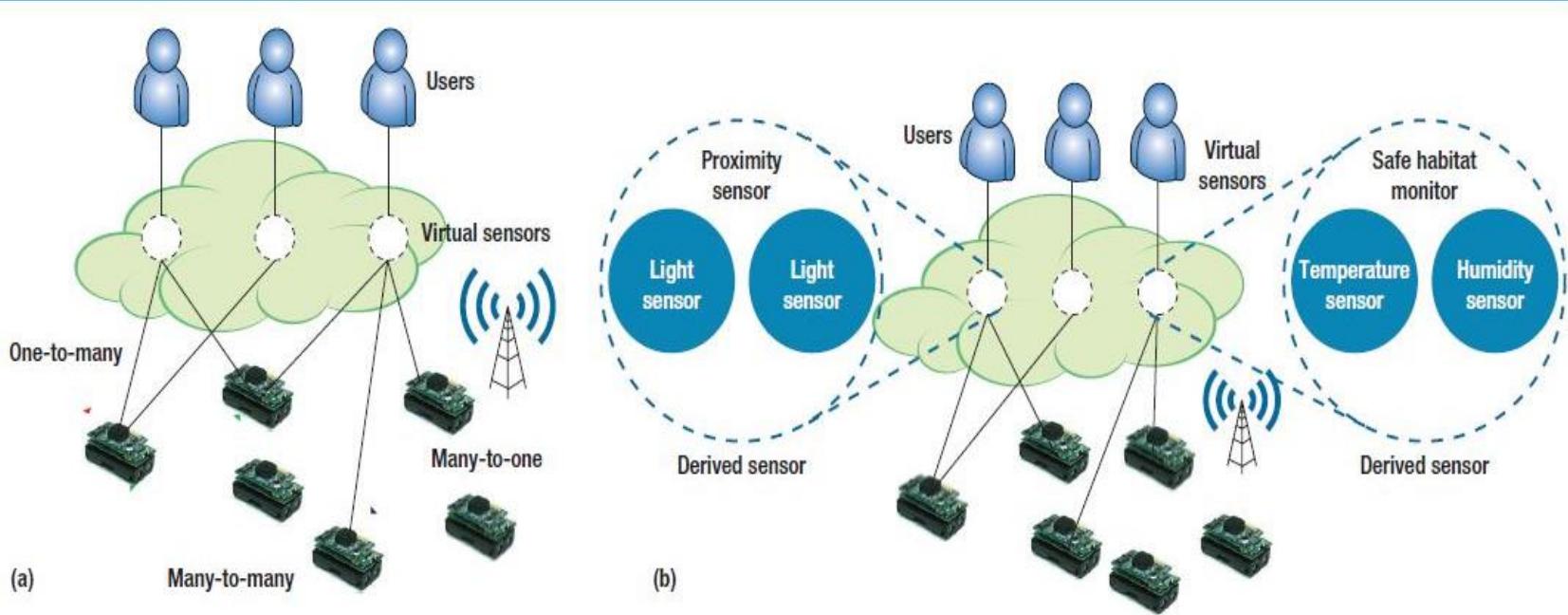
Madoka et al. "Sensor-Cloud Infrastructure Physical Sensor Management with Virtualized Sensors on Cloud Computing"

# System Architecture of Sensor Cloud Infrastructure



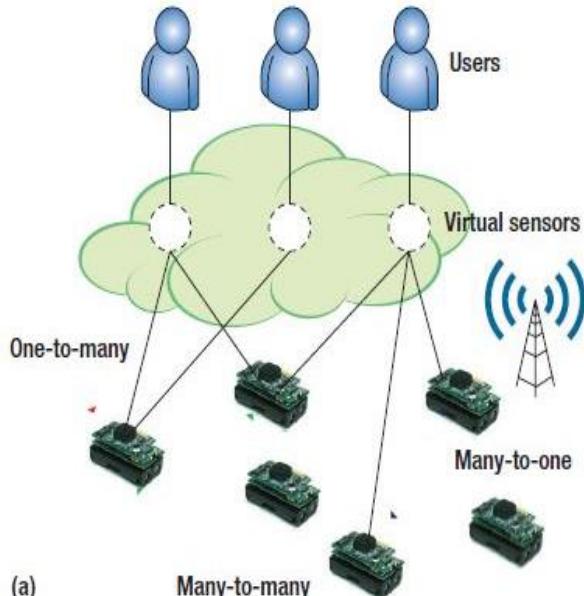
## Virtual Sensor Configurations

(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.



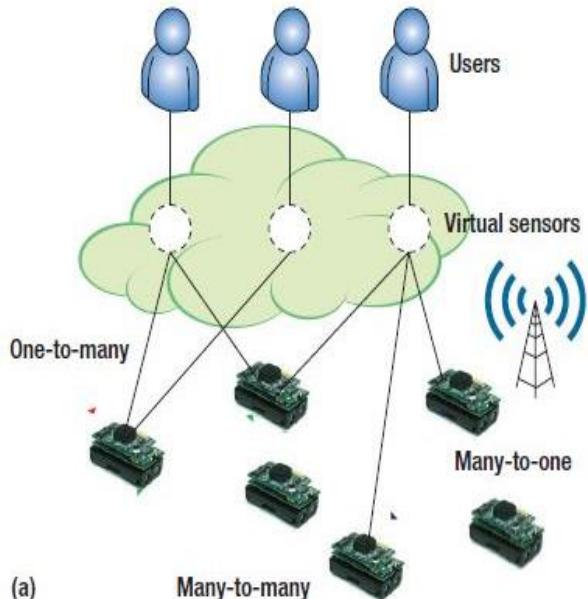
IIT KHARAGPUR



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



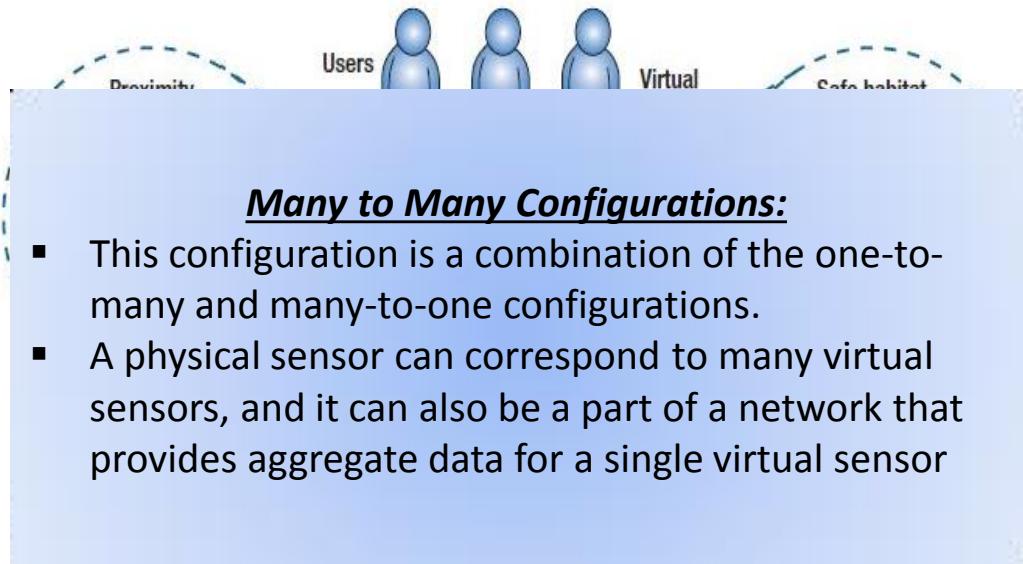
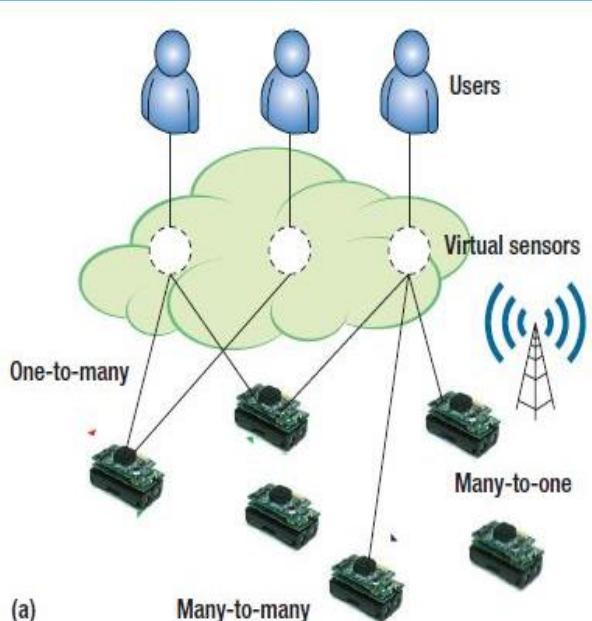
IIT KHARAGPUR



NPTEL  
NPTEL ONLINE  
CERTIFICATION COURSES

# Virtual Sensor Configurations

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



IIT KHARAGPUR



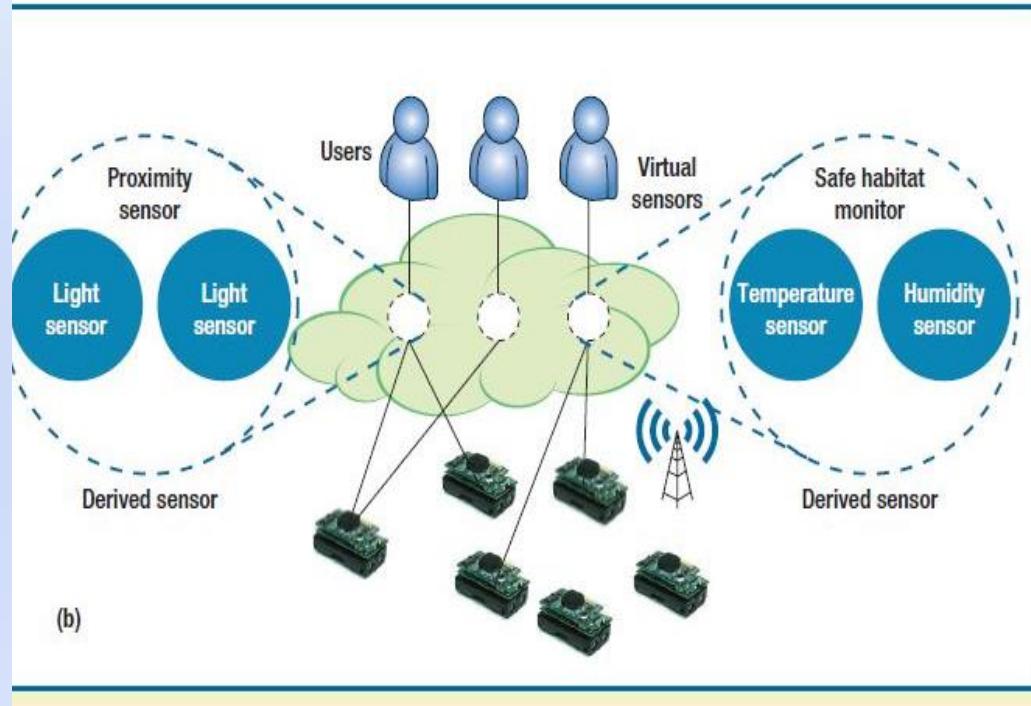
NPTEL ONLINE  
CERTIFICATION COURSES

# 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.



# 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.

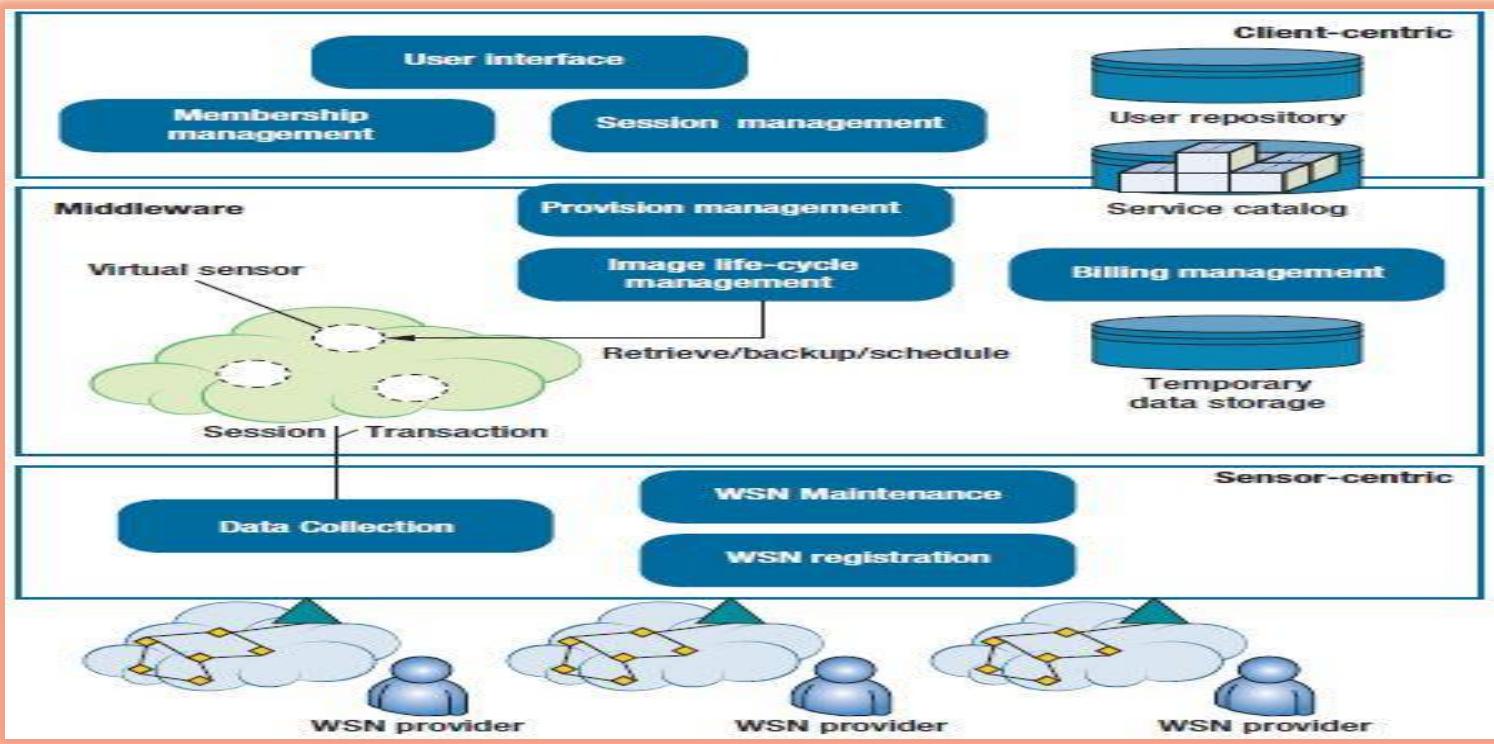


IIT KHARAGPUR



NPTEL  
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
  - arm 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...

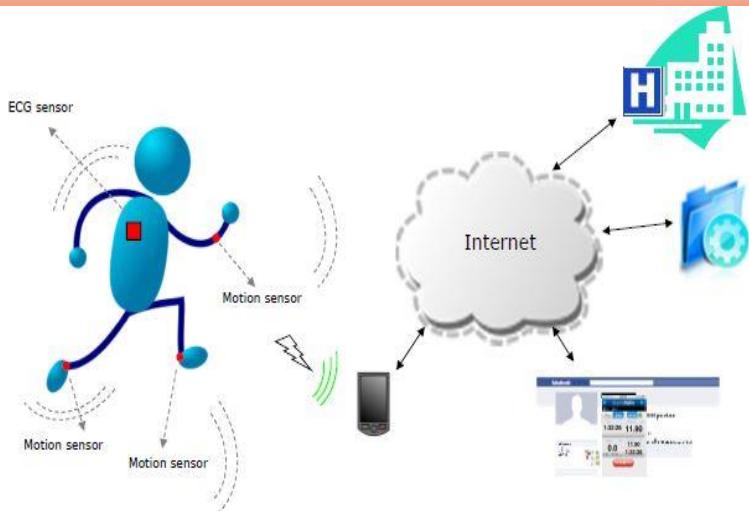


IIT KHARAGPUR

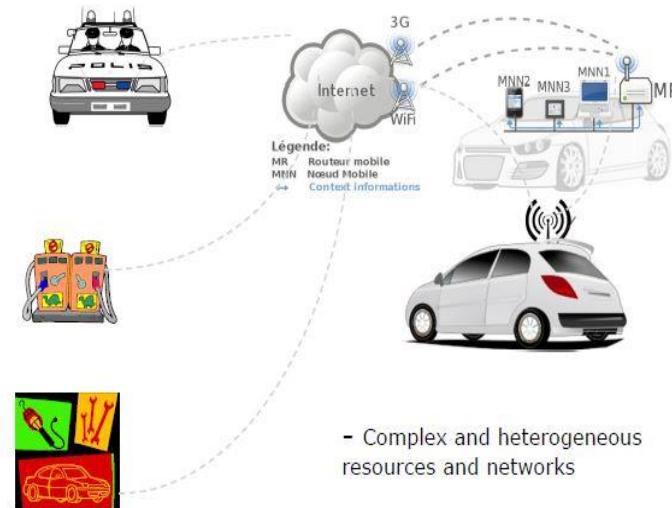


NPTEL ONLINE  
CERTIFICATION COURSES

Any time, Any place connectivity for Anyone and Anything!

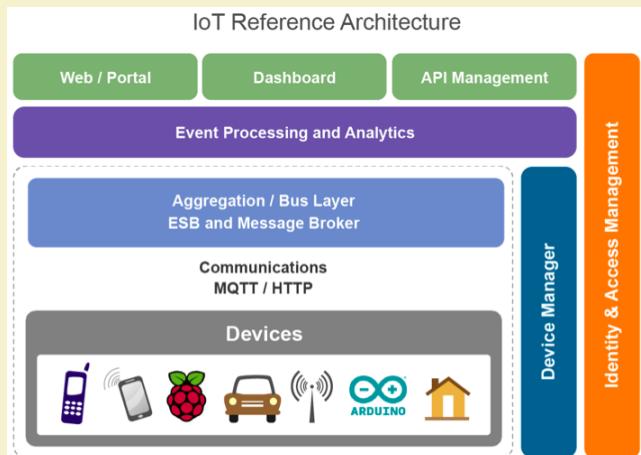


“People” Connecting to “Things”!



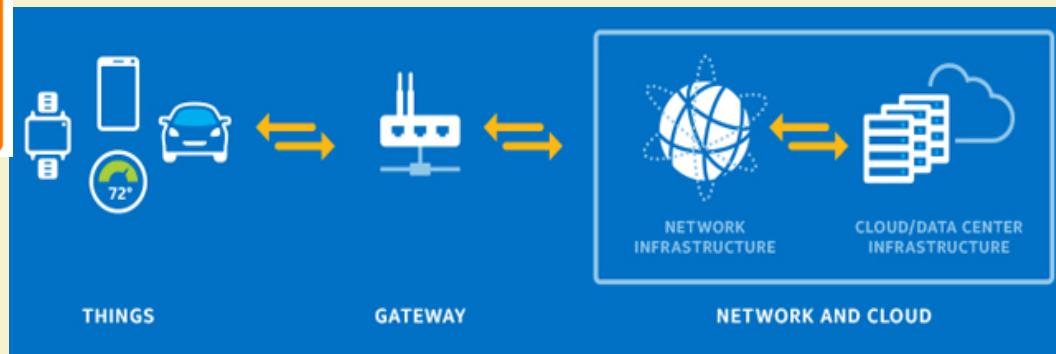
“Things” Connecting to “Things”!

# Basic IoT Architecture



An IoT platform has basically three building blocks

- Things
- Gateway
- Network and Cloud



# Several Aspects of IoT systems!

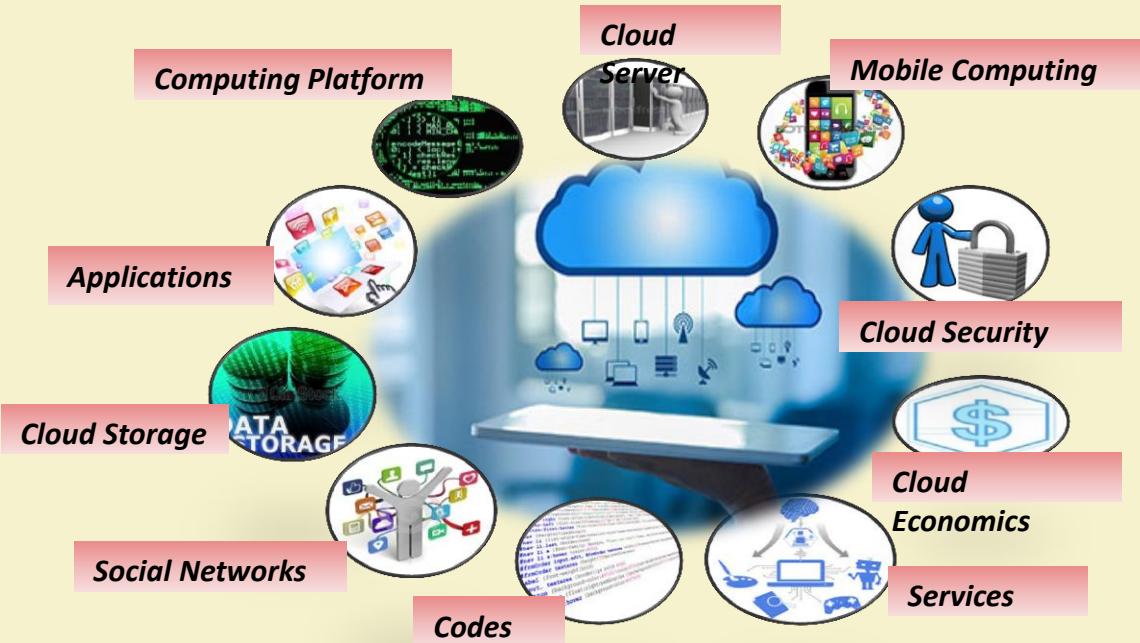
- **Scalability:** Scale for IoT system applies in terms of the numbers of sensors and actuators connected to the system, in terms of the networks which connect them together, in terms of the amount of data associated with the system and its speed of movement and also in terms of the amount of processing 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

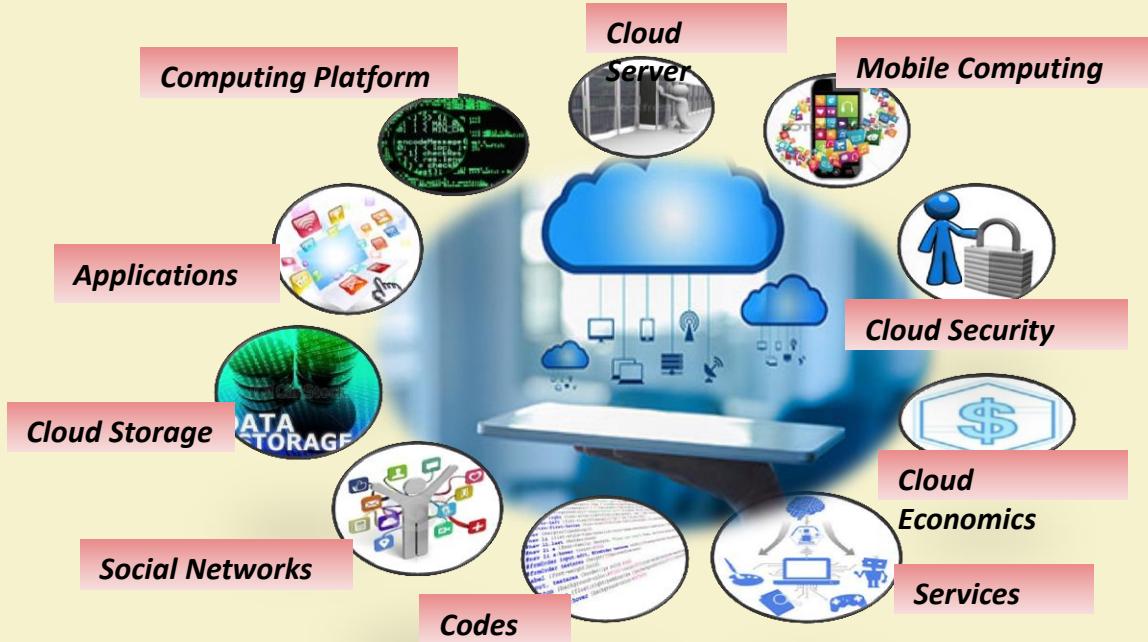
# Cloud Computing!

- 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!

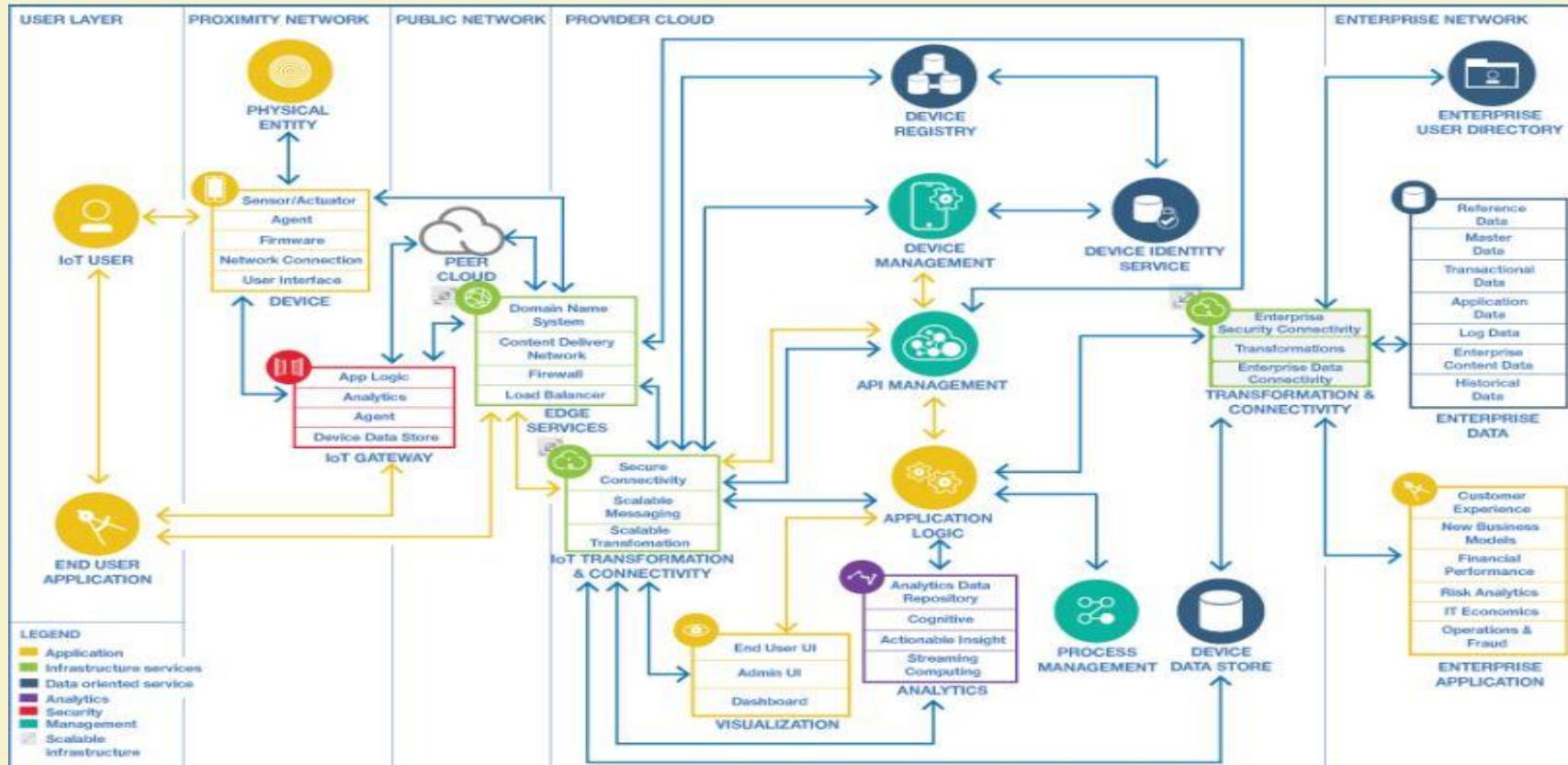
- It enables services to be used without any understanding of the infrastructure.
- Cloud computing works using economies of scale
- Data and services are stored remotely but accessible from “anywhere”.



# 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

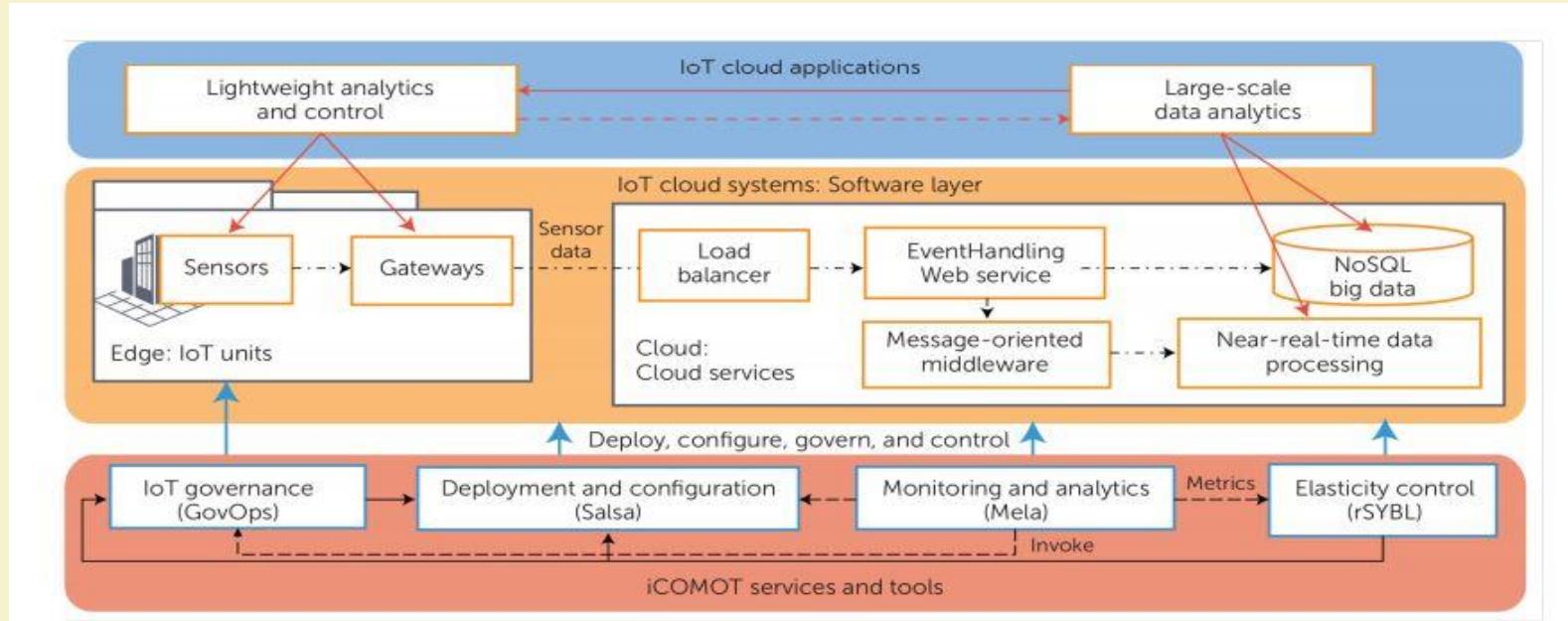


IIT KHARAGPUR



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.

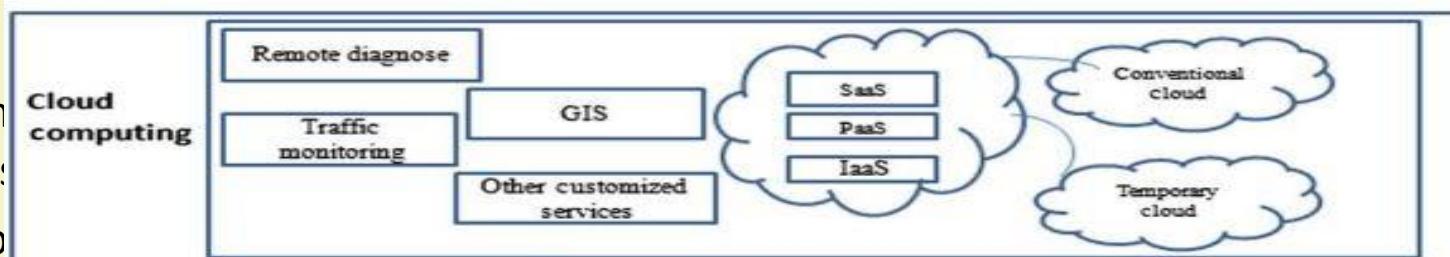
*Bottom layer* shows different tools and services from iCOMOT that can be used to monitor, control, and configure the software layer.

# Infrastructure, Protocols and Software Platforms for establishing an Internet of Things (IoT) Cloud system

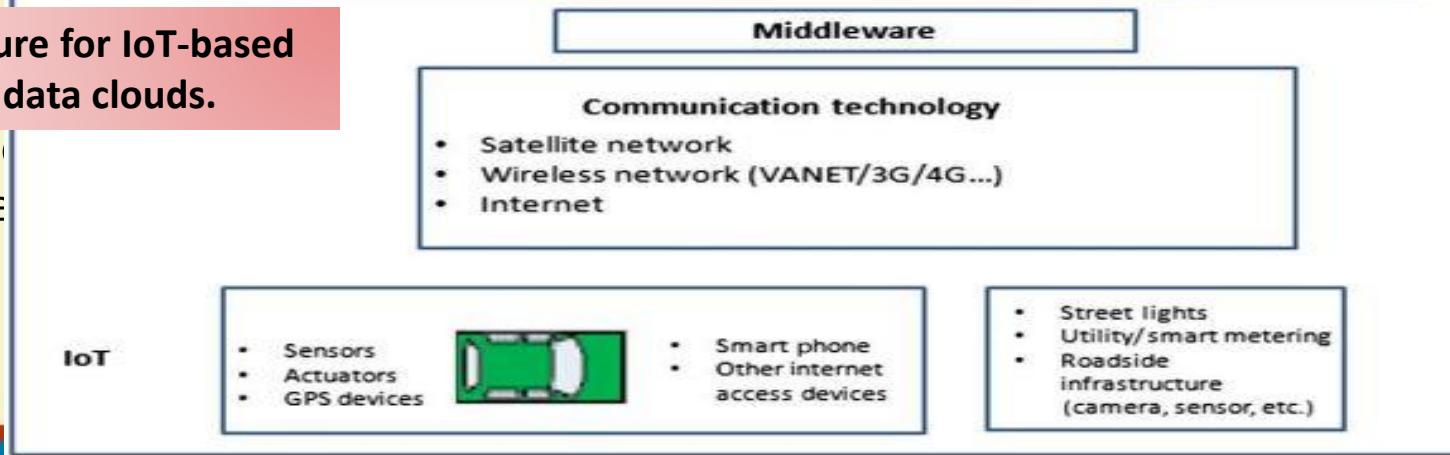
Types	IoT	Clouds	Purpose
Infrastructure machines	Industrial and common gateways (for example, Intel IoT Gateway) and operating system containers (such as Dockers)	Virtual machines and operating system containers	Enable (virtual) machines where software components will be executed
Connectivity protocols	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



Architecture for IoT-based  
vehicular data clouds.



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

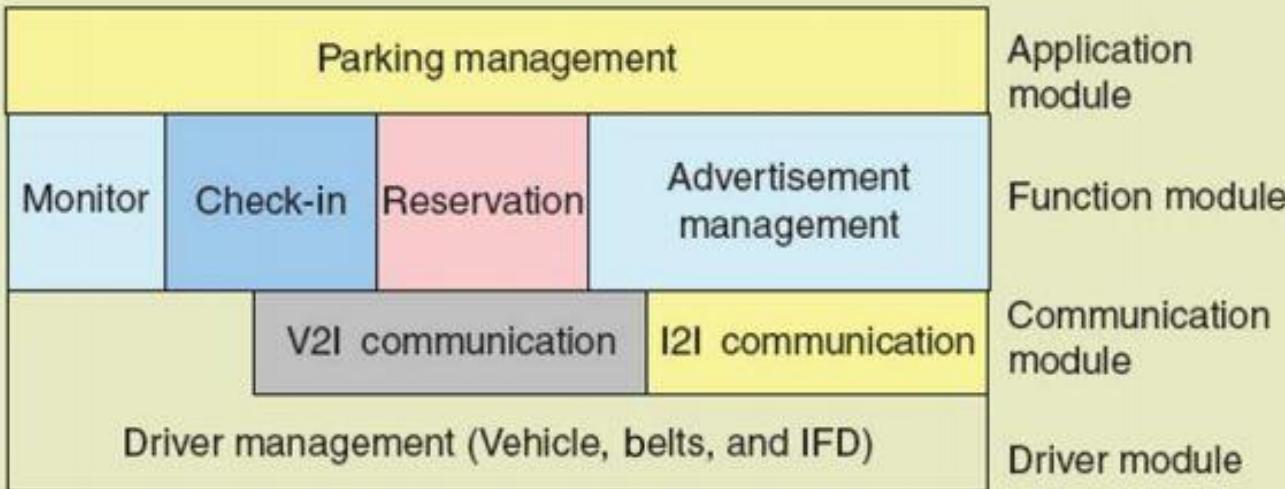


IIT KHARAGPUR



NPTEL  
NPTEL ONLINE  
CERTIFICATION COURSES

# Architecture for Intelligent Parking Cloud service

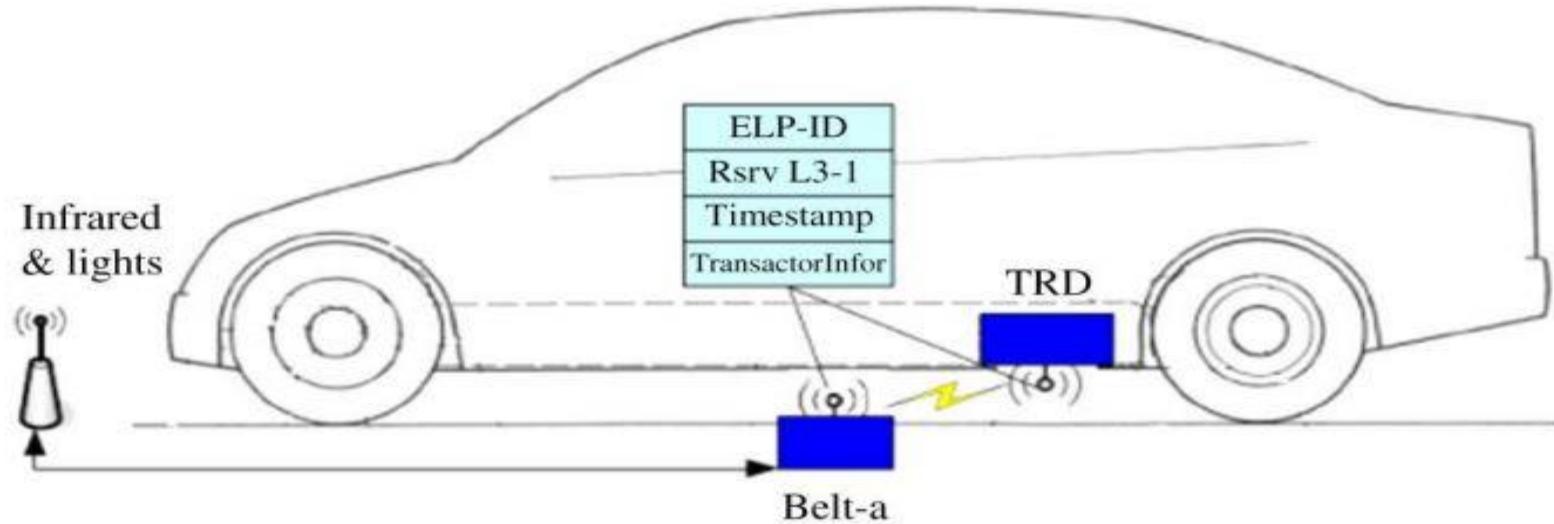


IIT KHARAGPUR



NPTEL  
NPTEL ONLINE  
CERTIFICATION COURSES

# Vacancy detections by Sensors

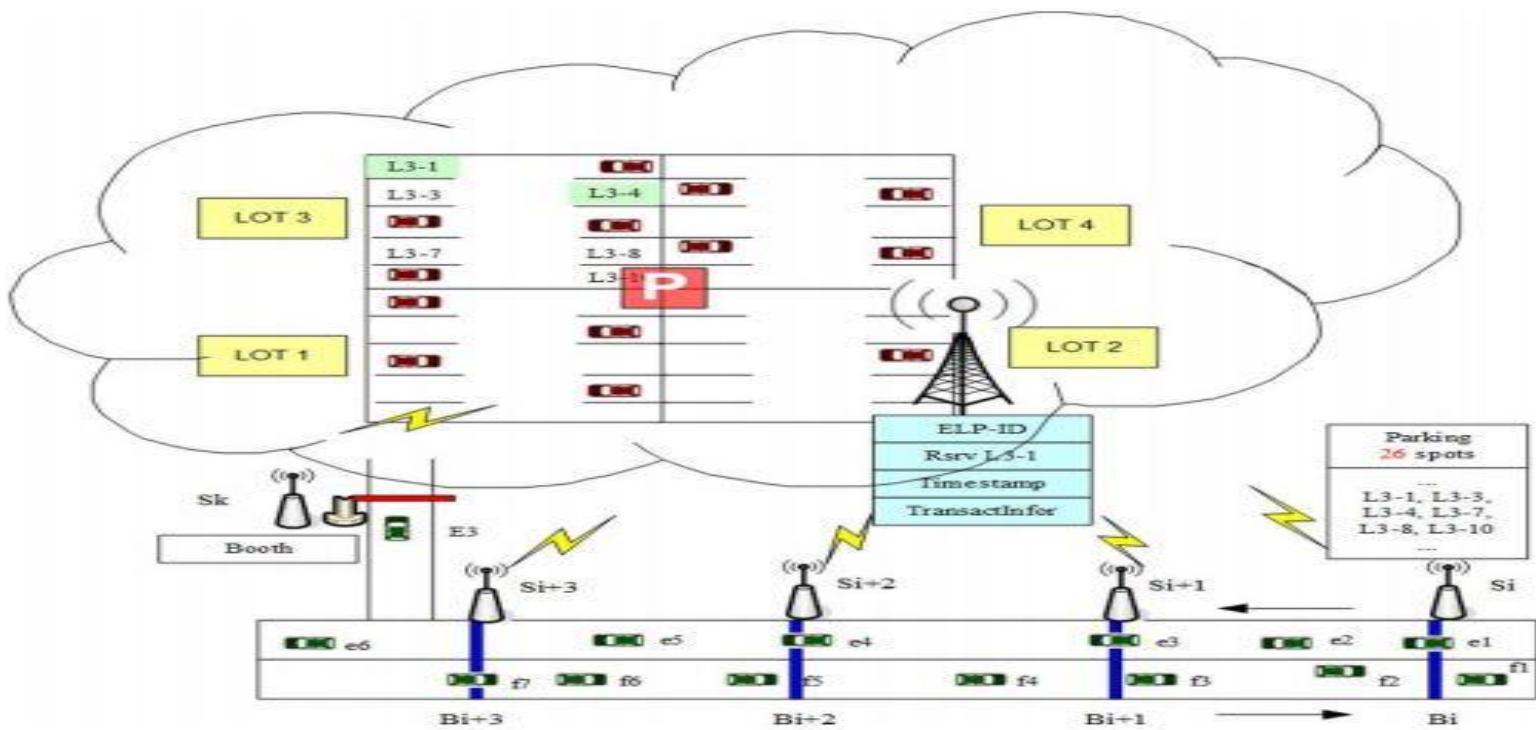


IIT KHARAGPUR



NPTEL  
NPTEL ONLINE  
CERTIFICATION COURSES

# Parking cloud service



IIT KHARAGPUR



NPTEL  
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!



IIT KHARAGPUR



NPTEL  
NPTEL ONLINE  
CERTIFICATION COURSES



IIT KHARAGPUR



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  
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  
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



IIT KHARAGPUR



NPTEL ONLINE  
CERTIFICATION COURSES

# Performance, Scalability, Reliability

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



IIT KHARAGPUR



NPTEL  
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



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  
NPTEL ONLINE  
CERTIFICATION COURSES

# Cloud and Other Technologies

- Fog Computing
- IoT Cloud
- Container Technology



IIT KHARAGPUR



NPTEL ONLINE  
CERTIFICATION COURSES

# Thank You!



IIT KHARAGPUR



NPTEL  
NPTEL ONLINE  
CERTIFICATION COURSES

# Fundamental of Cloud Security

Salim Hariri, UA Site Director

NSF Center for Cloud and Autonomic Computing

The University of Arizona

[nsfcac.arizona.edu](http://nsfcac.arizona.edu)

[hariri@ece.arizona.edu](mailto:hariri@ece.arizona.edu)



# Presentation Outline

- Introduction
- Cloud Computing Standards
- Cloud Security Issues
- Cloud Attack Mechanisms
- Cloud Protection and Solutions

# ■ INTRODUCTION



# Cloud Computing – Motivation

## Car rental services

- For short period
- Before you get your own car
- No need to maintain and upgrade
- Is popular

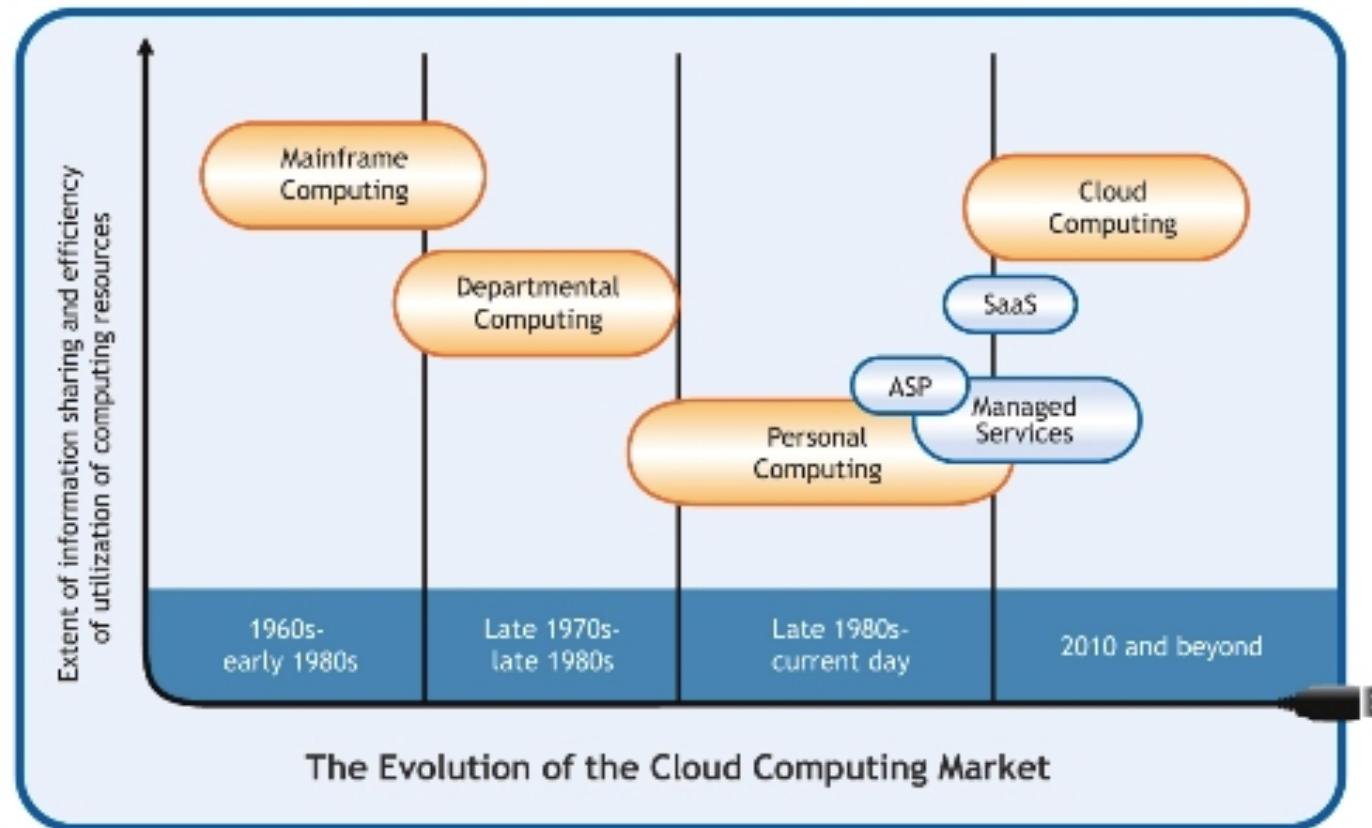
## Cloud rental services

- For short period
- Before you get your own devices
- No need to maintain and upgrade

# Cloud Computing Potential Benefits

- Increased Reliability – Duplicated data, logs, better maintenance
- Reduction in IT operating costs (**Pay-as-you-Go**)
- Scalability and Agility
- Ubiquitous Accessibility – Internet, and perform same task from any where and using any network device
- Levels the playing field
- Fast request-driven provisioning (**On Demand**)
- Improves collaboration

# How the Cloud is growing?



\* Source: <http://www.forbescustom.com/TechnologyPgs/CloudComputingP1.html> [accessed: May 26, 2013]

# Cloud Computing Growth

- Cloud usage is like having a customized cellular plan with all the features and functionality that you want, paying only for what you use, and with the ability to cancel at anytime without penalties or additional fees.
- Worldwide cloud service revenue grew at 16.6% in 2010, reaching \$68.3 billion, according to Gartner report.
  - It is expected that enterprises will spend in the next five years around \$112 billion on cloud technologies and services

# ■ CLOUD COMPUTING STANDARD



# NIST definition of cloud computing

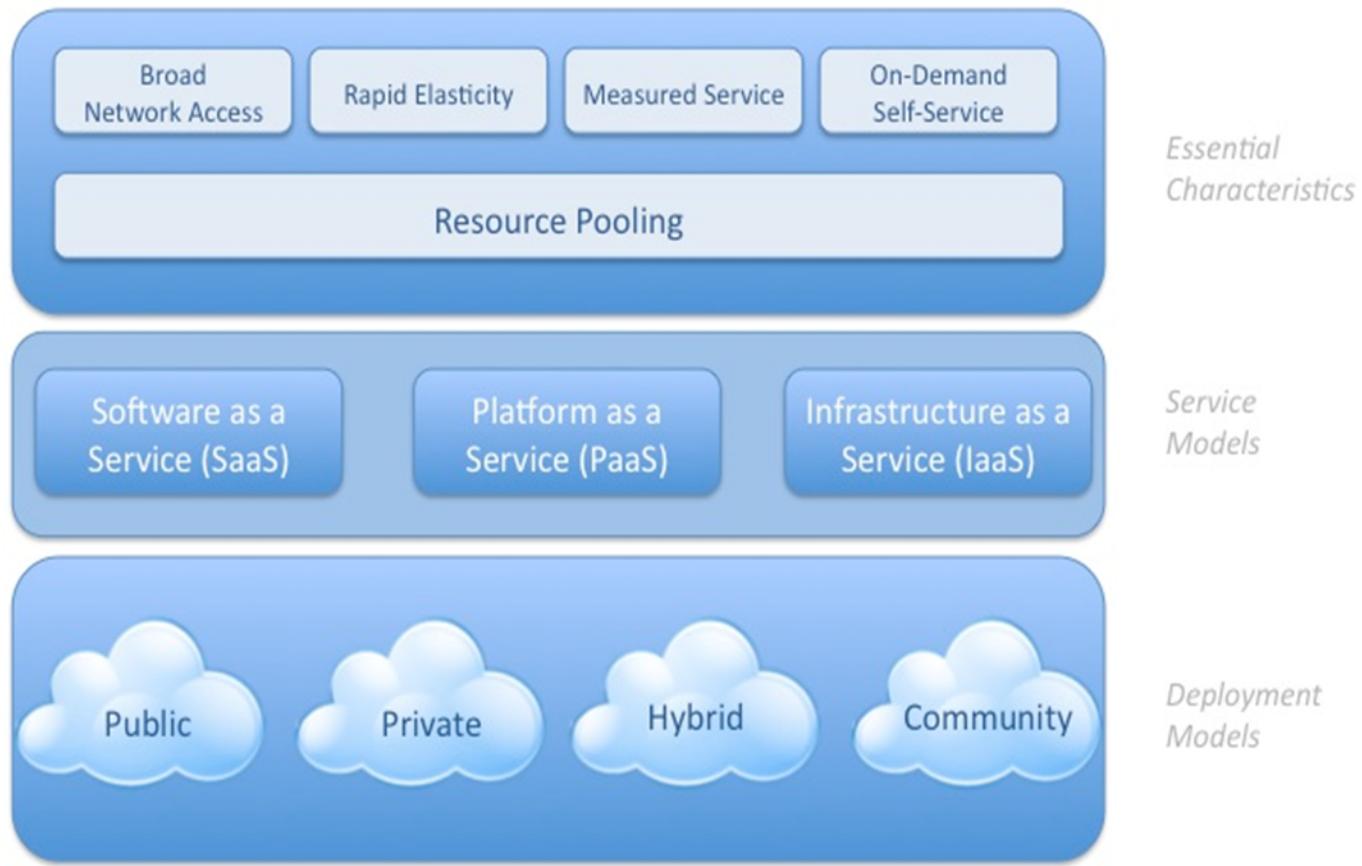
- Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction.



# What Comprises Cloud Computing?

- ❖ NIST defines:
  - Five essential cloud characteristics
  - Three cloud service models
  - Four cloud deployment models.

# NIST Model of Cloud



# Five Essential Cloud Characteristics

- On-demand self-service
- Broad network access
- Resource pooling
- Rapid elasticity
- Measured service

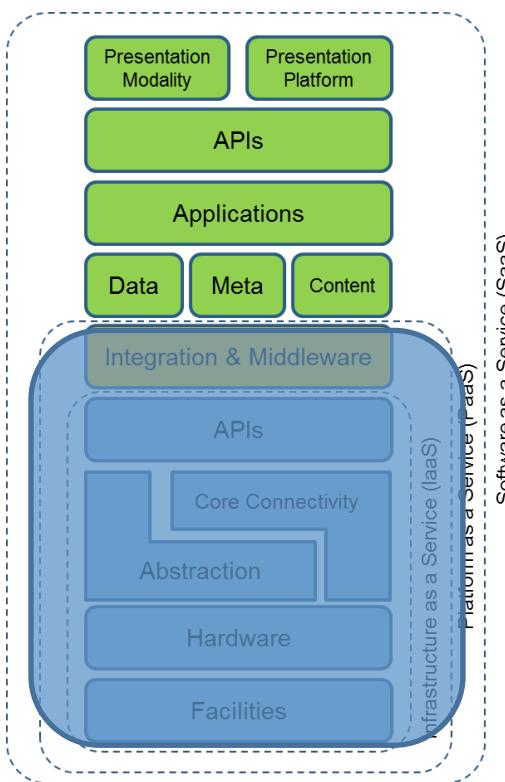
# Three Cloud Service Models

- Cloud Software as a Service (SaaS)<sup>\*</sup>
  - To use the provider's applications
- Cloud Platform as a Service (PaaS)<sup>\*</sup>
  - To deploy customer-created and acquired applications
- Cloud Infrastructure as a Service (IaaS)
  - To provision processing, storage, networks, and other fundamental computing resources

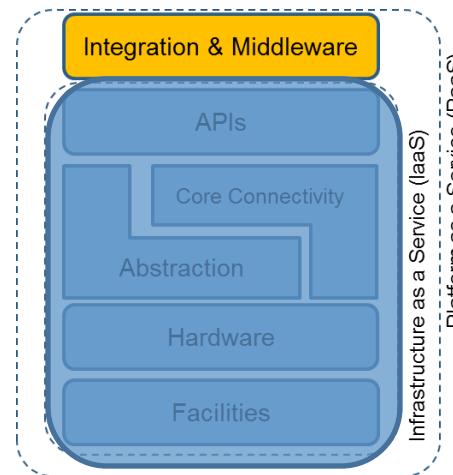
\* *To be considered as cloud services, they must be running on top of an cloud infrastructure.*

# Cloud Service Delivery Models

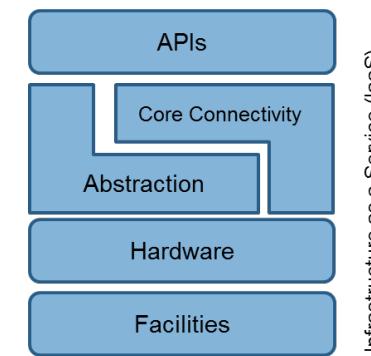
## SaaS



## PaaS

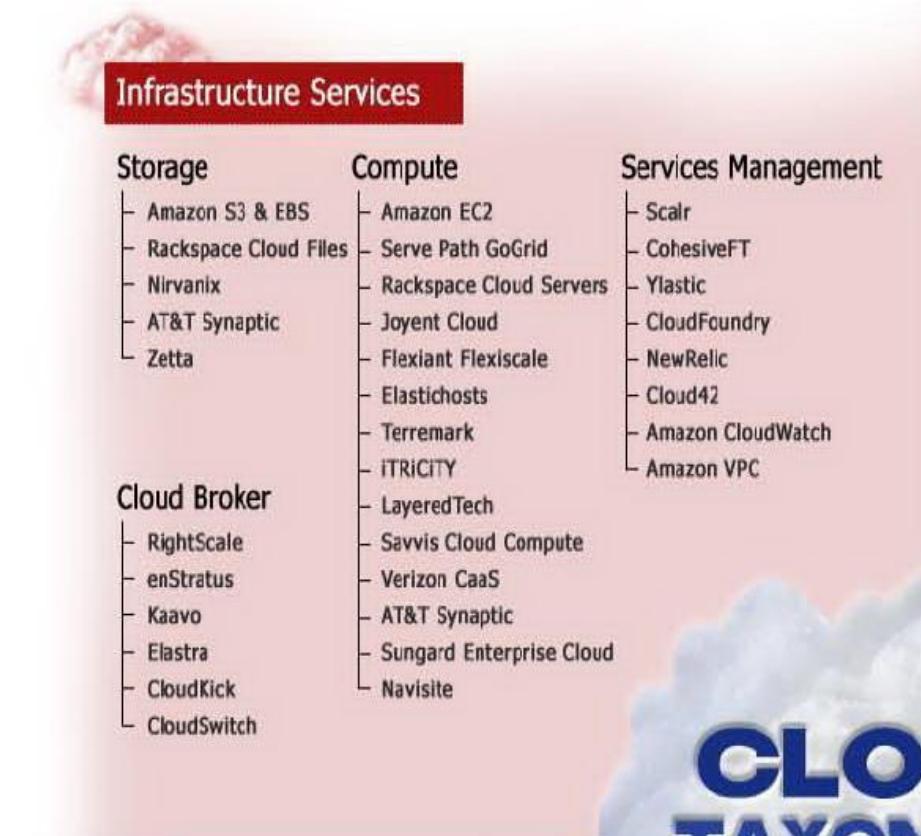


## IaaS

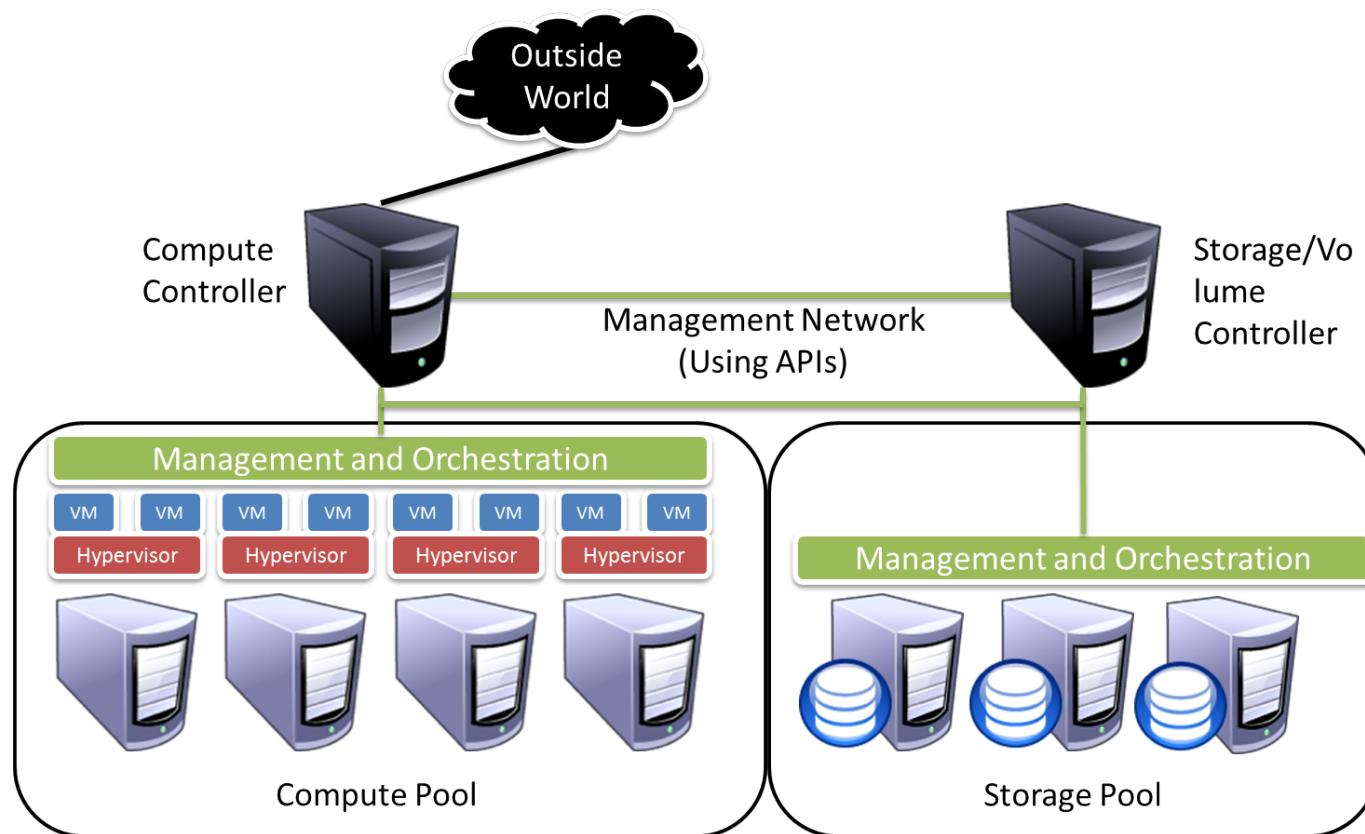


# Cloud Service Models –IaaS

- It delivers computer infrastructure as a service, along with raw storage and networking
- Rather than purchasing servers, software, data-center space, or network equipment, clients buy them as a fully outsourced service



# What is IaaS?



Source: Securisys, L.L.C. / Cloud Security Alliance

# IaaS

## Benefits

- Tremendous control to use whatever content system makes sense.
- Flexibility to secure data to whatever degree necessary.

## Issues

- Involves integrating all aspects of an application (DB, plug-ins, etc.)
- Responsible for all configurations implemented on the server (and in apps)
- Responsible to keep software up to date
- Multi-tenancy at hypervisor level

Src: Securisis, L.L.C. / Cloud Security Alliance

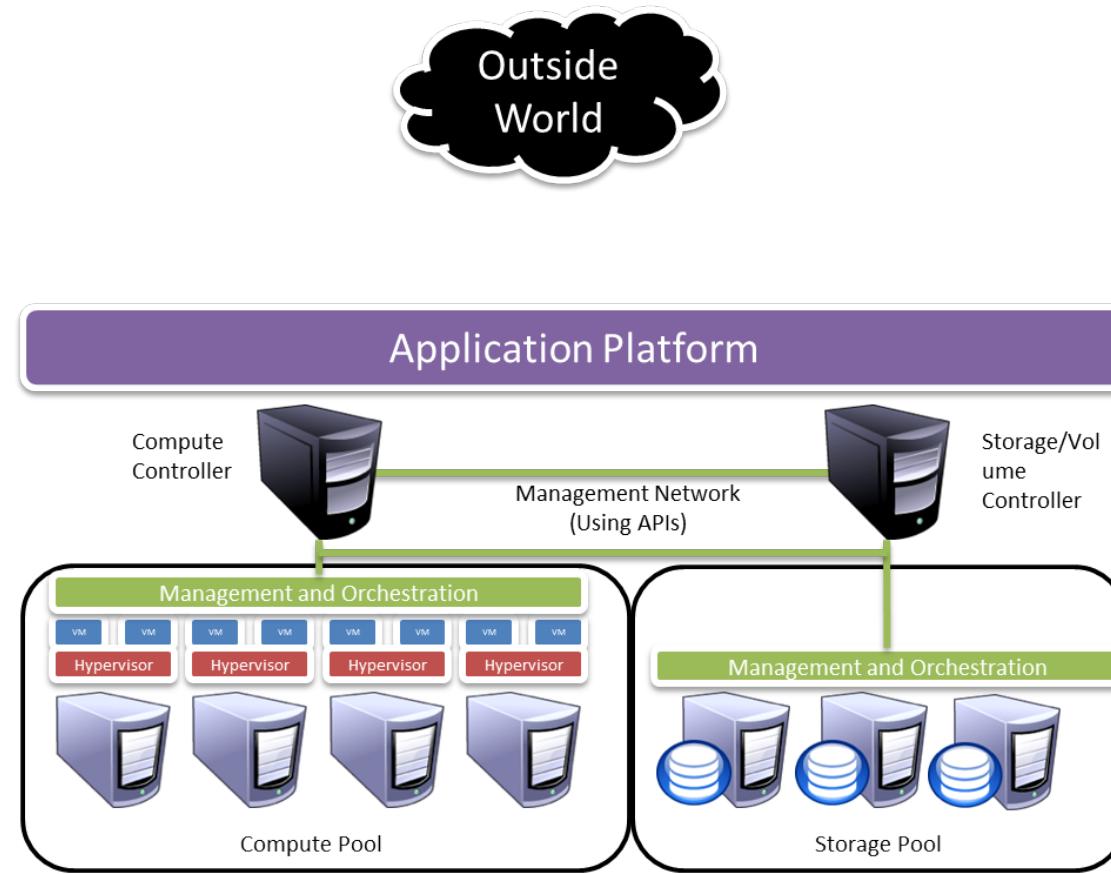


# Cloud Service Models – PaaS

- It delivers a computing platform and solution stack as a service. PaaS offering facilitate deployment of applications without the cost and complexity of buying and managing the underlying hardware and software and provisioning hosting capabilities

Platform Services				
General Purpose	Business Intelligence	Integration	Development & Testing	Database
<ul style="list-style-type: none"><li>- Force.com</li><li>- Etelos</li><li>- LongJump</li><li>- Rollbase</li><li>- Bungee Connect</li><li>- Google App Engine</li><li>- Engine Yard</li><li>- Caspio</li><li>- Qrimp</li><li>- MS Azure</li><li>- Mosso Cloud Sites</li><li>- VMforce</li><li>- Intuit Partner Platform</li><li>- Joyent Smart Platform</li></ul>	<ul style="list-style-type: none"><li>- Aster DB</li><li>- Quantivo</li><li>- Cloud9 Analytics</li><li>- K2 Analytics</li><li>- LogiXML</li><li>- Ooo</li><li>- PivotLink</li><li>- Clario Analytics</li><li>- ColdLight Neuron</li><li>- Vertica</li></ul>	<ul style="list-style-type: none"><li>- Amazon SQS</li><li>- Amazon SNS</li><li>- Boomi</li><li>- SnapLogic</li><li>- IBM Cast Iron</li><li>- gnip</li><li>- Apian Anywhere</li><li>- HubSpan</li><li>- Informatica On-Demand</li></ul>	<ul style="list-style-type: none"><li>- Keynote Systems</li><li>- SOASTA</li><li>- SkyTap</li><li>- Aptana</li><li>- LoadStorm</li><li>- Collabnet</li><li>- Rational Software Delivery Services</li></ul>	<ul style="list-style-type: none"><li>- Amazon SimpleDB</li><li>- Mosso Drizzle</li><li>- Amazon RDS</li></ul>

# What is PaaS?



Src: Securosis, L.L.C. / Cloud Security Alliance

# PaaS

## Benefits

- Packaged application “stack” reduces some complexity (configuration, components)
- If application vendor supports cloud APIs, streamlines implementation

## Issues

- Still responsible to keep stack updated
- Locked into providers API (which can change)
- Multi-tenancy at platform layer

Src: Securisis, L.L.C. / Cloud Security Alliance



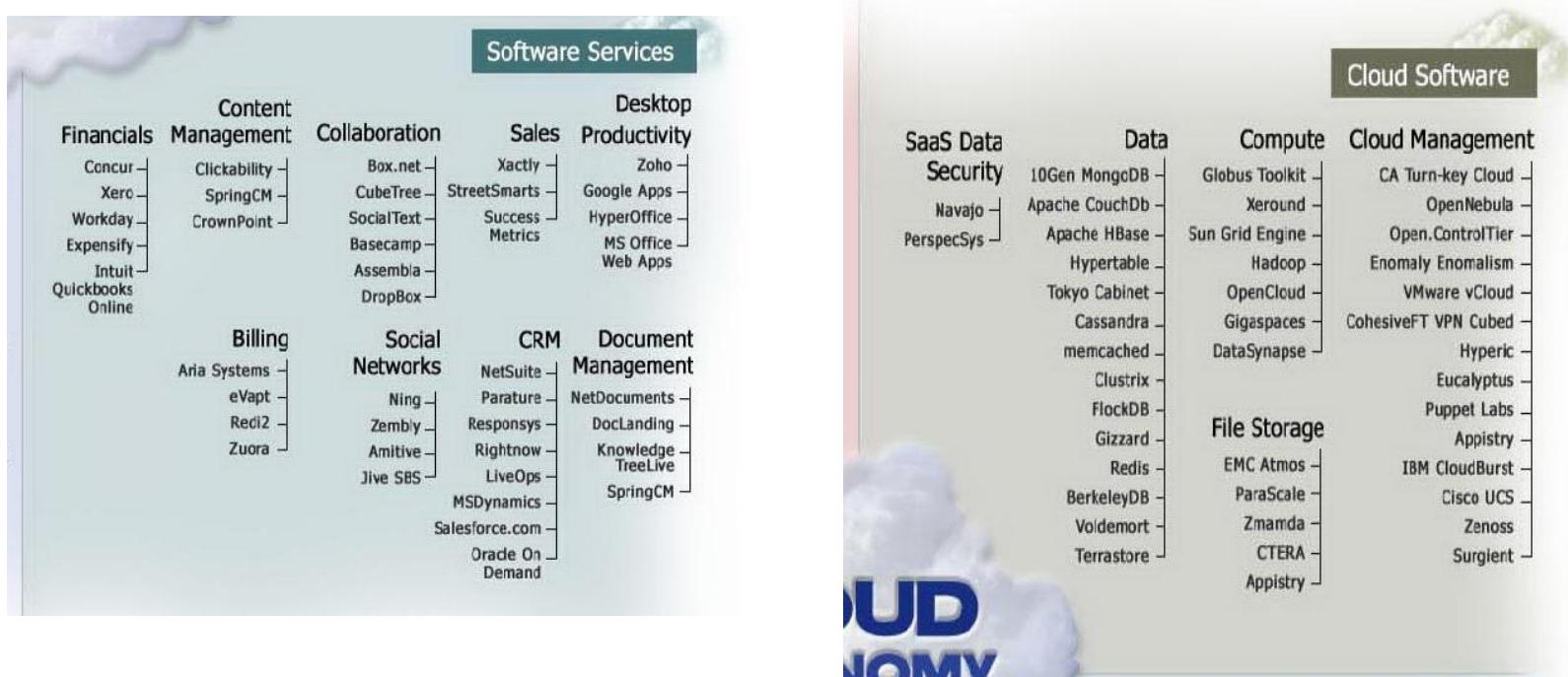
# Software as a Service SaaS

- Cloud computing services, such as Amazon's EC2 and Google Apps, are booming.
- With **Software as a Service**, you're not writing an app, just using someone else's.
- Changes the dynamic of pricing the software (pay on a per-use basis).
- 20% growth in SaaS products per year.

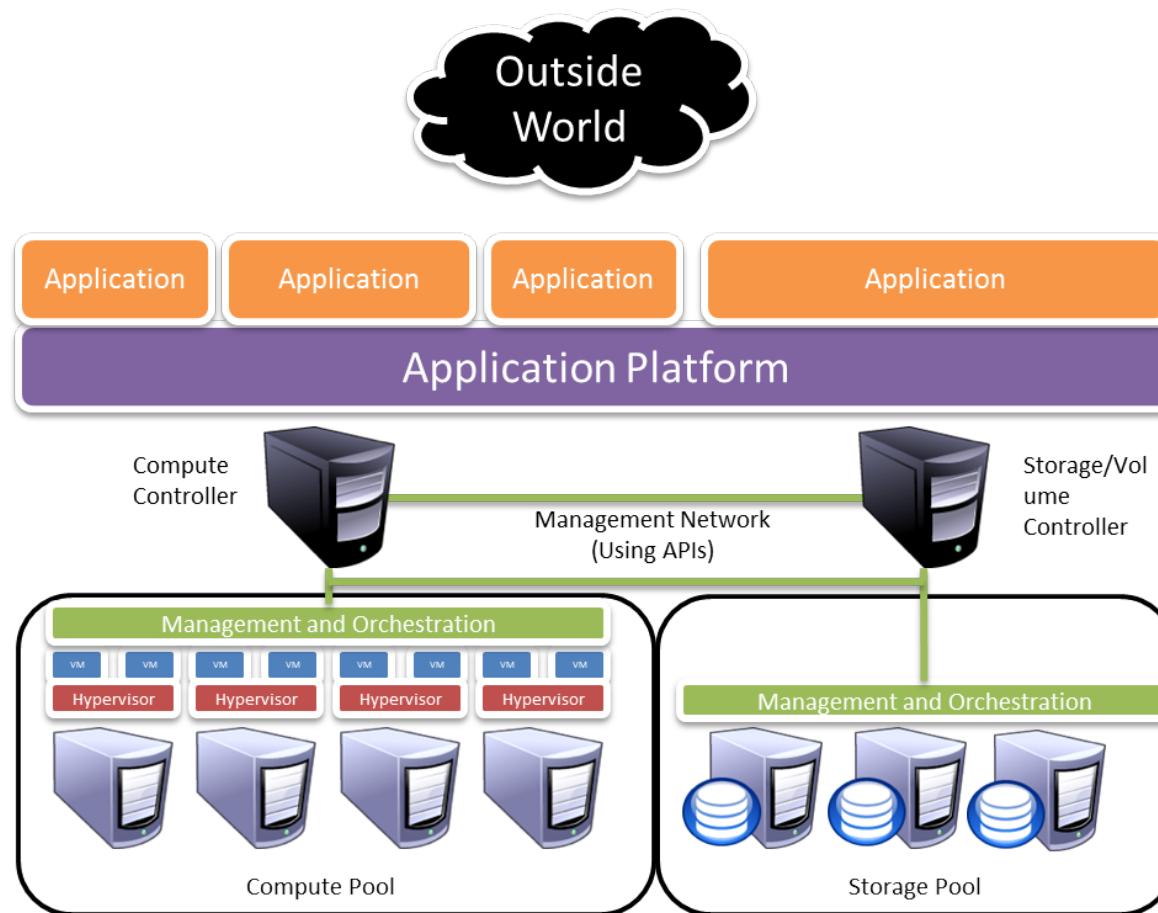


# Cloud Service Models – SaaS

Software and data are hosted on the cloud and are typically accessed by users using a thin client (browser with internet access)



# What is SaaS?



Src: Securosis, L.L.C. / Cloud Security Alliance

# SaaS

## Benefits

- Packaged solution reduces complexity
- Scaling environment isn't customer's problem.
- All updates/ configurations/security handled by provider.

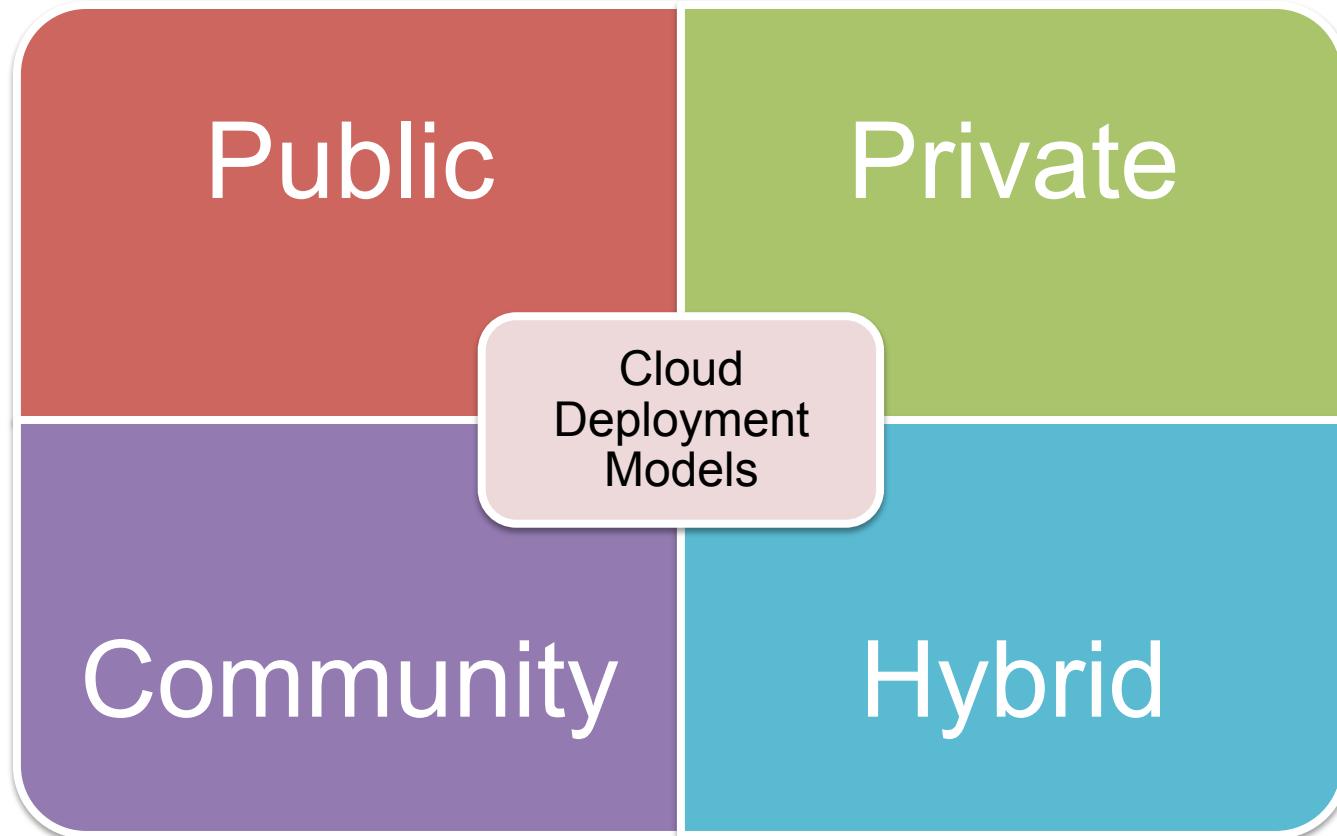
## Issues

- Very little app customization
- No control of components.
- No control of security (can only assess, not impact).
- Multi-tenancy issues at application layer.

Src: Securosis, L.L.C. / Cloud Security Alliance



# Cloud Deployment Models



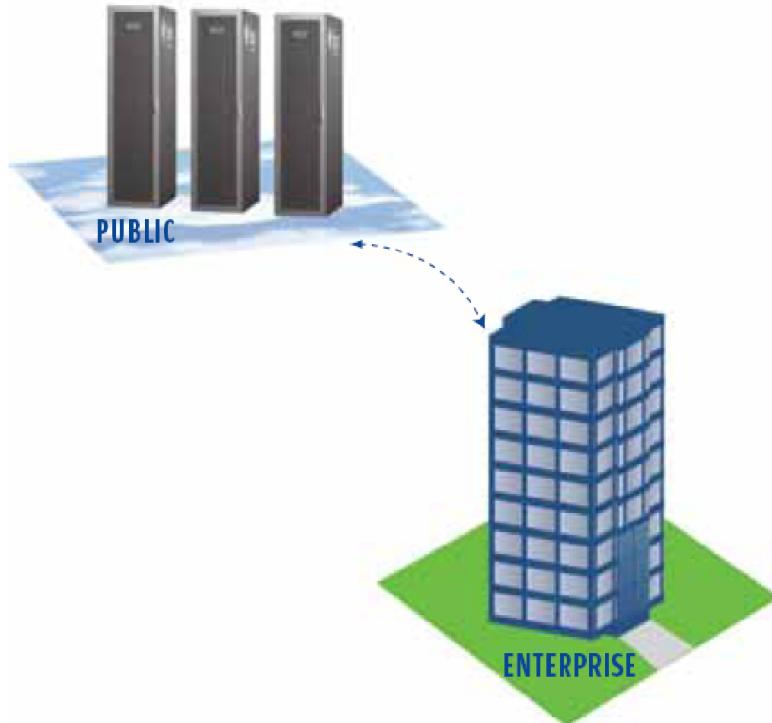
Src: Securosis, L.L.C. / Cloud Security Alliance



# Cloud Deployment Models

- Deployment Options
  - Private
  - Public
  - Community
  - Hybrid
- Controlled/Owned By
  - Internal
  - External

# Cloud Computing Infrastructures – Public Clouds



- Run by 3<sup>rd</sup> parties such as Amazon, Google or Microsoft.
- Employ statistical multiplexing to provide hardware and software resources.
- Are hosted away from user premises.
- For security, other applications running on the same clouds are transparent to cloud users.
- Public clouds guarantee improved performance, considerable & scalable resources, and growth flexibility.

# Public Cloud, Advantages, drawbacks

## Pros:

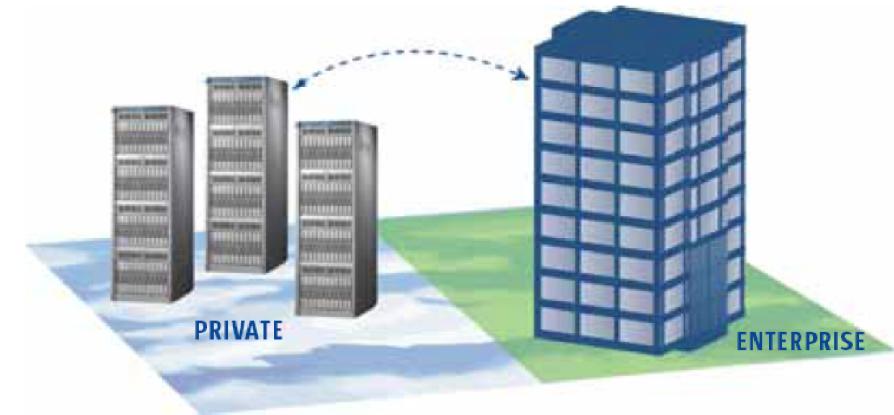
- Reliability
- Cost Efficiency
- Scalability and Agility

## Cons:

- Security
- Control

# Cloud Computing Infrastructures – Private Clouds

- Built for only one client.
- Provide complete control over data, security and QoS.
- Deployed on enterprise datacenter or co-location facility.



- Built by companies own IT organization or cloud service provider.
- Hosted private model- high level of control + technical expertise to establish and operate the cloud.

# Private Cloud, Advantages, drawbacks

 Pros:

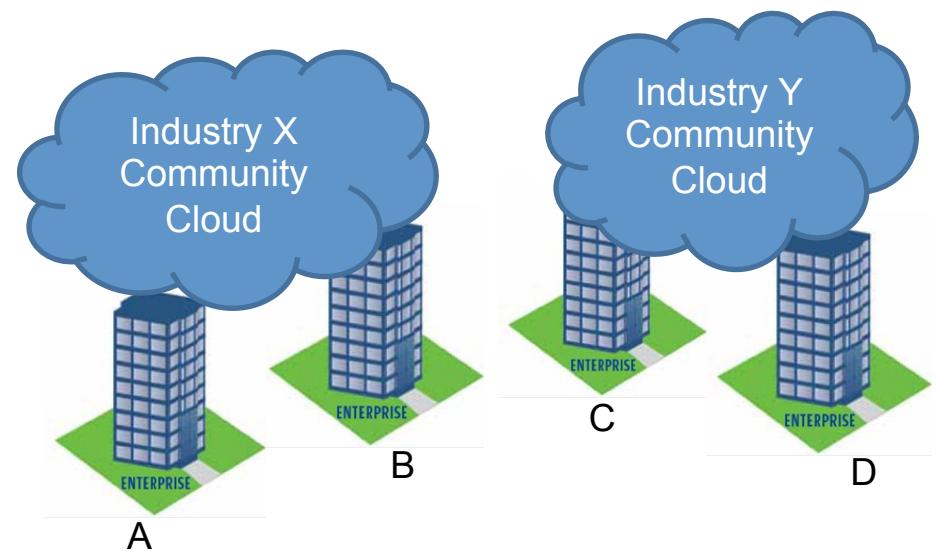
- Control / Security
- Availability
- Speed of Access

 Cons:

- Scalability
- Maintenance

# Community Cloud

- >In a community cloud Multiple organizations and infrastructures from the same community share the cloud infrastructure.
- They all have similar concerns and goals which helps to agree on the same cloud policies.



# Community Cloud, Advantages, drawbacks

## Pros

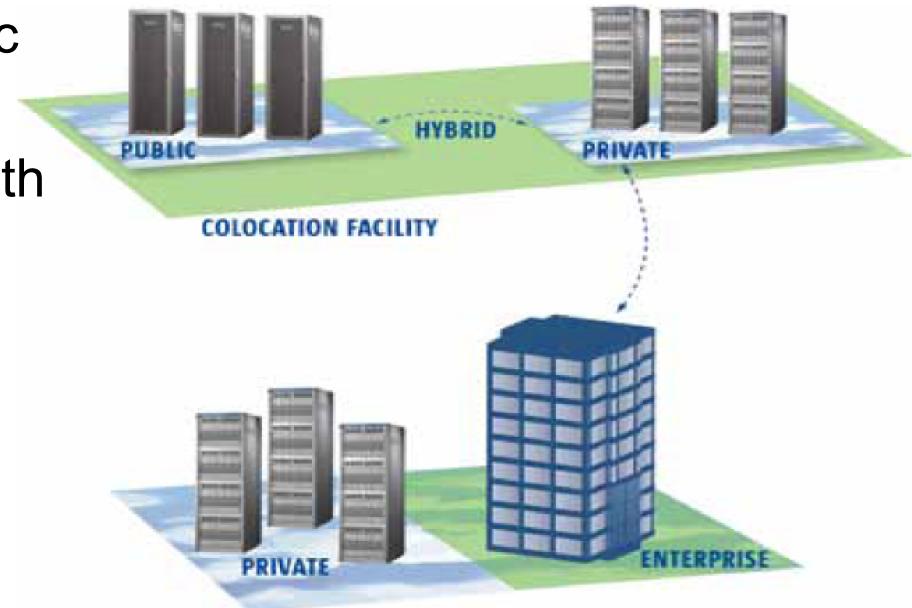
- Security
- Legal/compliance
- Same Policy and Concerns

## Cons

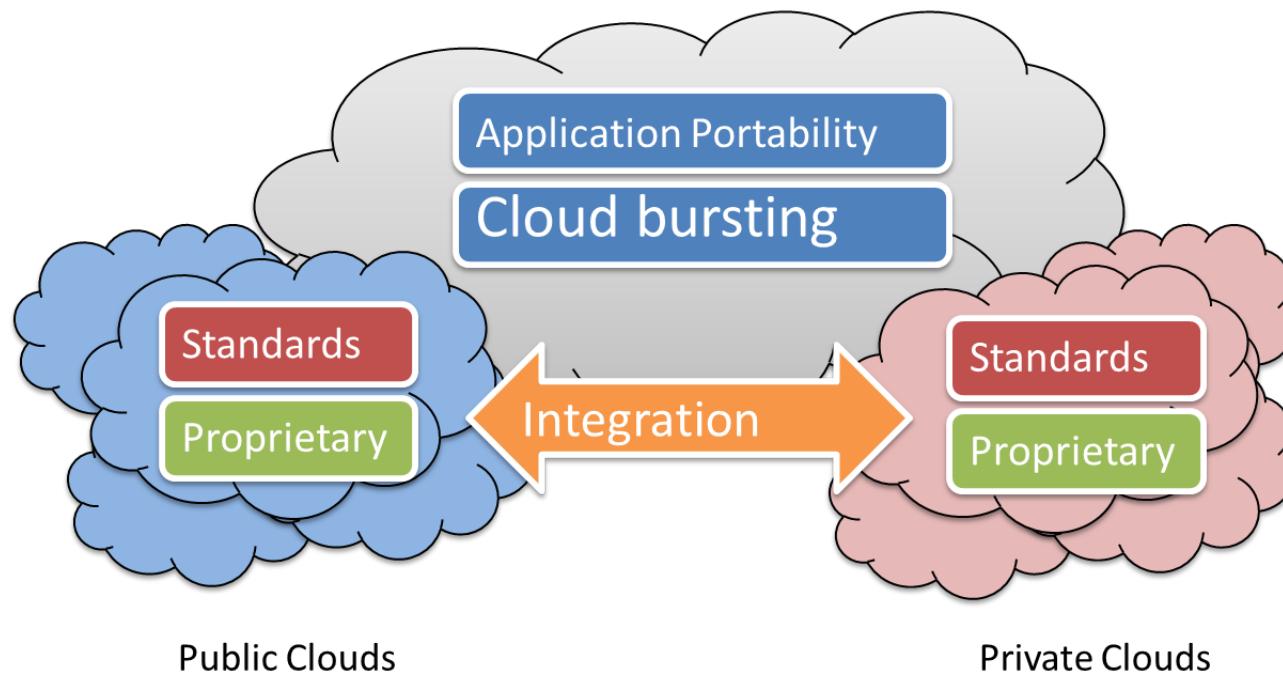
- Development
- Cost

# Cloud Computing Infrastructures – Hybrid Clouds

- Combines both private and public clouds.
- Private clouds are augmented with resources of public cloud.
- Are used to support Web 2.0 applications
- Also used to handle workload spikes, i.e. surge computing.
- More suitable for handling small data transfer or applications are stateless, than if large amount of data were transferred for small amount of processing.



# Hybrid Clouds



Source: Securosis, L.L.C. / Cloud Security Alliance

# Hybrid Cloud, Advantages, drawbacks

## Pros:

- High performance:
- Expanded capacity
- Scalability
- Security
- Low cost:

## Cons:

- Complex SLAs:
- Complex networking

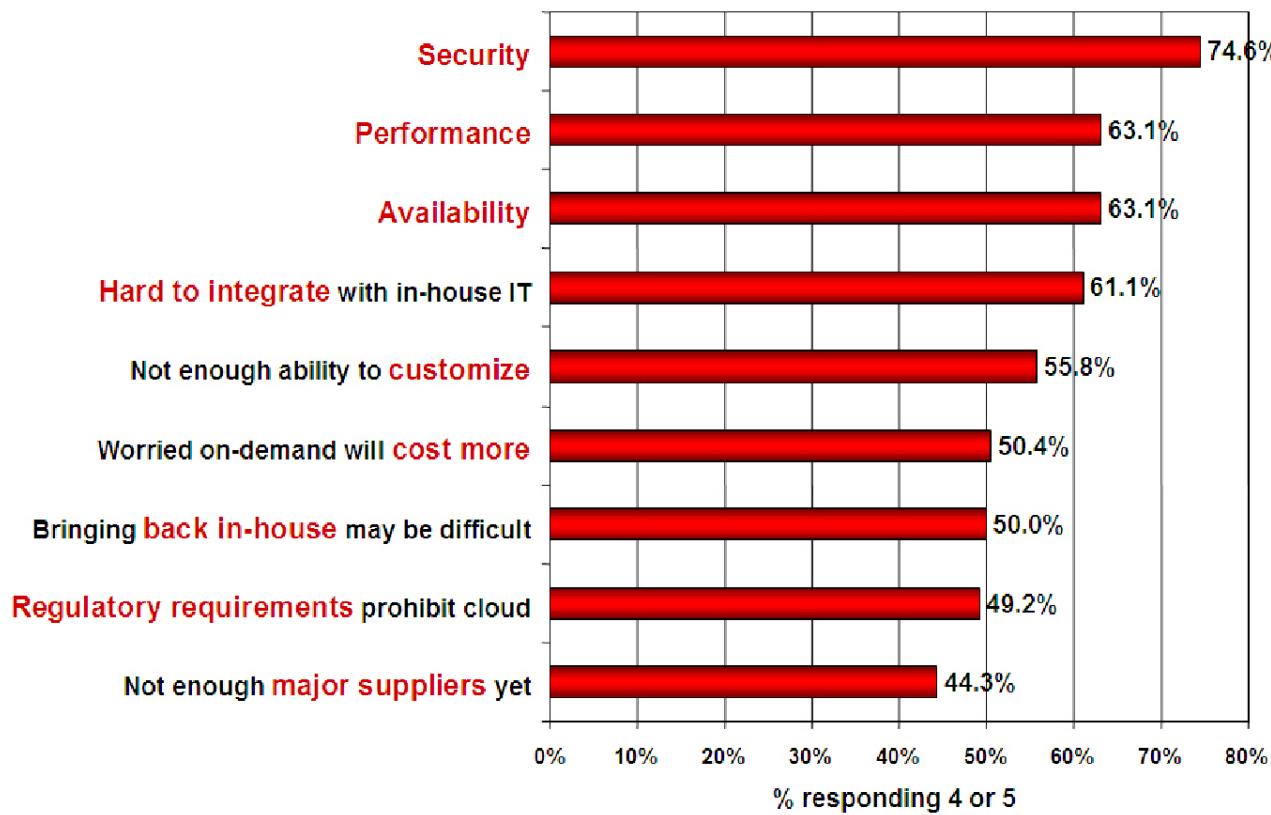
# ■ CLOUD SECURITY ISSUES

# If cloud computing is so great, why isn't everyone doing it?

- The cloud acts as a big black box, nothing inside the cloud is visible to the clients
- Clients have no idea or control over what happens inside a cloud
- Even if the cloud provider is honest, it can have malicious system admins who can tamper with the VMs and violate confidentiality and integrity
- Clouds are still subject to traditional data confidentiality, integrity, availability, and privacy issues, plus some additional attacks

# Companies are still afraid to use clouds

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

# Top Cyberattacks in 2014 so far!

- Analysts, Hold Security, startlingly announced in February that it had managed to obtain a list of 360 million account credentials for web services from the black market. That's just after three weeks of research.
- According to research from Arbor Networks, the number of DDoS events topping 20Gbps in the first half of 2014, are double that of 2013.
- Akamai Technologies State of the Internet report also showed that hacker attacks on websites went up 75% in the final quarter of 2013, with hackers in China responsible for 43% of all attacks
- This incredible [cybermap.kaspersky.com](http://cybermap.kaspersky.com)

interactive map from Antivirus software firm Kaspersky, which depicts all the current cyber attacks occurring around the world in real time, shows the growing intensity of hacks as the year progresses.

# Top Cyberattacks in 2014 - continue

- owl In May, eBay revealed that hackers had managed to steal personal records of 233 million users, with usernames, passwords, phone numbers and physical addresses compromised.
- owl **Community Health Services (health care).** The personal data for 4.5 million patients were compromised between April and June. The sophisticated malware used in the attack reportedly originated in China. (September 2014)
- owl **Google (communications).** Reportedly, 5 million Gmail usernames and passwords were compromised.<sup>[23]</sup> About 100,000 were released on a Russian forum site. (September 2014)
- owl **Apple iCloud (technology).** Hackers reportedly used passwords hacked with brute-force tactics and third-party applications to access Apple user's online data storage, leading to the subsequent posting of celebrities' private photos online. (September 2014)
- owl **J.P. Morgan Chase (financial).** The contact information for 76 million households and 7 million small businesses was compromised. The hackers may have originated in Russia and may have ties to the Russian government. (October 2014)

# Causes of Problems Associated with Cloud Computing

- Most security problems stem from:
  - Loss of control
  - Lack of trust (mechanisms)
  - Multi-tenancy
- These problems exist mainly in 3<sup>rd</sup> party management models
  - Self-managed clouds still have security issues, but not related to above

# Loss of Control in the Cloud

## owl Consumer's loss of control

- Data, applications, resources are located with provider
- User identity management is handled by the cloud
- User access control rules, security policies and enforcement are managed by the cloud provider
- Consumer relies on provider to ensure
  - Data security and privacy
  - Resource availability
  - Monitoring and repairing of services/resources

# Multi-tenancy Issues in the Cloud

- OWL Conflict between tenants' opposing goals
  - Tenants share a pool of resources and have opposing goals
  - OWL How does multi-tenancy deal with conflict of interest?
  - Can tenants get along together and 'play nicely' ?
  - If they can't, can we isolate them?
  - OWL How to provide separation between tenants?
  
- OWL Cloud Computing brings new threats
  - Multiple independent users share the same physical infrastructure
  - Thus an attacker can legitimately be in the same physical machine as the target

# Taxonomy of Fear

## Confidentiality

- Fear of loss of control over data
  - Will the sensitive data stored on a cloud remain confidential?
  - Will cloud compromises leak confidential client 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?

[www.cs.jhu.edu/~ragib/sp10/cs412](http://www.cs.jhu.edu/~ragib/sp10/cs412)

# Taxonomy of Fear (cont.)

## ❖ 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?
- Would cloud scale well-enough?

[www.cs.jhu.edu/~ragib/sp10/cs412](http://www.cs.jhu.edu/~ragib/sp10/cs412)



# Taxonomy of Fear (cont.)

- 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
- Increased attack surface
  - Entity outside the organization now stores and computes data, and so
  - Attackers can now target the communication link between cloud provider and client
  - Cloud provider employees can be phished

From [5] [www.cs.jhu.edu/~ragib/sp10/cs412](http://www.cs.jhu.edu/~ragib/sp10/cs412)



# Taxonomy of Fear (cont.)

- Auditability and forensics (out of control of data)
  - Difficult to audit data held outside organization in a cloud
  - Forensics also made difficult since now clients don't maintain data locally
- Legal quagmire and transitive trust issues
  - Who is responsible for complying with regulations?
    - e.g., SOX, HIPAA, PCI DSS ?
  - If cloud provider subcontracts to third party clouds, will the data still be secure?

[www.cs.jhu.edu/~ragib/sp10/cs412](http://www.cs.jhu.edu/~ragib/sp10/cs412)

# Cloud Computing: who should use it?

- Cloud computing definitely makes sense if your own security is weak, missing features, or below average.
- Ultimately, if
  - the cloud provider's security people are “better” than yours (and leveraged at least as efficiently),
  - the web-services interfaces don't introduce too many new vulnerabilities, and
  - the cloud provider aims at least as high as you do, at security goals,

then cloud computing has better security.

# ■ CLOUD ATTACK MECHANISMS

# Threat Model

- ❖ A threat model helps in analyzing a security problem, design mitigation strategies, and evaluate solutions

- ❖ Steps:

- Identify attackers, assets, threats and other components
- Rank the threats
- Choose mitigation strategies
- Build solutions based on the strategies

[www.cs.jhu.edu/~ragib/sp10/cs412](http://www.cs.jhu.edu/~ragib/sp10/cs412)

# Threat Model

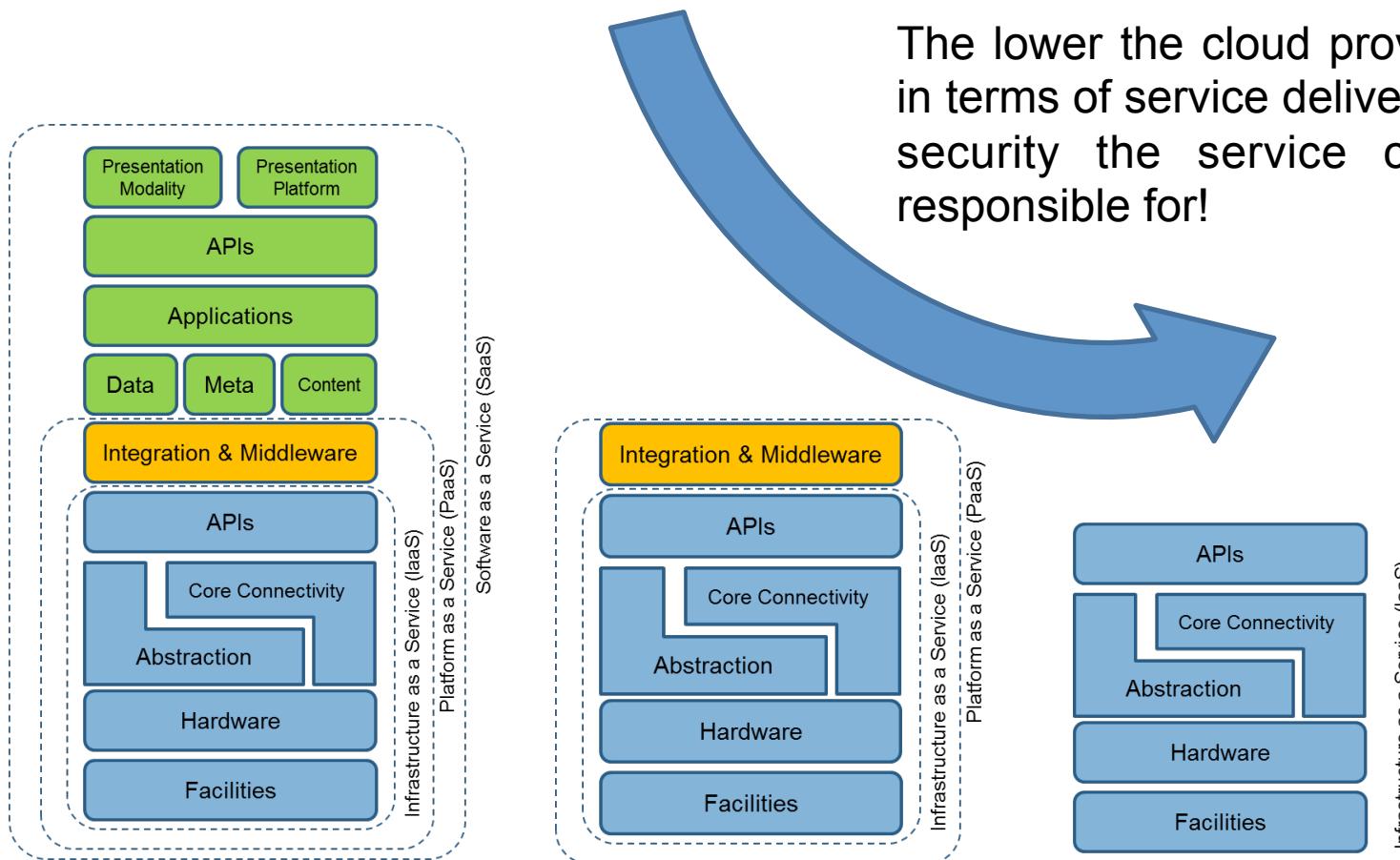
## Basic components

- Attacker modeling
  - Choose what attacker to consider
    - insider vs. outsider?
    - single vs. collaborator?
  - Attacker motivation and capabilities
- Attacker goals
- Vulnerabilities / threats

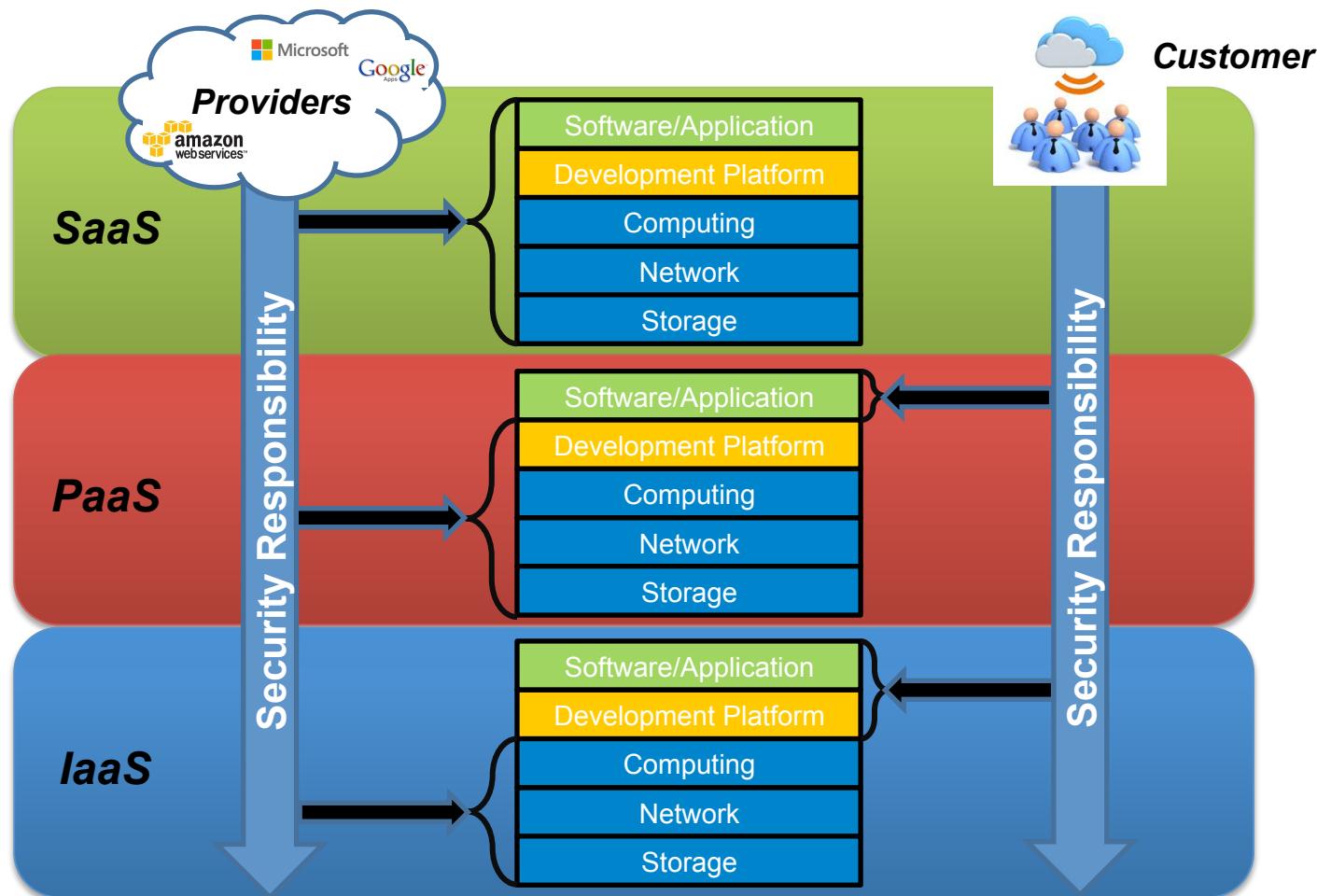
[www.cs.jhu.edu/~ragib/sp10/cs412](http://www.cs.jhu.edu/~ragib/sp10/cs412)



# Delivery model Security Issues



# Delivery model Security Issues



# Cloud Security Taxonomy

## Based on Service Models

### SaaS

- Cross Site Scripting
- Access Control Weaknesses
- SQL Injection Flaws
- Network Penetration
- Insecure SSL trust configuration
- Data Security

### PaaS

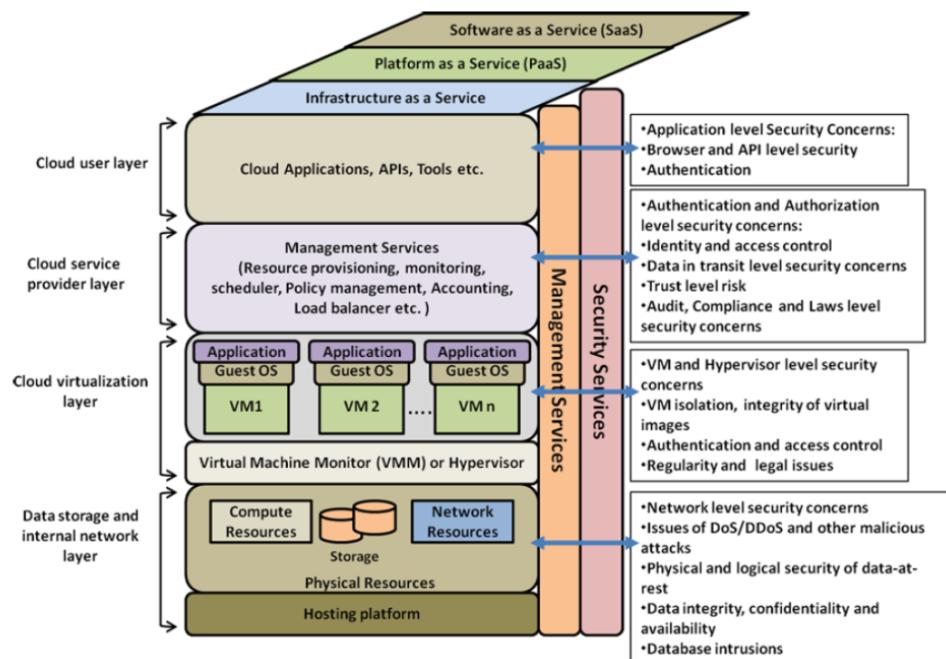
- Data Security Issues

### IaaS

- Data Reliability

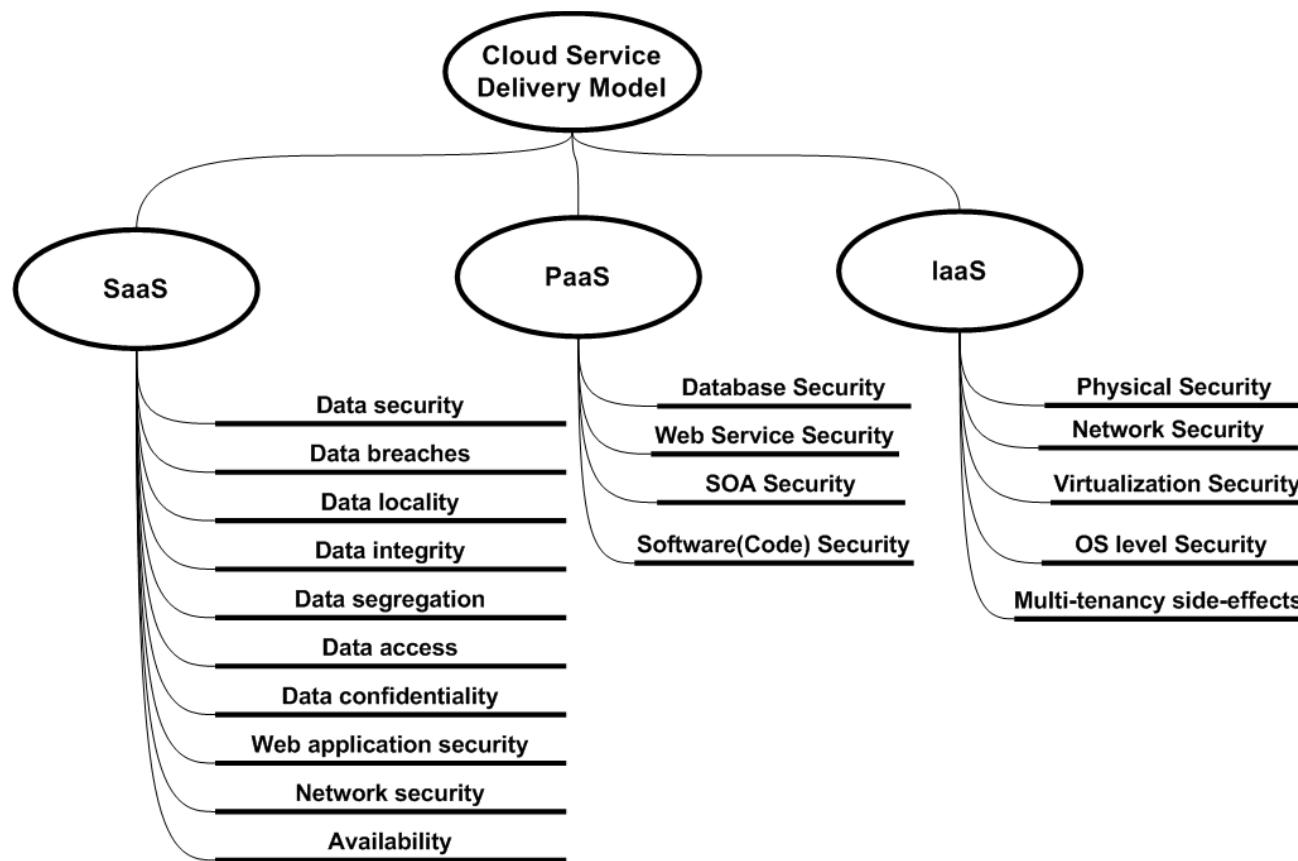
Source: V. S. Subashini, "A survey on security issues in service delivery models of cloud computing," *Journal of Network and Computer Applications*, vol. 34, pp. 1-11, 2011.

## Based on Layers

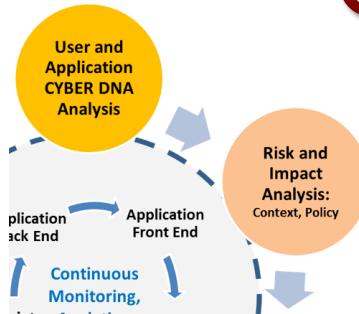


Source: C. Modi, D. Patel, B. Borisaniya, A. Patel and M. Rajarajan, "A survey on security issues and solutions at different layers of Cloud computing," *The Journal of Supercomputing*, pp. 1-32, 2012.

# Delivery model Security Issues



# Cloud Risk and Impact Analysis



Attacks	Attack Target	Impact
Cross Site Scripting SQL Injection Flaws Code malware, Worms Viruses, Flooding	<b>Cloud User Layer</b> Cloud Apps, APIs, Tools	Dollar and Reputation Loss, DoS, etc.
Identify and Access Control Insider Threats Insecure SSL trust configuration	<b>Cloud Service Provider Layer</b> Provisioning, Monitoring, Scheduling, Policy Management	Induced failures, availability, DoS, Inaccuracies, Reputation loss
VM and Hypervisor attacks VM integrity attacks Authentication and access control Backdoor implants	<b>Cloud Virtualization Layer</b> Applications, Guest OS, Virtual Machine Monitor,	VM privacy loss, Integrity of VM images
Dos/DDoS attacks Data integrity, confidentiality and availability Database intrusions	<b>Physical Resources Layer</b> Computing, Storage and Network Resources	Malfunctions, Data and Performance loss, Physical destruction

# The Notorious Nine

• The CSA(Cloud Security Alliance) has identified "The Notorious Nine", the top 9 cloud computing threats for 2013.

1. Data Breaches
2. Data Loss
3. Account Hijacking
4. Insecure APIs
5. Denial of Service
6. Malicious Insiders
7. Abuse of Cloud Services
8. Insufficient Due Diligence
9. Shared Technology Issues

# Data Breaches/Loss

- ❖ Deletion or alteration of records without a backup, Loss of an encoding key are some of the common examples which leads to data loss.
- ❖ As the data resides on the third parties data centers, security of data is becoming the main concern for cloud adoption.
- ❖ Thus it is the duty of Cloud security provider to prevent the unauthorized parties from gaining access to the sensitive data.

# Data Loss Remediation

- ❖ Implementing strong access controls
- ❖ Strong encryption and decryption for data.
- ❖ Implement strong key generation, storage and management, and destruction practices.
- ❖ Maintaining back up for the data and updating the changes timely.

# Data Breaches

TOP THREAT RANKING



## SERVICE MODEL

IaaS

PaaS

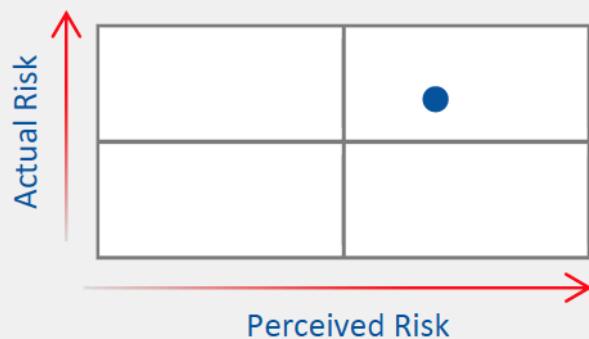
SaaS

## RISK ANALYSIS

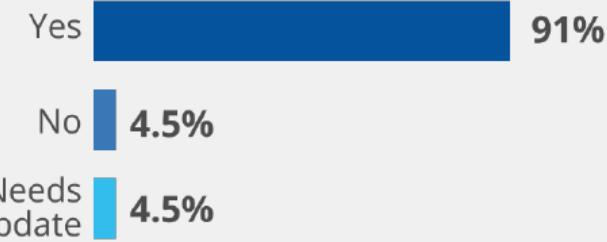
**CIANA:** Confidentiality

**STRIDE:** Information Disclosure

## RISK MATRIX



## IS THREAT STILL RELEVANT?



Source: [https://downloads.cloudsecurityalliance.org/initiatives/top\\_threats/The\\_Notorious\\_Nine\\_Cloud\\_Computing\\_Top\\_Threats\\_in\\_2013.pdf](https://downloads.cloudsecurityalliance.org/initiatives/top_threats/The_Notorious_Nine_Cloud_Computing_Top_Threats_in_2013.pdf)

# Taxonomy of Security

## CIA NA

- stands for Confidentiality, Integrity, Availability, Non-Repudiation, and Authentication (Information Assurance, Information Security)

## STRIDE is a system developed by Microsoft threat analysis:

- Spoofing of user identity
- Tampering
- Repudiation
- Information disclosure (privacy breach or data leak)
- Denial of service (D.o.S)
- Elevation of privilege

# Data Loss

TOP THREAT RANKING

5  
2010



2  
2013

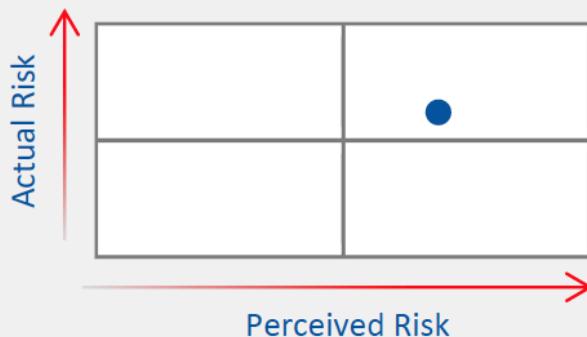
## SERVICE MODEL

IaaS

PaaS

SaaS

## RISK MATRIX

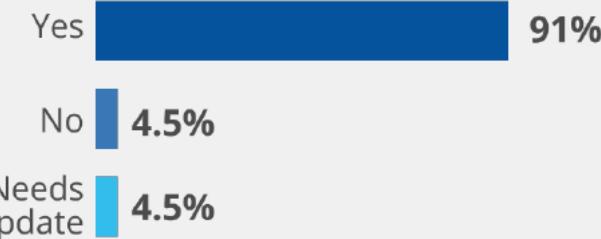


## RISK ANALYSIS

**CIANA:** Availability, Non-Repudiation

**STRIDE:** Repudiation, Denial of Service

## IS THREAT STILL RELEVANT?



Source: [https://downloads.cloudsecurityalliance.org/initiatives/top\\_threats/The\\_Notorious\\_Nine\\_Cloud\\_Computing\\_Top\\_Threats\\_in\\_2013.pdf](https://downloads.cloudsecurityalliance.org/initiatives/top_threats/The_Notorious_Nine_Cloud_Computing_Top_Threats_in_2013.pdf)

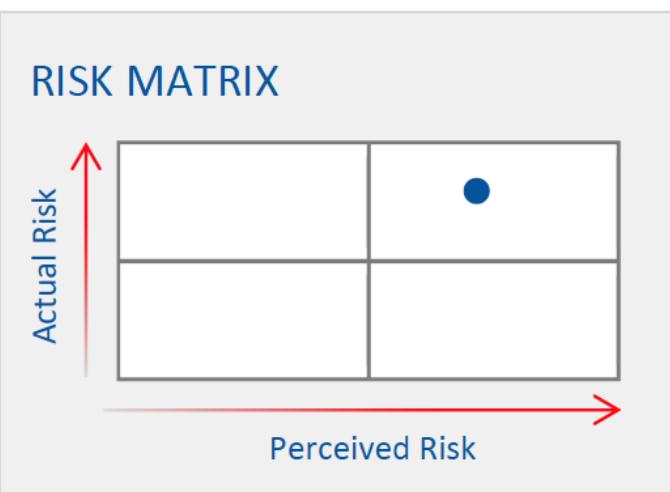
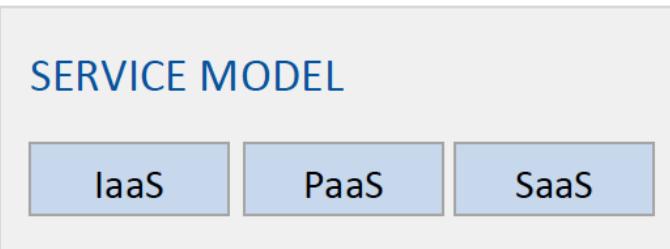
# Account, Service and Traffic Hijacking

- ❖ If an attacker gains access to the credentials, they can eavesdrop on your activities and transactions, manipulate data, return falsified information, and redirect your clients to illegitimate sites.
- ❖ Using the credentials and passwords for longer time without changing and reusing the same for different accounts makes this type of attack easy.

# Remediation

- Following the password rules to create strong passwords
- Changing the passwords timely
- Prohibiting the use of passwords on unknown machines and sharing of the passwords with other users

# Account or Service Traffic Hijacking

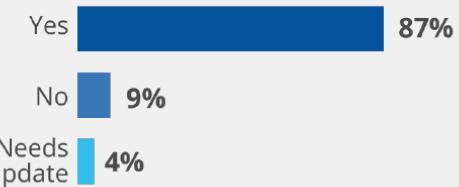


## RISK ANALYSIS

**CIANA:** Authenticity, Integrity, Confidentiality, Non-repudiation, Availability

**STRIDE:** Tampering with Data, Repudiation, Information Disclosure, Elevation of Privilege, Spoofing Identity

## IS THREAT STILL RELEVANT?



Source: [https://downloads.cloudsecurityalliance.org/initiatives/top\\_threats/The\\_Notorious\\_Nine\\_Cloud\\_Computing\\_Top\\_Threats\\_in\\_2013.pdf](https://downloads.cloudsecurityalliance.org/initiatives/top_threats/The_Notorious_Nine_Cloud_Computing_Top_Threats_in_2013.pdf)

# Insecure APIs

- ▀ The security of the cloud services is dependent on how secure is their API's
- ▀ accidental and malicious attempts must be taken into consideration when designing the APIs
- ▀ Organizations are facing a variety of authenticity, confidentiality, and integrity, issues due to their dependence on a weak set of APIs

# Insecure APIs



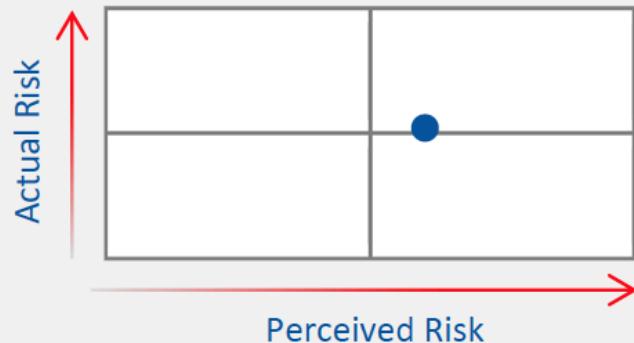
## SERVICE MODEL

IaaS

PaaS

SaaS

## RISK MATRIX

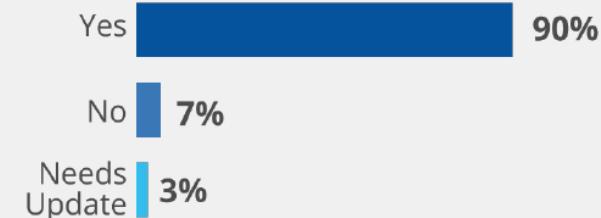


## RISK ANALYSIS

**CIANA:** Authenticity, Integrity, Confidentiality

**STRIDE:** Tampering with Data, Repudiation, Information Disclosure, Elevation of Privilege

## IS THREAT STILL RELEVANT?



Source: [https://downloads.cloudsecurityalliance.org/initiatives/top\\_threats/The\\_Notorious\\_Nine\\_Cloud\\_Computing\\_Top\\_Threats\\_in\\_2013.pdf](https://downloads.cloudsecurityalliance.org/initiatives/top_threats/The_Notorious_Nine_Cloud_Computing_Top_Threats_in_2013.pdf)

# Remediation

- 🐺 Analyze the security model of cloud provider interfaces.
- 🐺 Ensure strong authentication and access controls are implemented in concert with encrypted transmission.

# Denial of Service

- ▀ Preventing users from accessing cloud services.
- ▀ Using resource exhaustion attacks or software vulnerability attacks.
- ▀ The cloud becomes irresponsible or legal users will pay more for using more resources.

# Denial of Service

TOP THREAT RANKING

N/A  
2010

5  
2013

## SERVICE MODEL

IaaS

PaaS

SaaS

## RISK ANALYSIS

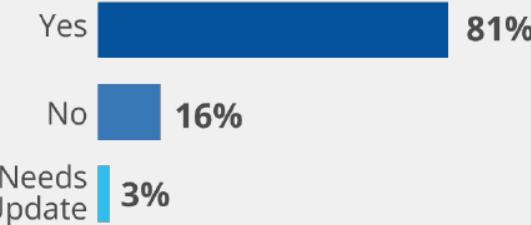
CIANA: Availability

STRIDE: Denial of Service

## RISK MATRIX



## IS THREAT STILL RELEVANT?



Source: [https://downloads.cloudsecurityalliance.org/initiatives/top\\_threats/The\\_Notorious\\_Nine\\_Cloud\\_Computing\\_Top\\_Threats\\_in\\_2013.pdf](https://downloads.cloudsecurityalliance.org/initiatives/top_threats/The_Notorious_Nine_Cloud_Computing_Top_Threats_in_2013.pdf)

# Remediation

- None is provided by CSA
- Anomaly Behavior Analysis (ABA)
- Intrusion Tolerance by using diversity and redundancy

# Malicious Insiders

- ❖ Malicious insider threat is well-known to most organizations.
- ❖ A provider may not reveal how it grants employees access to physical and virtual assets, how it monitors these employees, or how it analyzes and reports on policy compliance.
- ❖ This kind of situation clearly creates an attractive opportunity for hobbyist hacker.

# Malicious Insiders

TOP THREAT RANKING

3  
2010



6  
2013

## SERVICE MODEL

IaaS

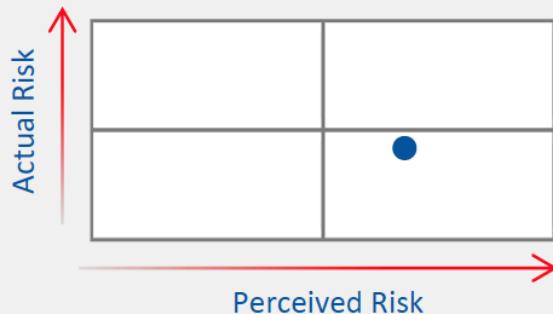
PaaS

SaaS

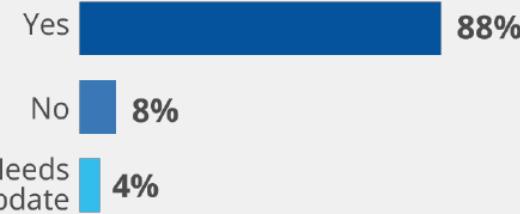
## RISK ANALYSIS

**STRIDE:** Spoofing, Tampering,  
Information Disclosure

## RISK MATRIX



## IS THREAT STILL RELEVANT?



Source: [https://downloads.cloudsecurityalliance.org/initiatives/top\\_threats/The\\_Notorious\\_Nine\\_Cloud\\_Computing\\_Top\\_Threats\\_in\\_2013.pdf](https://downloads.cloudsecurityalliance.org/initiatives/top_threats/The_Notorious_Nine_Cloud_Computing_Top_Threats_in_2013.pdf)

# Remediation

- ❖ Human resource required specifications should be part of legal contract.
- ❖ Cloud Service Provider should provide transparently all security and management practices.

# Abuse of Cloud Services

- ❖ The registration process for cloud resources has become so easy that anyone with a valid credit card can register and immediately begin using services.
- ❖ Thus, spammers, malicious code authors, and other criminals have been able to conduct their activities with relative impunity
- ❖ Thus PaaS and IaaS providers are suffering from these kind of attacks.

# Abuse of Cloud Services



## SERVICE MODEL

IaaS

PaaS

SaaS

## RISK ANALYSIS

CIANA: N/A

STRIDE: N/A

## RISK MATRIX

N/A

## IS THREAT STILL RELEVANT?

Yes  84%

No  14%

Needs Update  2%

Source: [https://downloads.cloudsecurityalliance.org/initiatives/top\\_threats/The\\_Notorious\\_Nine\\_Cloud\\_Computing\\_Top\\_Threats\\_in\\_2013.pdf](https://downloads.cloudsecurityalliance.org/initiatives/top_threats/The_Notorious_Nine_Cloud_Computing_Top_Threats_in_2013.pdf)

# Impact

- ❖ Attackers are coming up with new technologies to improve their reach, avoid detection and improve the effectiveness of their activities.
- ❖ The reasons for this type of attacks are:
  - Weak registration systems that are facilitating the anonymity.
  - Limited capabilities of service providers to fraud detection capabilities

# Remediation

- ❖ Strict initial registration and validation
- ❖ Enhanced credit card fraud monitoring and coordination
- ❖ Constant monitoring of customer network traffic.
- ❖ Monitoring public blacklists for one's own network blocks.

# Insufficient Due Diligence

- ❖ Organizations moving fast toward the cloud for its cost reductions, operational efficiencies and improved security.
- ❖ However, without a full understanding of the cloud service provider environment and responsibilities, they are increasing their risk.

# Insufficient Due Diligence



## SERVICE MODEL

IaaS

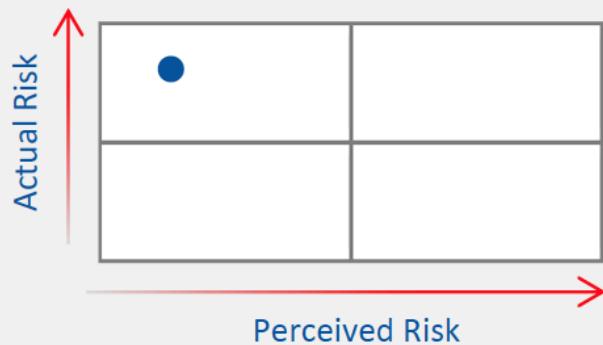
PaaS

SaaS

## RISK ANALYSIS

STRIDE: All

## RISK MATRIX



## IS THREAT STILL RELEVANT?



Source: [https://downloads.cloudsecurityalliance.org/initiatives/top\\_threats/The\\_Notorious\\_Nine\\_Cloud\\_Computing\\_Top\\_Threats\\_in\\_2013.pdf](https://downloads.cloudsecurityalliance.org/initiatives/top_threats/The_Notorious_Nine_Cloud_Computing_Top_Threats_in_2013.pdf)

# Remediation

- » Organizations need to understand the risk of moving to the cloud.
- » 24/7 Continuous Monitoring, Analysis, and Mitigation

# Shared Technology Vulnerabilities

- Cloud Service Providers deliver their services in a scalable way by sharing infrastructure.
- Cloud services depend on utilizing virtualization.
- Virtualization Hypervisors, like any other software, have flaws that allow attackers with access to the guest operating system to attack the host.
- This impacts the operations of other cloud customers and allow attackers to gain access to unauthorized data.

# Shared Technology Issues



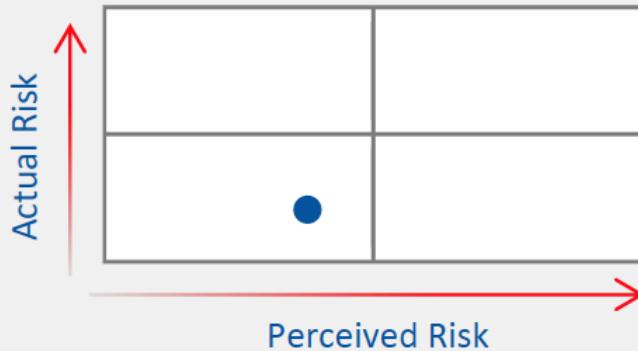
## SERVICE MODEL

IaaS

PaaS

SaaS

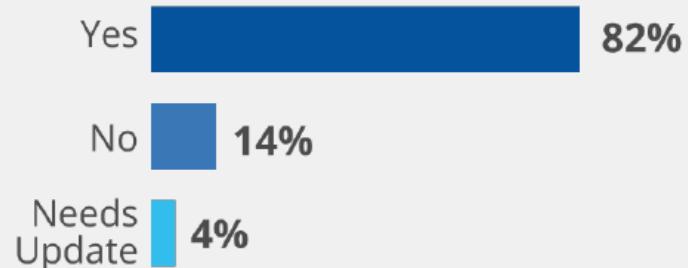
## RISK MATRIX



## RISK ANALYSIS

**STRIDE:** Information Disclosure,  
Elevation of Privilege

## IS THREAT STILL RELEVANT?



Source: [https://downloads.cloudsecurityalliance.org/initiatives/top\\_threats/The\\_Notorious\\_Nine\\_Cloud\\_Computing\\_Top\\_Threats\\_in\\_2013.pdf](https://downloads.cloudsecurityalliance.org/initiatives/top_threats/The_Notorious_Nine_Cloud_Computing_Top_Threats_in_2013.pdf)

# Remediation

- ❖ Implementing and applying security best practices for both the installation and configuration processes
- ❖ Continuously monitoring for the environment to detect unauthorized activities.
- ❖ Enforcing strict access control and strong authentication for all critical operations.
- ❖ Continuously searching for vulnerabilities and threats.

# Unknown risk Profile

- ❖ The features and functionality of the cloud services are well informed to the customer, but the details of internal security procedures, auditing, logging, internal access control remains unanswered leaving customers with an unknown risk profile

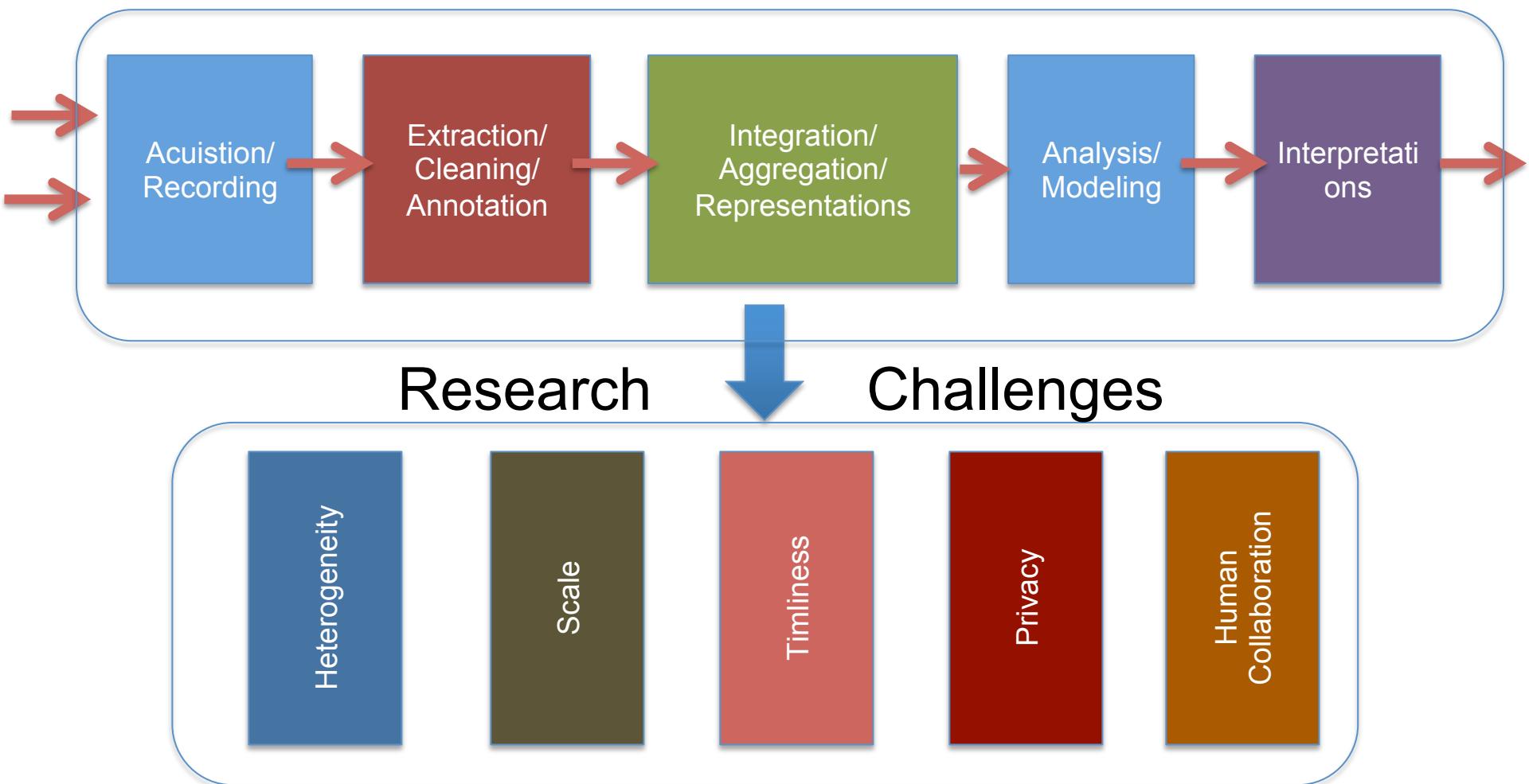
# General Security Issues

- OWL In addition to the above mentioned top threats there are many other threats that are effecting cloud computing. They are:
  - OWL Insider Threats
  - OWL Hypervisor vulnerabilities
  - OWL Denial of Service attacks
  - OWL Malware Injection attacks
  - OWL Man-In-The Middle Cryptographic attacks

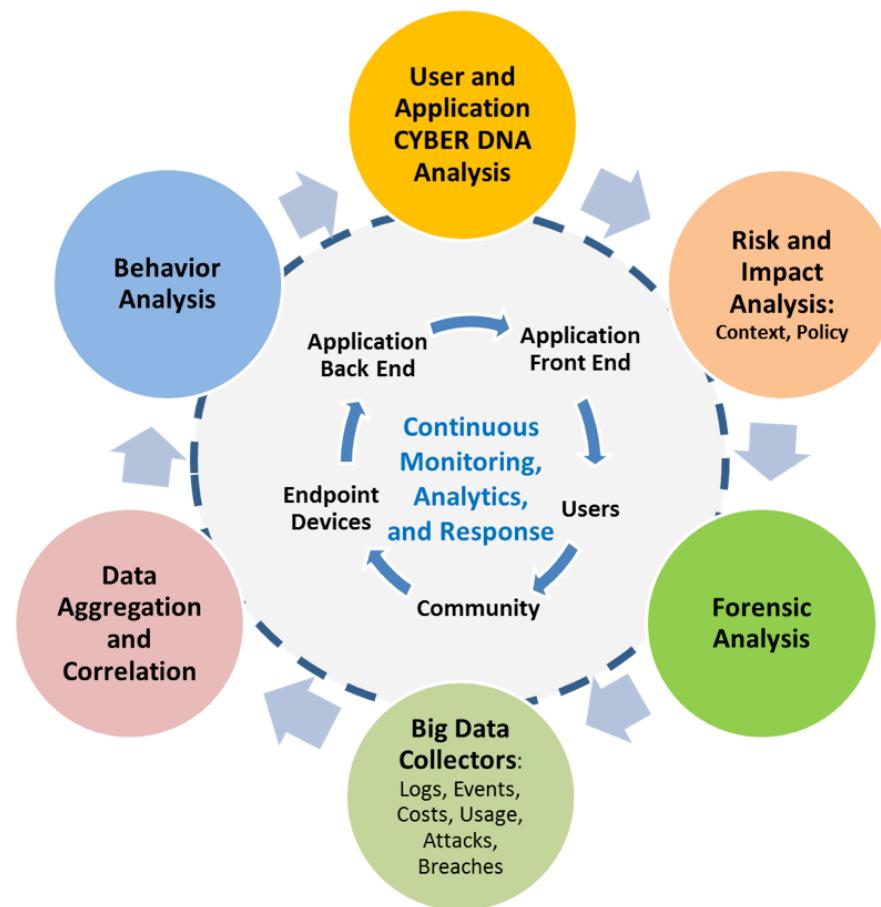
# UA Ongoing Cybersecurity Research Projects



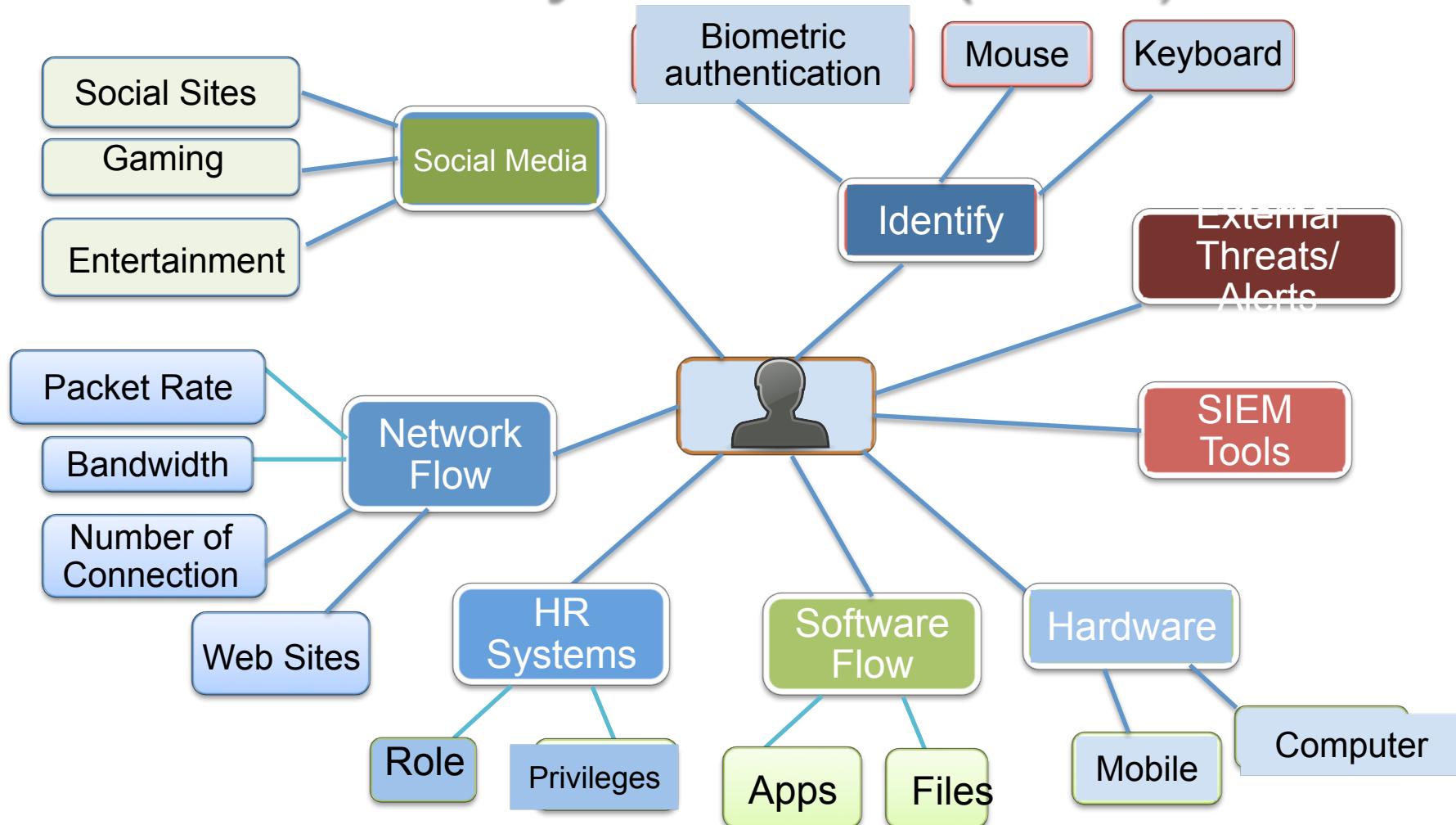
# Big Data Analytics Pipeline



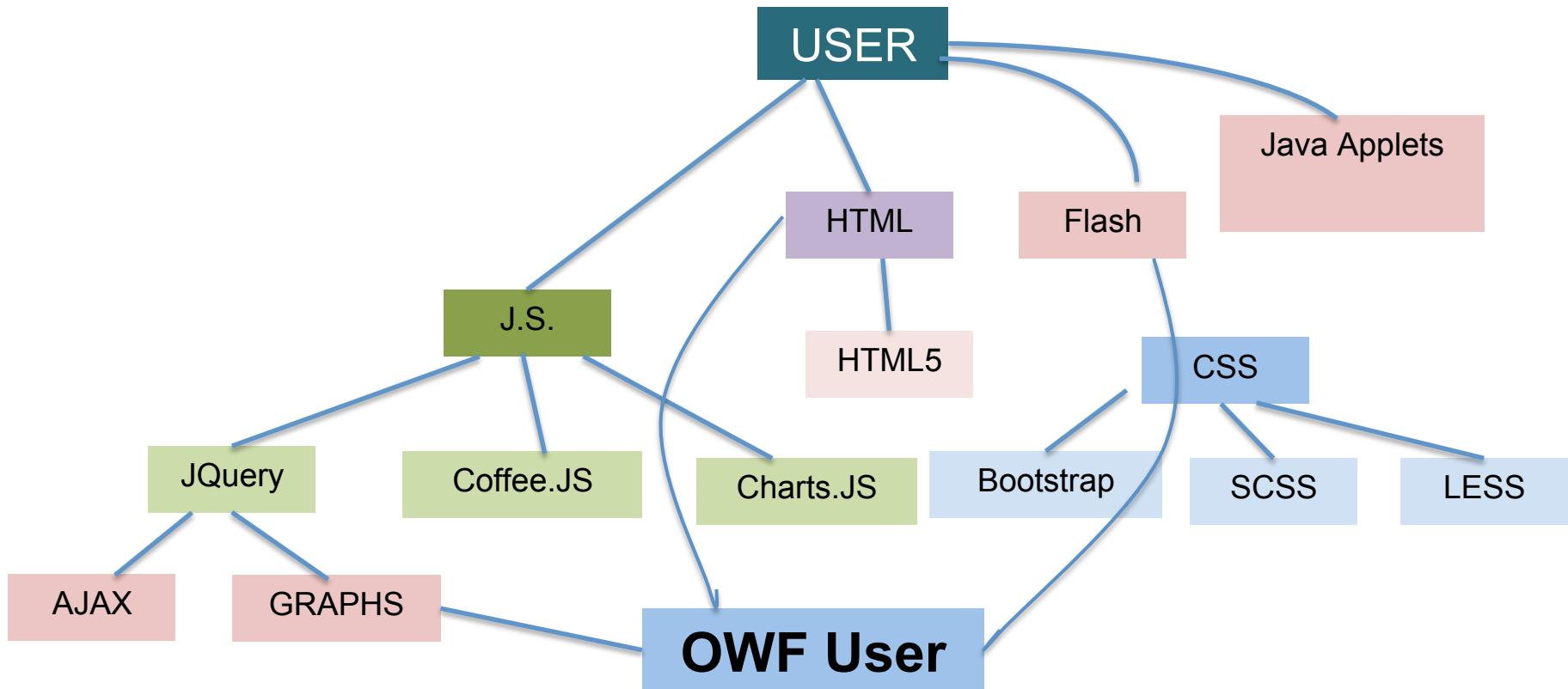
# Big Data Analytics for Cybersecurity Architecture



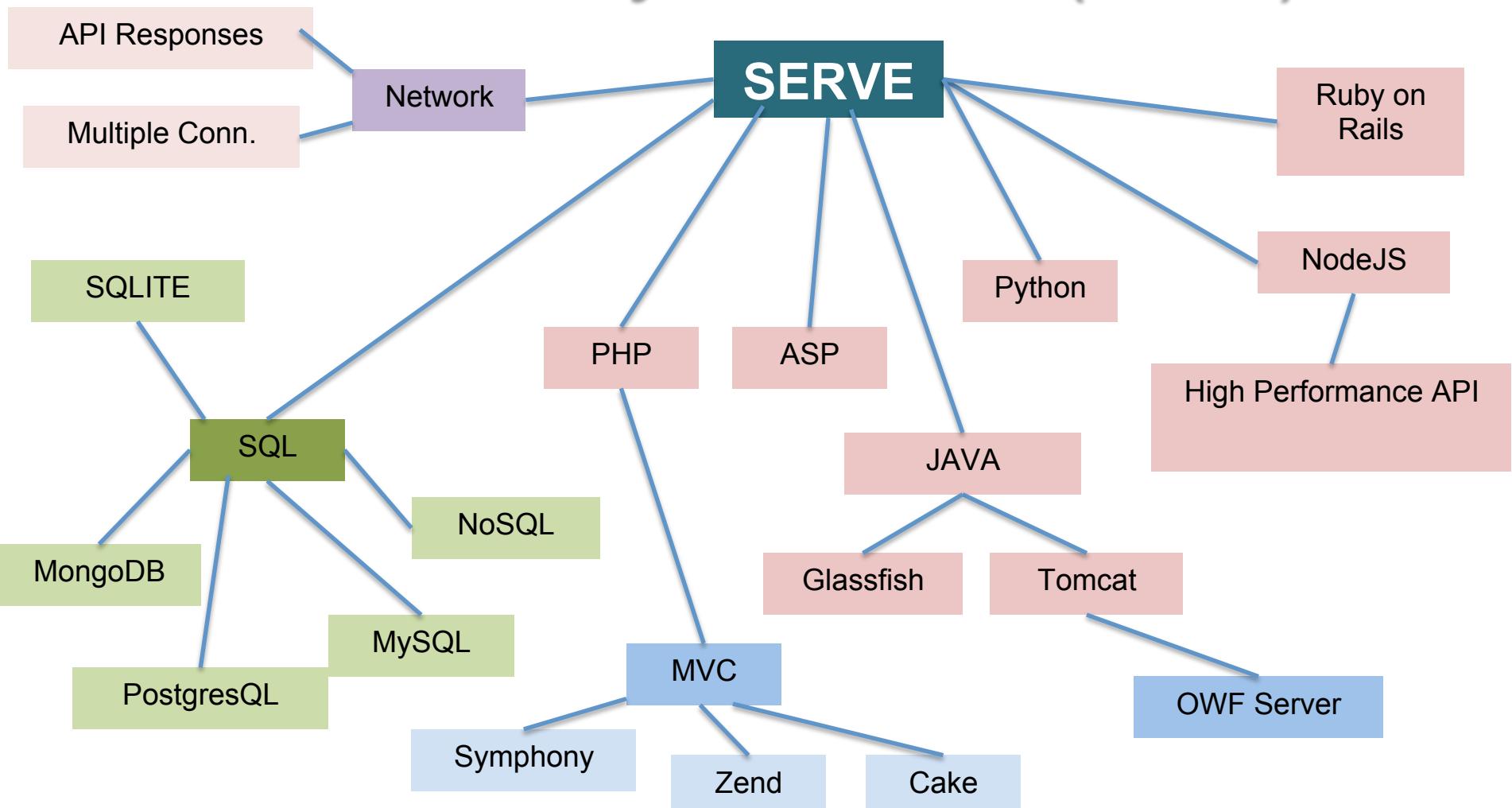
# User Cyber Flow (UCF)



# Application-Cyber Flow (ACF)

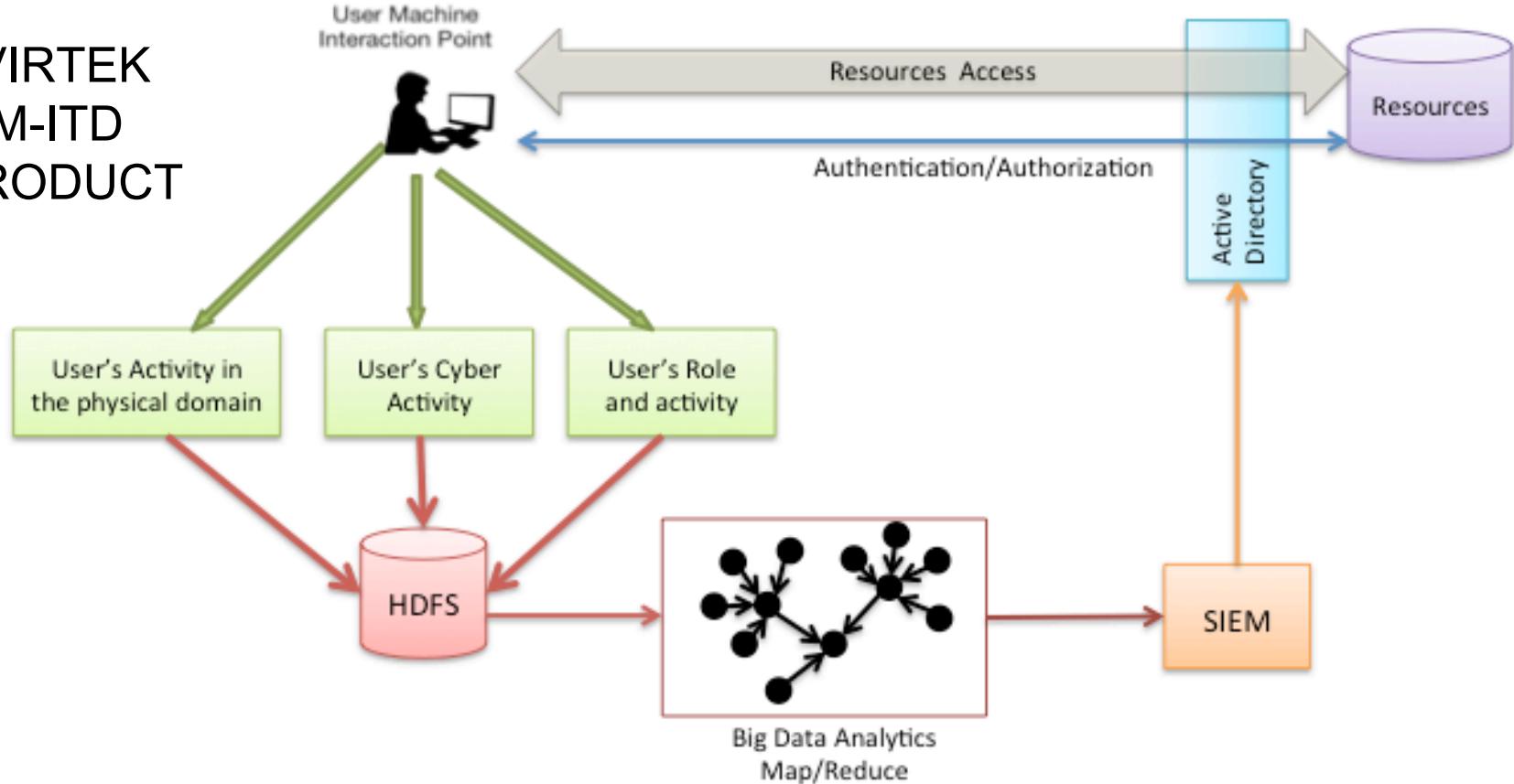


# Server-Cyber Flow (SCF)



# Insider Threat Detection (ITD) with Smart Big Data Analytics

AVIRTEK  
AIM-ITD  
PRODUCT



# Conclusion

- Cloud computing is sometimes viewed as a reincarnation of the classic mainframe client-server model
  - However, resources are ubiquitous, scalable, highly virtualized
  - Contains all the traditional threats, as well as new ones
- In developing solutions to cloud computing security issues it may be helpful to identify the problems and approaches in terms of
  - Loss of control
  - Lack of trust
  - Multi-tenancy problems



Thank You



# Hadoop File System

1

# Reference

2

- The Hadoop Distributed File System: Architecture and Design by Apache Foundation Inc.

# Basic Features: HDFS

3

- Highly fault-tolerant
- High throughput
- Suitable for applications with large data sets
- Streaming access to file system data
- Can be built out of commodity hardware

# Fault tolerance

4

- Failure is the norm rather than exception
- A HDFS instance may consist of thousands of server machines, each storing part of the file system's data.
- Since we have huge number of components and that each component has non-trivial probability of failure means that there is always some component that is non-functional.
- Detection of faults and quick, automatic recovery from them is a core architectural goal of HDFS.

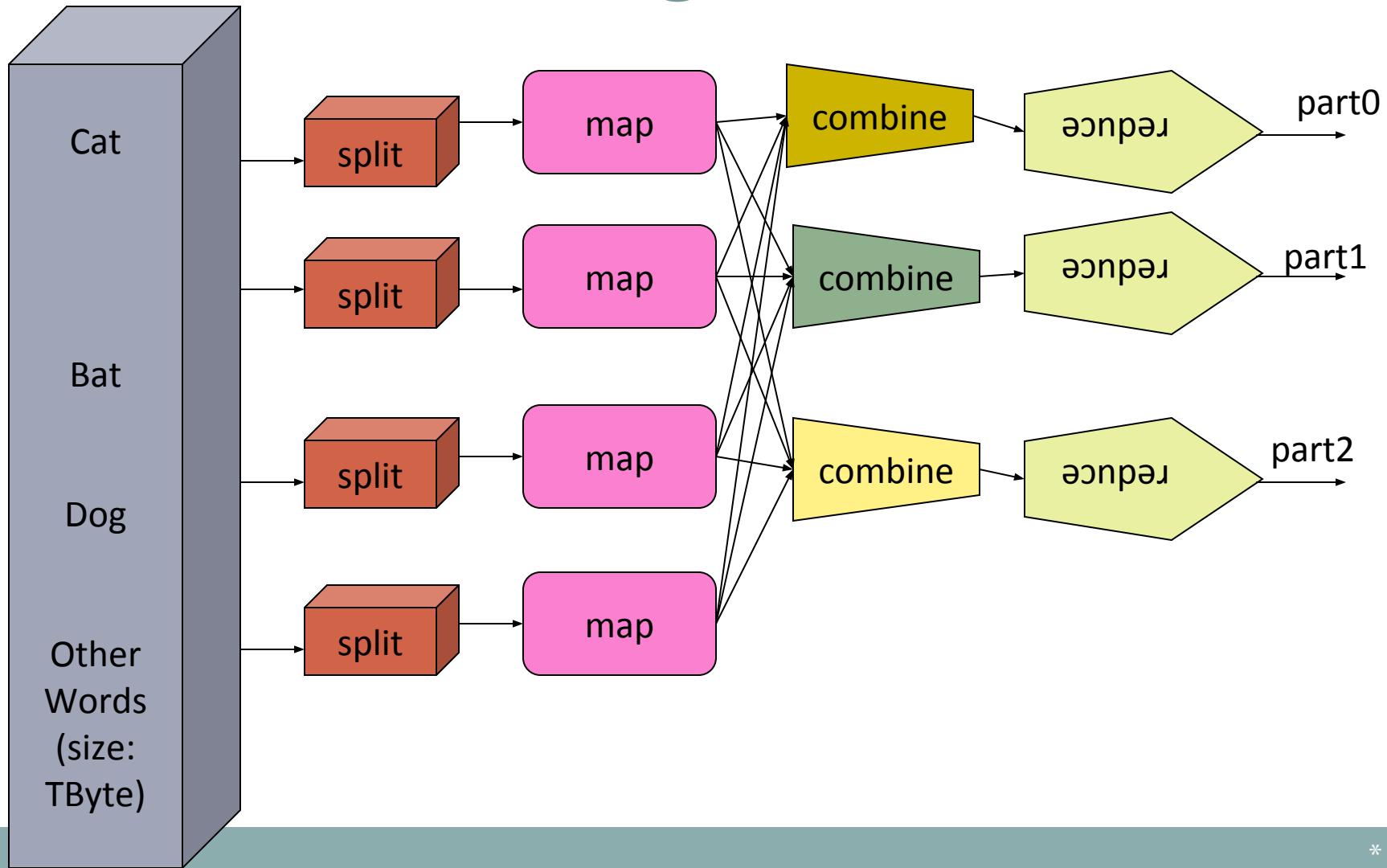
# Data Characteristics

5

- Streaming data access
- Applications need streaming access to data
- Batch processing rather than interactive user access.
- Large data sets and files: gigabytes to terabytes size
- High aggregate data bandwidth
- Scale to hundreds of nodes in a cluster
- Tens of millions of files in a single instance
- Write-once-read-many: a file once created, written and closed need not be changed – this assumption simplifies coherency
- A map-reduce application or web-crawler application fits perfectly with this model.

# MapReduce

6



# Architecture

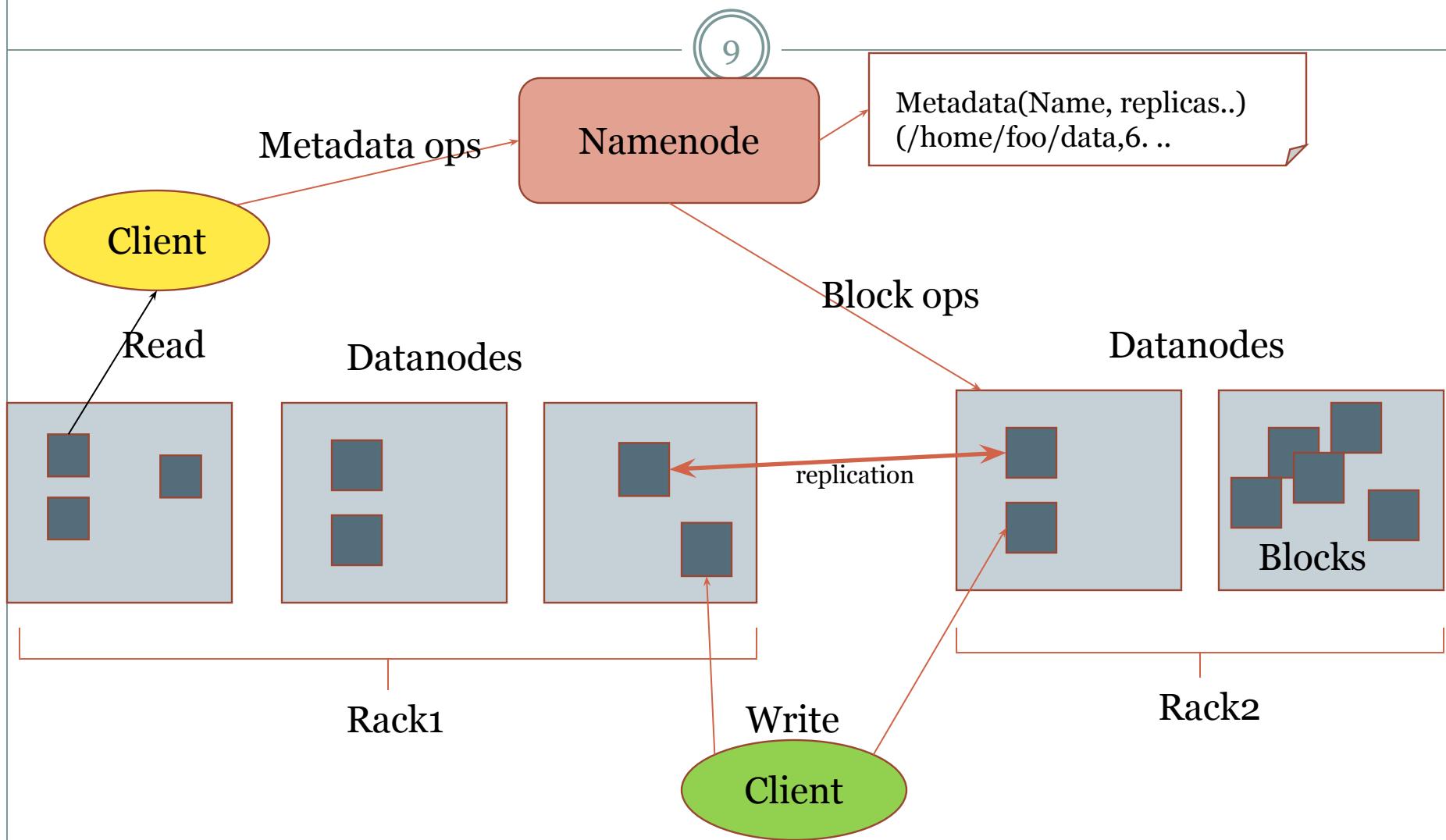
7

# Namenode and Datanodes

8

- Master/slave architecture
- HDFS cluster consists of a single **Namenode**, a master server that manages the file system namespace and regulates access to files by clients.
- There are a number of **DataNodes** usually one per node in a cluster.
- The DataNodes manage storage attached to the nodes that they run on.
- HDFS exposes a file system namespace and allows user data to be stored in files.
- A file is split into one or more blocks and set of blocks are stored in DataNodes.
- DataNodes: serves read, write requests, performs block creation, deletion, and replication upon instruction from Namenode.

# HDFS Architecture



# File system Namespace

10

- Hierarchical file system with directories and files
- Create, remove, move, rename etc.
- Namenode maintains the file system
- Any meta information changes to the file system recorded by the Namenode.
- An application can specify the number of replicas of the file needed: replication factor of the file. This information is stored in the Namenode.

# Data Replication

11

- HDFS is designed to store very large files across machines in a large cluster.
- Each file is a sequence of blocks.
- All blocks in the file except the last are of the same size.
- Blocks are replicated for fault tolerance.
- Block size and replicas are configurable per file.
- The Namenode receives a Heartbeat and a BlockReport from each DataNode in the cluster.
- BlockReport contains all the blocks on a Datanode.

# Replica Placement

12

- The placement of the replicas is critical to HDFS reliability and performance.
- Optimizing replica placement distinguishes HDFS from other distributed file systems.
- Rack-aware replica placement:
  - Goal: improve reliability, availability and network bandwidth utilization
  - Research topic
- Many racks, communication between racks are through switches.
- Network bandwidth between machines on the same rack is greater than those in different racks.
- Namenode determines the rack id for each DataNode.
- Replicas are typically placed on unique racks
  - Simple but non-optimal
  - Writes are expensive
  - Replication factor is 3
  - Another research topic?
- Replicas are placed: one on a node in a local rack, one on a different node in the local rack and one on a node in a different rack.
- 1/3 of the replica on a node, 2/3 on a rack and 1/3 distributed evenly across remaining racks.

# Replica Selection

13

- Replica selection for READ operation: HDFS tries to minimize the bandwidth consumption and latency.
- If there is a replica on the Reader node then that is preferred.
- HDFS cluster may span multiple data centers: replica in the local data center is preferred over the remote one.

# Safemode Startup

14

- On startup Namenode enters Safemode.
- Replication of data blocks do not occur in Safemode.
- Each DataNode checks in with Heartbeat and BlockReport.
- Namenode verifies that each block has acceptable number of replicas
- After a configurable percentage of safely replicated blocks check in with the Namenode, Namenode exits Safemode.
- It then makes the list of blocks that need to be replicated.
- Namenode then proceeds to replicate these blocks to other Datanodes.

# Filesystem Metadata

15

- The HDFS namespace is stored by Namenode.
- Namenode uses a transaction log called the EditLog to record every change that occurs to the filesystem meta data.
  - For example, creating a new file.
  - Change replication factor of a file
  - EditLog is stored in the Namenode's local filesystem
- Entire filesystem namespace including mapping of blocks to files and file system properties is stored in a file FsImage. Stored in Namenode's local filesystem.

# Namenode

16

- Keeps image of entire file system namespace and file Blockmap in memory.
- 4GB of local RAM is sufficient to support the above data structures that represent the huge number of files and directories.
- When the Namenode starts up it gets the FsImage and Editlog from its local file system, update FsImage with EditLog information and then stores a copy of the FsImage on the filesystem as a checkpoint.
- Periodic checkpointing is done. So that the system can recover back to the last checkpointed state in case of a crash.

# Datanode

17

- A Datanode stores data in files in its local file system.
- Datanode has no knowledge about HDFS filesystem
- It stores each block of HDFS data in a separate file.
- Datanode does not create all files in the same directory.
- It uses heuristics to determine optimal number of files per directory and creates directories appropriately:
  - Research issue?
- When the filesystem starts up it generates a list of all HDFS blocks and send this report to Namenode:  
Blockreport.

# Protocol

18

# The Communication Protocol

19

- All HDFS communication protocols are layered on top of the TCP/IP protocol
- A client establishes a connection to a configurable TCP port on the Namenode machine. It talks ClientProtocol with the Namenode.
- The Datanodes talk to the Namenode using Datanode protocol.
- RPC abstraction wraps both ClientProtocol and Datanode protocol.
- Namenode is simply a server and never initiates a request; it only responds to RPC requests issued by DataNodes or clients.

# Robustness

20

# Objectives

21

- Primary objective of HDFS is to store data reliably in the presence of failures.
- Three common failures are: Namenode failure, Datanode failure and network partition.

# DataNode failure and heartbeat

22

- A network partition can cause a subset of Datanodes to lose connectivity with the Namenode.
- Namenode detects this condition by the absence of a Heartbeat message.
- Namenode marks Datanodes without Hearbeat and does not send any IO requests to them.
- Any data registered to the failed Datanode is not available to the HDFS.
- Also the death of a Datanode may cause replication factor of some of the blocks to fall below their specified value.

# Re-replication

23

- The necessity for re-replication may arise due to:
  - A Datanode may become unavailable,
  - A replica may become corrupted,
  - A hard disk on a Datanode may fail, or
  - The replication factor on the block may be increased.

# Cluster Rebalancing

24

- HDFS architecture is compatible with data rebalancing schemes.
- A scheme might move data from one Datanode to another if the free space on a Datanode falls below a certain threshold.
- In the event of a sudden high demand for a particular file, a scheme might dynamically create additional replicas and rebalance other data in the cluster.
- These types of data rebalancing are not yet implemented: **research issue**.

# Data Integrity

25

- Consider a situation: a block of data fetched from Datanode arrives corrupted.
- This corruption may occur because of faults in a storage device, network faults, or buggy software.
- A HDFS client creates the checksum of every block of its file and stores it in hidden files in the HDFS namespace.
- When a client retrieves the contents of file, it verifies that the corresponding checksums match.
- If does not match, the client can retrieve the block from a replica.

# Metadata Disk Failure

26

- FsImage and EditLog are central data structures of HDFS.
- A corruption of these files can cause a HDFS instance to be non-functional.
- For this reason, a Namenode can be configured to maintain multiple copies of the FsImage and EditLog.
- Multiple copies of the FsImage and EditLog files are updated synchronously.
- Meta-data is not data-intensive.
- The Namenode could be single point failure: automatic failover is NOT supported! Another research topic.

# Data Organization

27

# Data Blocks

28

- HDFS support write-once-read-many with reads at streaming speeds.
- A typical block size is 64MB (or even 128 MB).
- A file is chopped into 64MB chunks and stored.

# Staging

29

- A client request to create a file does not reach Namenode immediately.
- HDFS client caches the data into a temporary file. When the data reached a HDFS block size the client contacts the Namenode.
- Namenode inserts the filename into its hierarchy and allocates a data block for it.
- The Namenode responds to the client with the identity of the Datanode and the destination of the replicas (Datanodes) for the block.
- Then the client flushes it from its local memory.

# Staging (contd.)

30

- The client sends a message that the file is closed.
- Namenode proceeds to commit the file for creation operation into the persistent store.
- If the Namenode dies before file is closed, the file is lost.
- This client side caching is required to avoid network congestion; also it has precedence in AFS (Andrew file system).

# Replication Pipelining

31

- When the client receives response from Namenode, it flushes its block in small pieces (4K) to the first replica, that in turn copies it to the next replica and so on.
- Thus data is pipelined from Datanode to the next.

# API (Accessibility)

32

# Application Programming Interface

33

- HDFS provides [Java API](#) for application to use.
- [Python](#) access is also used in many applications.
- A C language wrapper for Java API is also available.
- A HTTP browser can be used to browse the files of a HDFS instance.

# FS Shell, Admin and Browser Interface

34

- HDFS organizes its data in files and directories.
- It provides a command line interface called the FS shell that lets the user interact with data in the HDFS.
- The syntax of the commands is similar to bash and csh.
- Example: to create a directory /foodir  
`/bin/hadoop dfs –mkdir /foodir`
- There is also DFSAdmin interface available
- Browser interface is also available to view the namespace.

# Space Reclamation

35

- When a file is deleted by a client, HDFS renames file to a file in be the /trash directory for a configurable amount of time.
- A client can request for an undelete in this allowed time.
- After the specified time the file is deleted and the space is reclaimed.
- When the replication factor is reduced, the Namenode selects excess replicas that can be deleted.
- Next heartbeat(?) transfers this information to the Datanode that clears the blocks for use.

# Summary

36

- We discussed the features of the Hadoop File System, a peta-scale file system to handle big-data sets.
- What discussed: Architecture, Protocol, API, etc.
- Missing element: Implementation
  - The Hadoop file system (internals)
  - An implementation of an instance of the HDFS (for use by applications such as web crawlers).

# Identity-as-a-Service

## Overview

**E**

mployees in a company require to login into system to perform various tasks. These systems may be based on local server or cloud based. Following are the problems that an employee might face:

- Remembering different username and password combinations for accessing multiple servers.
- If an employee leaves the company, it's required to ensure that each of the user's account has been disabled. This increases workload on IT staff.

To solve above problems, a new technique emerged which is known as **Identity as a Service (IDaaS)**.

IDaaS offers management of identity (information) as a digital entity. This identity can be used during electronic transactions.

## Identity

**Identity** refers to set of attributes associated with something and make it recognizable. All objects may have same attributes, but their identity cannot be the same. This unique identity is assigned through unique identification attribute.

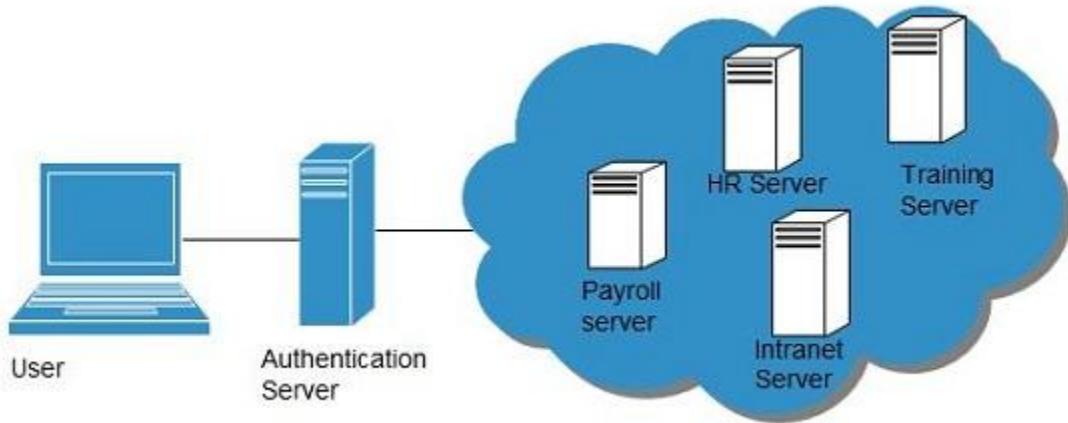
There are several **identity services** that have been deployed to validate services such as validating web sites, transactions, transaction participants, client, etc. Identity as a Service may include the following:

- Directory Services
- Federated Services
- Registration
- Authentication Services
- Risk and Event monitoring
- Single sign-on services
- Identity and Profile management

# Single Sign-On (SSO)

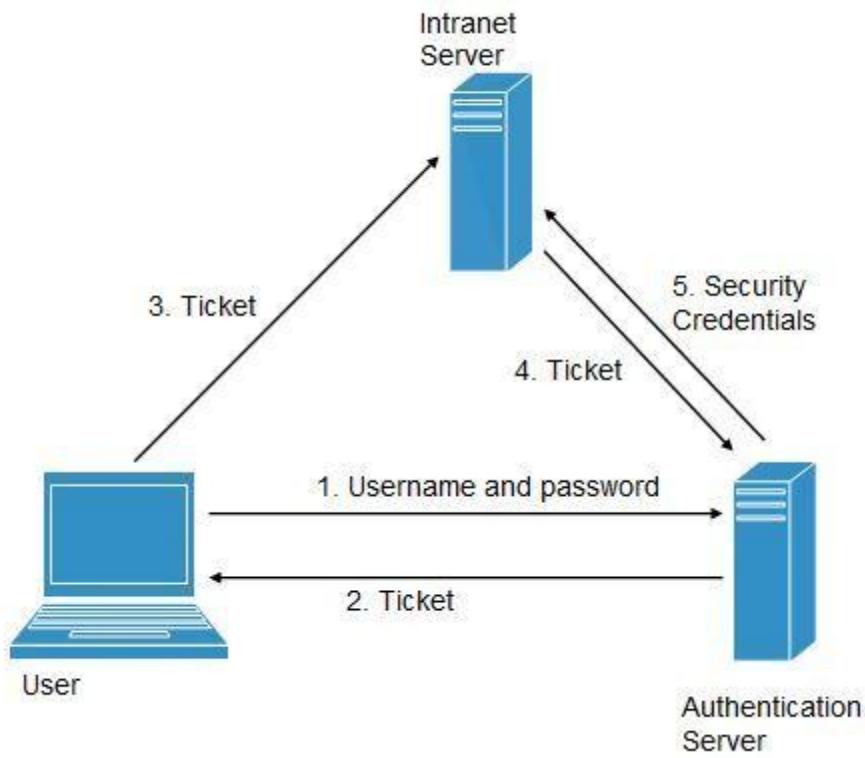
To solve the problem of using different username & password combination for different servers, companies now employ Single Sign-On software, which allows the user to login only one time and manages the user's access to other systems.

**SSO** has single authentication server, managing multiple accesses to other systems, as shown in the following diagram:



## SSO WORKING

There are several implementations of SSO. Here, we will discuss the common working of SSO:



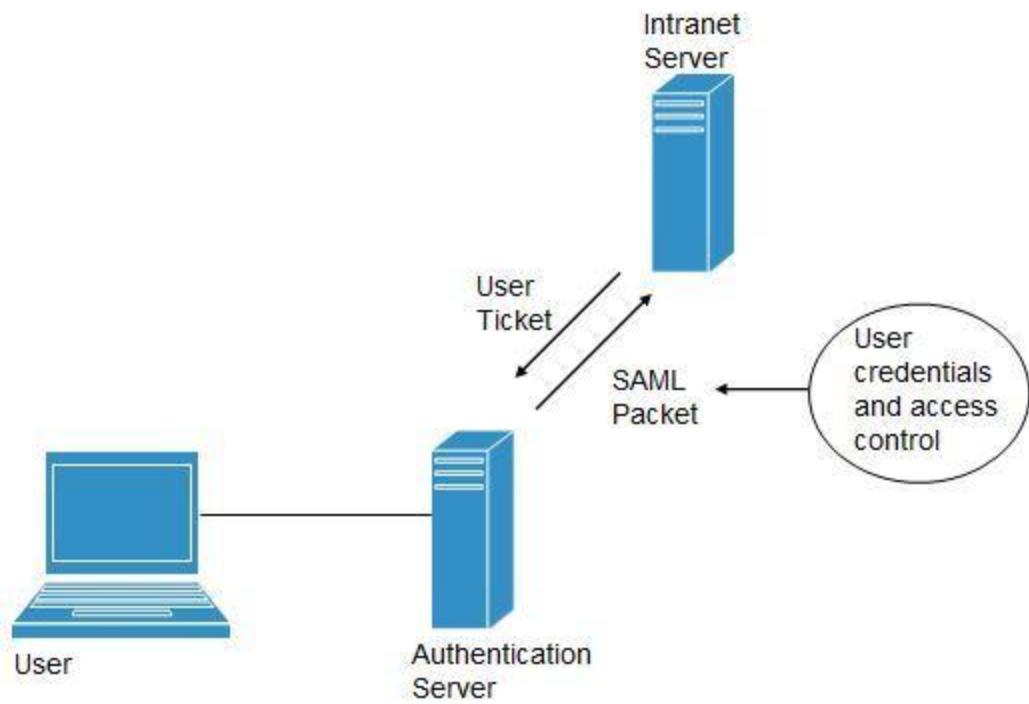
Following steps explain the working of Single Sign-On software:

1. User logs into the authentication server using a username and password.
2. The authentication server returns the user's ticket.
3. User sends the ticket to intranet server.
4. Intranet server sends the ticket to the authentication server.
5. Authentication server sends the user's security credentials for that server back to the intranet server.

If an employee leaves the company, then it just required to disable the user at the authentication server, which in turn disables the user's access to all the systems.

## Federated Identity Management (FIDM)

**FIDM** describes the technologies and protocols that enable a user to package security credentials across security domains. It uses **Security Markup Language (SAML)** to package a user's security credentials as shown in the following diagram:



## OpenID

It offers users to login into multiple websites with single account. Google, Yahoo!, Flickr, MySpace, WordPress.com are some of the companies that support OpenID.

## Benefits

- Increased site conversation rates.
- Access to greater user profile content.
- Fewer problems with lost passwords.
- Ease of content integration into social networking sites.

# Cloud Computing Management

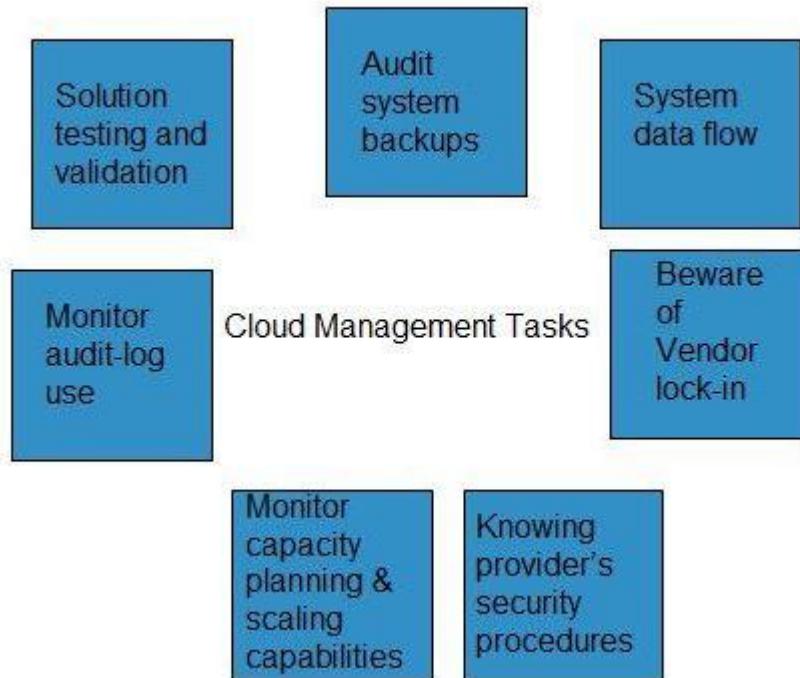
## Overview

I

t is the responsibility of cloud provider to manage resources and their performance. Management may include several aspects of cloud computing such as **load balancing, performance, storage and backups, capacity, deployment**, etc. Management is required to access full functionality of resources in the cloud.

## Cloud Management Tasks

Cloud Management involves a number of tasks to be performed by the cloud provider to ensure efficient use of cloud resources. Here, we will discuss some of these tasks:



## AUDIT SYSTEM BACKUPS

It is required to timely audit the backups to ensure you can successfully restore randomly selected files of different users. Backups can be performed in following ways:

- Backing up files by the company, from on-site computers to the disks that reside within the cloud.
- Backing up files by the cloud provider.

It is necessary to know if cloud provider has encrypted the data, who has access to that data and if the backup is taken at different locations, you must know where.

## SYSTEM'S DATA FLOW

The managers should develop a diagram describing a detailed process flow. This process flow will describe the movement of company's data throughout the cloud solution.

## BEWARE OF VENDOR LOCK-IN

The managers must know the procedure to exit from services of a particular cloud provider. There must exist procedures, enabling the managers to export company's data to a file and importing it to another provider.

## KNOWING PROVIDER'S SECURITY PROCEDURES

The managers should know the security plans of the provider for different services:

- Multitenant use
- E-commerce processing
- Employee screening
- Encryption policy

## MONITOR CAPACITY PLANNING AND SCALING CAPABILITIES

The managers should know the capacity planning in order to ensure whether the cloud provider will meet the future capacity requirement for his business or not.

It is also required to manage scaling capabilities in order to ensure services can be scaled up or down as per the user need.

## MONITOR AUDIT-LOG USE

In order to identify the errors in the system, managers must audit the logs on a regular basis.

## SOLUTION TESTING AND VALIDATION

It is necessary to test the solutions provided by the provider in order to validate that it gives the correct result and is error-free. This is necessary for a system to be robust and reliable.

# Cloud Computing Security

**S**

ecurity in cloud computing is a major concern. Data in cloud should be stored in encrypted form. To restrict client from direct accessing the shared data, proxy and brokerage services should be employed.

## Security Planning

Before deploying a particular resource to cloud, one should need to analyze several attributes about the resource such as:

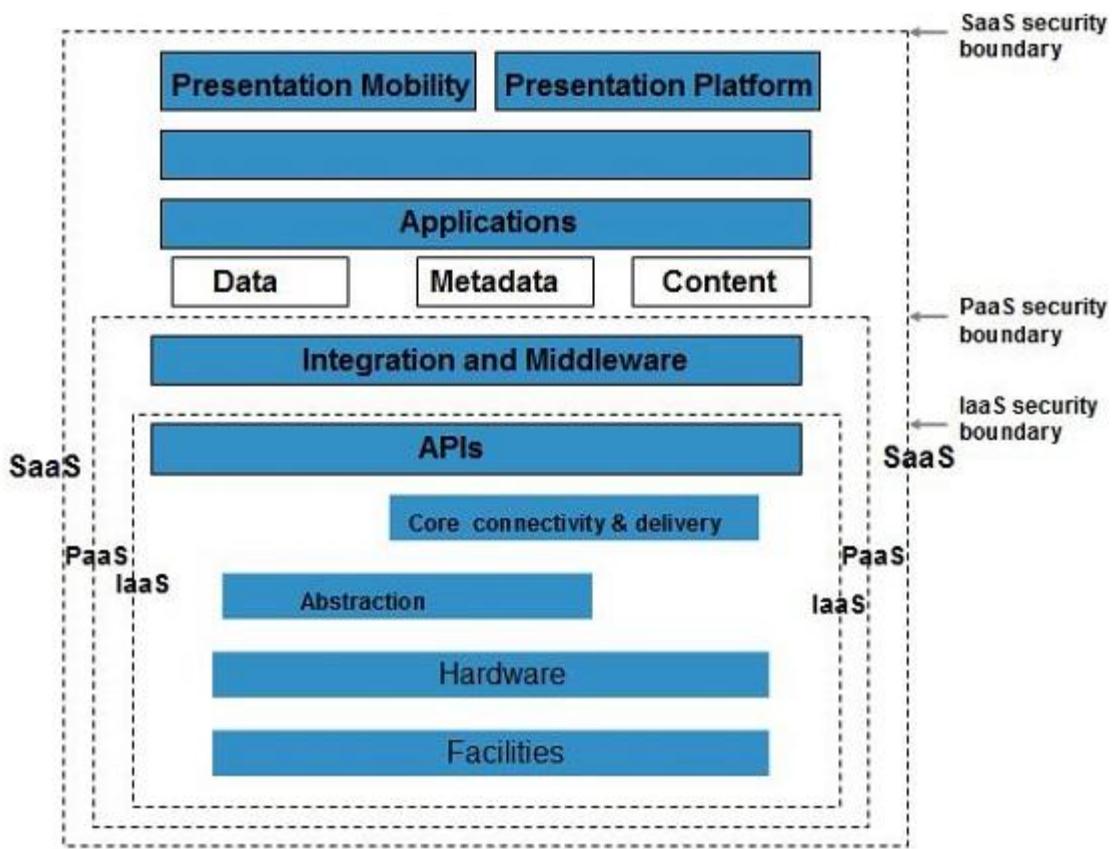
- Select which resources he is going to move to cloud and analyze its sensitivity to risk.
- Consider cloud service models such as **IaaS**, **PaaS**, and **SaaS**. These models require consumer to be responsible for security at different levels of service.
- Consider which cloud type such as **public**, **private**, **community** or **hybrid**.
- Understand the cloud service provider's system that how data is transferred, where it is stored and how to move data into and out of cloud.

Mainly the risk in cloud deployment depends upon the service models and cloud types.

## Understanding Security of Cloud

### SECURITY BOUNDARIES

A particular service model defines the boundary between the responsibilities of service provider and consumer. **Cloud Security Alliance (CSA)** stack model defines the boundaries between each service model and shows how different functional units relate to each other. The following diagram shows the **CSA stack model**:



#### KEY POINTS TO CSA MODEL:

- IaaS is the most basic level of service with PaaS and SaaS next two above levels of service.
- Moving upwards each of the service inherits capabilities and security concerns of the model beneath.
- IaaS provides the infrastructure, PaaS provides platform development environment and SaaS provides operating environment.
- IaaS has the least level of integrated functionalities and integrated security while SaaS has the most.
- This model describes the security boundaries at which cloud service provider's responsibility ends and the consumer's responsibilities begin.
- Any security mechanism below the security boundary must be built into the system and above should be maintained by the consumer.

Although each service model has security mechanism but security needs also depends upon where these services are located, in private, public, hybrid or community cloud.

## UNDERSTANDING DATA SECURITY

Since all the data is transferred using Internet, data security is of major concern in cloud. Here are key mechanisms for protecting data mechanisms listed below:

- Access Control

- Auditing
- Authentication
- Authorization

All of the service models should incorporate security mechanism operating in all above-mentioned areas.

## ISOLATED ACCESS TO DATA

Since data stored in cloud can be accessed from anywhere, therefore to protect the data, we must have a mechanism to isolate data from direct client access.

**Brokered Cloud Storage Access** is one of the approaches for isolating storage in cloud. In this approach, two services are created:

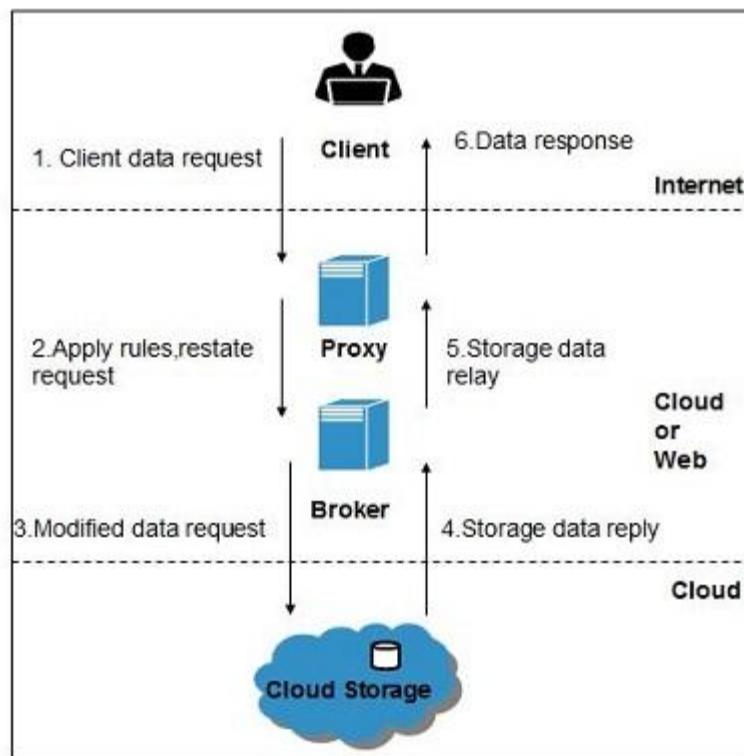
- A broker with full access to storage but no access to client.
- A proxy with no access to storage but access to both client and broker.

## WORKING OF BROKERED CLOUD STORAGE ACCESS SYSTEM

When the client issue request to access data:

- The client data request goes to proxy's external service interface.
- The proxy forwards the request to the broker.
- The broker requests the data from cloud storage system.
- The cloud storage system returns the data to the broker.
- The broker returns the data to proxy.
- Finally the proxy sends the data to the client.

All of the above steps are shown in the following diagram:



## Encryption

Encryption helps to protect data from being compromised. It protects data that is being transferred as well as data stored in the cloud. Although encryption helps to protect data from any unauthorized access, it does not prevent from data loss.

# Cloud Computing Operations

## Overview

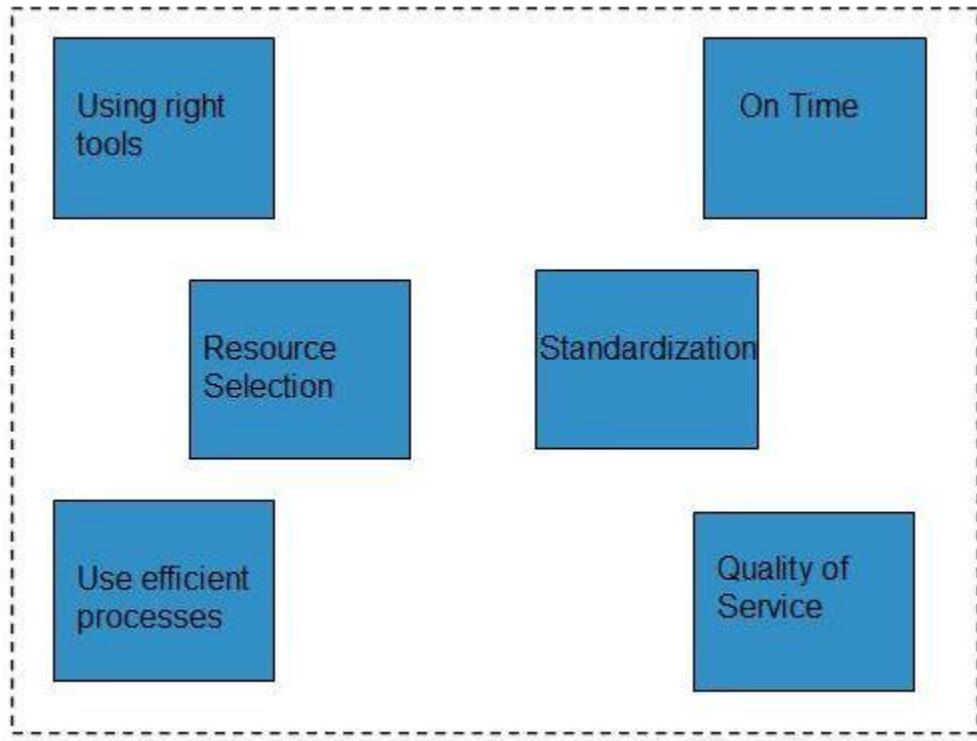
Cloud Computing operation refers to delivering superior cloud service. Today, cloud computing operations have become very popular and widely employed by many of the organizations just because it allows to perform all business operations over the Internet.

These operations can be performed using a web application or mobile based applications. There are a number of operations that are performed in cloud, some of them are shown in the following diagram:



## Managing Cloud Operations

There are several ways to manage day-to-day cloud operations, as shown in the following diagram:



- Always employ right tools and resources to perform any function in the cloud.
- Things should be done at right time and at right cost.
- Selecting an appropriate resource is mandatory for operation management.
- The process should be standardized and automated to avoid repetitive tasks.
- Using efficient process will eliminate the waste and redundancy.
- One should maintain the quality of service to avoid re-work later.

# Cloud Computing Applications

**C**loud Computing has its applications in almost all the fields such as **business, entertainment, data storage, social networking, management, entertainment, education, art** and **global positioning system**, etc. Some of the widely famous cloud computing applications are discussed here in this tutorial:

## Business Applications

Cloud computing has made businesses more collaborative and easy by incorporating various apps such as **MailChimp, Chatter, Google Apps for business, and Quickbooks**.

SN	Application Description
1	<b>MailChimp</b> It offers an <b>e-mail publishing platform</b> . It is widely employed by the businesses to design and send their e-mail campaigns.
2	<b>Chatter</b> <b>Chatter app</b> helps the employee to share important information about organization in real time. One can get the instant feed regarding any issue.
3	<b>Google Apps for Business</b> <b>Google</b> offers <b>creating text documents, spreadsheets, presentations</b> , etc., on <b>Google Docs</b> which allows the business users to share them in collaborating manner.
4	<b>Quickbooks</b> It offers <b>online accounting solutions</b> for a business. It helps in <b>monitoring cash flow, creating VAT returns and creating business reports</b> .

## Data Storage and Backup

**Box.com, Mozy, Joukuu** are the applications offering data storage and backup services in cloud.

SN	Application Description
1	<b>Box.com</b> <b>Box.com</b> offers drag and drop service for files. It just required to drop the files into Box and access from anywhere.
2	<b>Mozy</b> <b>Mozy</b> offers online backup service for files during a data loss.
3	<b>Joukuu</b>

	<b>Joukuu</b> is a web-based interface. It allows to display a single list of contents for files stored in <b>Google Docs, Box.net and Dropbox</b> .
--	--

## Management Applications

There are apps available for management task such as **time tracking, organizing notes**. Applications performing such tasks are discussed below:

SN	Application Description
1	<b>Toggl</b> It helps in tracking time period assigned to a particular project.
2	<b>Evernote</b> Evernote is an application that organizes the sticky notes and even can read the text from images which helps the user to locate the notes easily.
3	<b>Outright</b> It is an accounting app. It helps to track income, expenses, profits and losses in real time.

## Social Applications

There are several social networking services providing websites such as Facebook, Twitter, etc.

SN	Application Description
1	<b>Facebook</b> <b>Facebook</b> offers social networking service. One can share photos, videos, files, status and much more.
2	<b>Twitter</b> <b>Twitter</b> helps to interact directly with the public. One can follow any celebrity, organization and any person, who is on twitter and can have latest updates regarding the same.

## Entertainment Applications

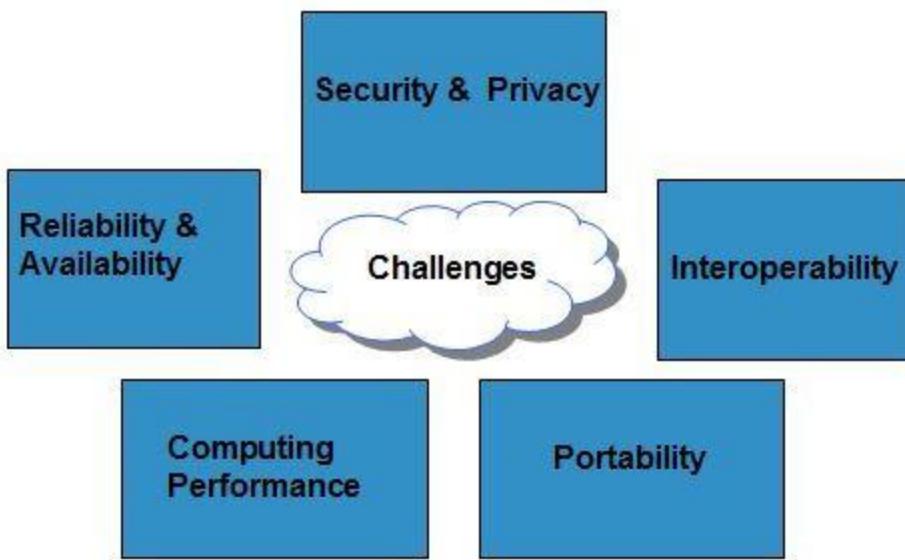
SN	Application Description
1	<b>Audiobox.fm</b> It offers streaming service, i.e., music can be stored online and can be played from cloud using service's own media player.

## Art Applications

SN	Application Description
1	<b>Moo</b> It offers art services such as designing and printing <b>business cards, postcards</b> and <b>minicards</b> .

# Cloud Computing Challenges

Cloud Computing, an emergence technology, has placed many challenges in different aspects. Some of these are shown in the following diagram:



## SECURITY & PRIVACY

Security and Privacy of information is the biggest challenge to cloud computing. Security and privacy issues can be overcome by employing encryption, security hardware and security applications.

## PORATABILITY

This is another challenge to cloud computing that applications should easily be migrated from one cloud provider to another. There should not be vendor lock-in. However, it is not yet made possible because each of the cloud provider uses different standard languages for their platforms.

## **INTEROPERABILITY**

Application on one platform should be able to incorporate services from other platform. It is made possible via web services. But writing such web services is very complex.

## **COMPUTING PERFORMANCE**

To deliver data intensive applications on cloud requires high network bandwidth, which results in high cost. If done at low bandwidth, then it does not meet the required computing performance of cloud application.

## **RELIABILITY AND AVAILABILITY**

It is necessary for cloud systems to be reliable and robust because most of the businesses are now becoming dependent on services provided by third-party.

CHAPTER

# 23

# Mobile Cloud Computing

**C**loud Computing offers such smartphones that have rich Internet media experience and require less processing, less power. In term of Mobile Cloud Computing, processing is done in cloud, data is stored in cloud. And the mobile devices serve as a media for display.

Today smartphones are employed with rich cloud services by integrating applications that consume web services. These web services are deployed in cloud.

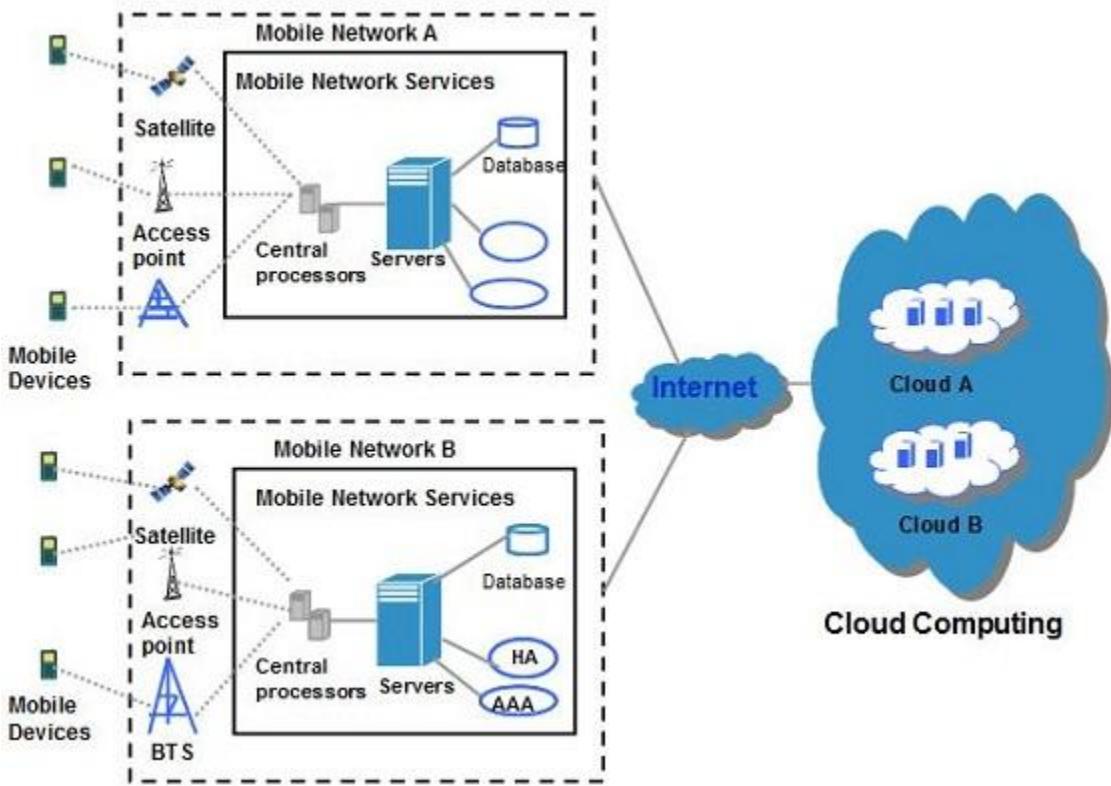
There are several Smartphone operating systems available such as **Google's Android**, **Apple's iOS**, **RIM BlackBerry**, **Symbian**, and **Windows Mobile Phone**. Each of these platforms support third-party applications that are deployed in cloud.

## Architecture

**MCC** includes four types of cloud resources:

- Distant mobile cloud
- Distant immobile cloud
- Proximate mobile computing entities
- Proximate immobile computing entities
- Hybrid

The following diagram shows the framework for mobile cloud computing architecture:



## Issues

Despite of having significant development in field of mobile computing, there still exists many issues:

### EMERGENCY EFFICIENT TRANSMISSION

There should be a frequent transmission of information between cloud and the mobile devices.

### ARCHITECTURAL ISSUES

Mobile cloud computing is required to make architectural neutral because of heterogeneous environment.

### LIVE VM MIGRATION

It is challenging to migrate an application, which is resource-intensive to cloud and to execute it via Virtual Machine.

### MOBILE COMMUNICATION CONGESTION

Due to continuous increase demand for mobile cloud services, the workload to enable smooth communication between cloud and mobile devices has been increased.

# Inter Cloud /Cloud Federation

- While the Cloud approach provides many benefits it still has limitations.
  - Limited amount of resources.
  - Limited types of services provided.
  - Limited geographical presence.
  - Good but not perfect fault tolerance.
- Solution – combine the clouds.
  - Federate individual clouds to allow resource sharing
  - Governed by pre-arranged peering / exchange relationship.
  - Requires standardization and communication between clouds.

# Inter Cloud /Cloud Federation

- Intercloud – “the cloud of clouds”. Name derived from the Internet (network of networks).
- Does not dictate the internal organization or structure used inside of a cloud (intracloud), but rather only the connection between clouds.
- Coordinating the delivery of ubiquitous and interoperable services for content, storage, computation, etc.
- Modeled after the Internet (network of networks) infrastructure.
- Intercloud relies on the generation, maintenance and usage of gathered information about the federated clouds.
- Create among federated clouds common: naming, addressing, Identity, trust, presence, messaging, multicast, time domain and application messaging.

# Inter Cloud /Cloud Federation

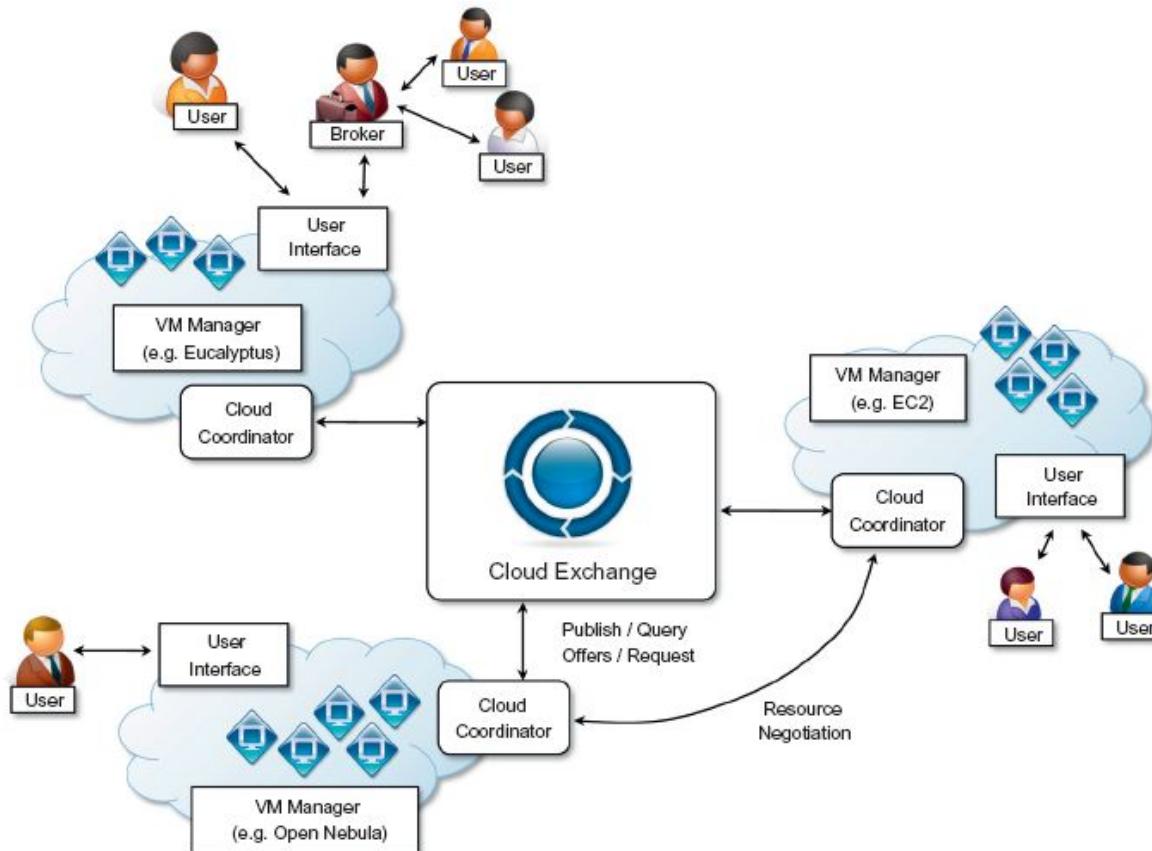
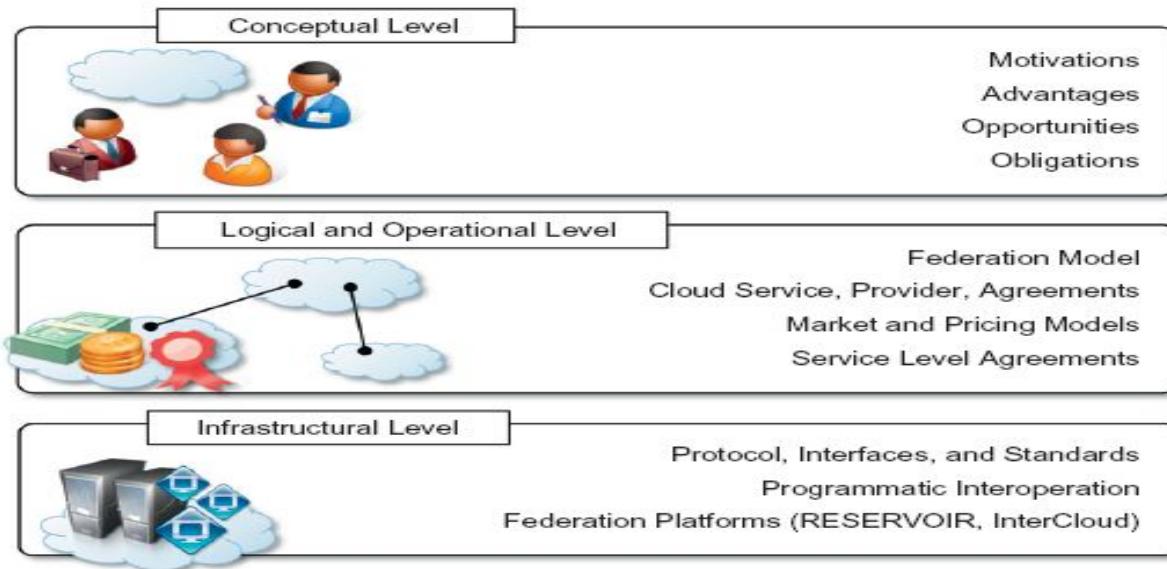


FIGURE 11.12

InterCloud architecture.

# Cloud Federation stack

*Cloud federation manages consistency and access controls when two or more independent geographically distinct Clouds share either authentication, files, computing resources, command and control or access to storage resources.*



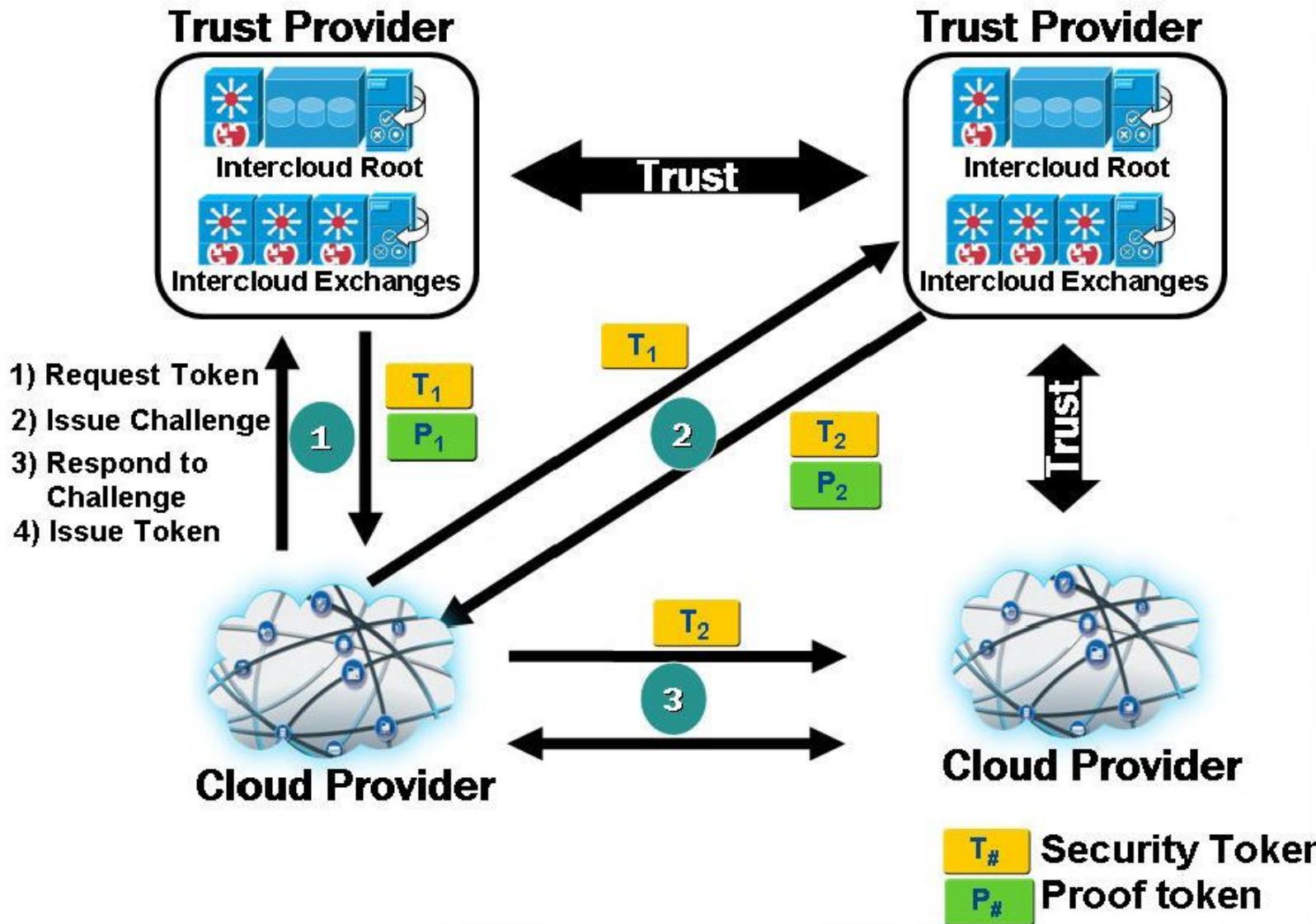
**FIGURE 11.7**

Cloud federation reference stack.

# Security Issues in Intercloud

- The goal of Intercloud is the ability to dynamically manage workload between cloud providers with maximum flexibility and choice given to users.
- The primary security concern is the ability of tasks to cross from one administrative domain to another and be serviced (at some cost) for the user.
- A trust model is required to allow tasks to seamlessly migrate from one cloud to another without user intervention.
- Additionally, sensitive information about the tasks (and user) should not be disclosed during the migration.

- Another key problem in federating clouds is the management of identify and access permissions across clouds.
  - Key functionality provided by IAM includes: user provisioning, user management, authorization and identify data integration/virtualization.
- Intercloud exchanges facilitating this functionality by being the trusted third parties between cloud providers to establish cryptographic session keys for communication.
- Currently, most cloud providers only have proprietary means of controlling granular (resource level) access.
- It is proposed that the XACML language (standardized by OASIS) is used to standardize the communication of access controls and policies between clouds.  
eXtensible Access Control Markup Language



**Table 11.1** Cloudbus Toolkit Components and Technologies

Technology	Description
Aneka	Middleware for cloud applications development and deployment.
Broker	Middleware for scheduling distributed applications across heterogeneous systems based on the bag-of-tasks model.
Workflow management system	Middleware for the execution, composition, management, and monitoring of workflows across heterogeneous systems.
Market Maker/Meta-Broker	A matchmaker that matches the user's requirements with service providers' capabilities within the context of a marketplace.
InterCloud	A framework for the federation of independent computing clouds.
MetaCDN	Middleware that leverages storage clouds for intelligently delivering users' content based on their QoS and budget preferences.
Energy-efficient computing	Ongoing research on developing techniques and technologies for addressing scalability and energy efficiency.

# Some Important Terminology

## Distributed management task force

The *Distributed Management Task Force (DMTF)* is an organization involving more than 4,000 active members, 44 countries, and nearly 200 organizations. It is the industry organization leading the development, adoption, and promotion of interoperable management standards and initiatives. With specific reference to cloud computing, the DMTF introduced the *Open Virtualization Format (OVF)* and supported several initiatives for interoperable cloud technologies, such as the *Open Cloud Standards Incubator*, the *Cloud Management Working Group (CMWG)*, and the *Cloud Audit Data Federation Working Group (CADFWG)*.

The *Open Virtualization Format (OVF)* [51] is a vendor-independent format for packaging standards designed to facilitate the portability and deployment of virtual appliances across different virtualization platforms. OVF can be used by independent software vendors (ISVs) to package and securely distribute applications, and system images can be imported and deployed on multiple platforms, thus enabling cross-platform portability. The specification is the result of the collaborative effort of Dell, HP, IBM, Microsoft, VMWare, and XenSource in defining a platform-independent

# Live Migration of Virtual Machines in Cloud

Ashima Agarwal, Shangruff Raina

Department of Computer, MIT College of Engineering Pune, India

**Abstract-** Migration of a virtual machine is simply moving the VM running on a physical machine (source node) to another physical machine (target node). It is done as, while the VM is running on the source node, and without disrupting any active network connections, even after the VM is moved to the target node. It is considered “live”, since the original VM is running, while the migration is in progress. Huge benefit of doing the live migration is the very small (almost zero) downtime in the order of milliseconds.

There exists a model in which the load is balanced among the servers according to their processor usage or their IO usage and keeping virtual machines zero downtime. To increase the throughput of the system, it is necessary that the virtual machines are load balanced statically, i.e. the load is distributed to each part of the system in proportion to their computing IO capacity.

To migrate a virtual machine from one physical host to another, the control of virtual machines is converted to the management of services in Red Hat Cluster Suite. This creates a high availability and load balancing cluster services, to migrate a virtual machine from one physical host to another.

Software services, file systems and network status can be monitored and controlled by the cluster suite, services and resources can be failed over to other network nodes in case of failure. The cluster suite forcibly terminates a cluster node's access to services or resources to ensure the node and data is in a known state. The node is terminated by removing power or access to the shared storage.

## I. INTRODUCTION

### What is virtual machine?

One computer containing multiple operating systems loaded on a single PC, each of which functions as a separate OS on a separate physical machine. Virtualization [1] software does just that by creating and managing one or more virtual machines on a single, physical host PC. It can run its own operating systems and applications as if it were a physical computer. A virtual machine [2] behaves exactly like a physical computer and contains its own virtual (i.e. software-based) CPU, RAM hard disk and network interface card (NIC).

### Benefits of virtual machines

1. **Isolation:** While virtual machines can share the physical resources of a single computer, they remain completely isolated from each other as if they were separate physical machines. If, for example, there are four virtual machines on a single physical server and one of the virtual machines crashes, the other three virtual machines remain available. Isolation is an important

reason why the availability and security of applications running in a virtual environment is far

2. **Encapsulation:** A virtual machine is essentially a software container that bundles or “encapsulates” a complete set of virtual hardware resources, as well as an operating system and all its applications, inside a software package. Encapsulation makes virtual machines incredibly portable and easy to manage.
3. **Hardware Independence:** Virtual machines are completely independent from their underlying physical hardware. For example, you can configure a virtual machine with virtual components (eg, CPU, network card, SCSI controller) that are completely different from the physical components that are present on the underlying hardware. Virtual machines on the same physical server can even run different kinds of operating systems (Windows, Linux, etc).

## II. WHAT IS LIVE MIGRATION?

Live migration [3] is the movement of a virtual machine from one physical host to another while continuously powered-up. When properly carried out, this process takes place without any noticeable effect from the point of view of the end user. Live migration allows an administrator to take a virtual machine offline for maintenance or upgrading without subjecting the system's users to downtime.

When a VM is running a live service it is important that this transfer occurs in a manner that balances the requirements of minimizing both downtime and total migration time. The former is the period during which the service is unavailable due to there being no currently executing instance of the VM; this period will be directly visible to clients of the VM as service interruption. The latter is the duration between when migration is initiated and when the original VM may be finally discarded and hence, the source host may potentially be taken down for maintenance, upgrade or repair. It is easiest to consider the trade-offs between these requirements by generalizing memory transfer into three phases:

**Push phase:** The 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 phase:** The source VM is stopped, pages are copied across to the destination VM, then the new VM is started. The time between stopping VM on source and resuming it on destination is called "down-time". Down-time of a VM during a live migration could be a few milliseconds to seconds according to the size of memory and applications running on the VM. There

are some techniques to reduce live migration down-time such as using probability density function of memory change.<sup>[3]</sup>

**Pull phase:** 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. VM migration is initiated by suspending the VM at the source. With the VM suspended, a minimal execution state of the VM (CPU, registers, and non-pageable memory) is transferred to the target. The VM is then resumed at the target, even though the entire memory state of the VM has not yet been transferred, and still resides at the source. At the target, when the VM tries to access pages that have not yet been transferred, it generates page-faults, which are trapped at the target and redirected towards the source over the network. Such faults are referred to as network faults. Source host responds to the network-fault by sending the faulted page. Since each page fault of the running VM is redirected towards the source, it can degrade the applications running inside the VM. However, when pure demand-paging accompanied with the techniques such as pre-paging can reduce this impact by a great extent. Live migration is widely used in virtualization ready data centers and enterprise IT departments. It separates software's from physical servers, and provides desirable abilities such as online server maintenance, dynamic load balancing, and etc.

Live migration can be applied in both local area network (LAN) environment and wide area network (WAN) environment. Live migration in LAN environment is simpler, because live migration process avoids virtual storage migration by sharing a network storage. Furthermore, live migration in LAN skips the migration of network topology, and the migration process only needs to broadcast an unsolicited ARP reply from the migrated VM, in order to advertise the MAC relocation.

### III. SEAMLESS LIVE MIGRATION

When down-time of a VM during a live migration is zero or a few millisecond which is not noticeable by end user, it is called a seamless live migration. Otherwise, the end user will feel a small or relatively long glitch in the service.

#### What is the purpose of live migration?

Migrating an entire OS and all of its applications as one unit allows us to avoid many of the difficulties faced by process-level migration approaches. With virtual machine migration, on the other hand, the original host may be decommissioned once migration has completed. This is particularly valuable when migration is occurring in order to allow maintenance of the original host.

Secondly, migrating at the level of an entire virtual machine means that in-memory state can be transferred in a consistent and (as will be shown) efficient fashion. This applies to kernel-internal state (e.g. the TCP control block for a currently active connection) as well as application-level state, even when this is shared between multiple cooperating processes.

Thirdly, live migration of virtual machines allows a separation of concerns between the users and operator of a data center or cluster. Users have 'carte blanche' regarding the software and services they run within their virtual machine, and need not provide the operator with any OS-level access at all. Live OS migration is a extremely powerful tool for cluster administrators,

allowing separation of hardware and software considerations, and consolidating clustered hardware into a single coherent management domain. If a physical machine needs to be removed from service an administrator may migrate OS instances including the applications that they are running to alternative machine(s), freeing the original machine for maintenance.

#### How is live migration achieved?

By using a *pre-copy* approach in which pages of memory are iteratively copied from the source machine to the destination host, all without ever stopping the execution of the virtual machine being migrated. Page level protection hardware is used to ensure a consistent snapshot is transferred, and a rate-adaptive algorithm is used to control the impact of migration traffic on running services. The final phase pauses the virtual machine, copies any remaining pages to the destination, and resumes execution there. We eschew a 'pull' approach which faults in missing pages across the network since this adds a residual dependency of arbitrarily long duration, as well as providing in general rather poor performance.

The logical steps that we execute when migrating an OS are summarized in Figure 3. We take a conservative approach to the management of migration with regard to safety and failure handling. Although the consequences of hardware failures can be severe, our basic principle is that safe migration should at no time leave a virtual OS more exposed to system failure than when it is running on the original single host. To achieve this, we view the migration process as a transactional interaction between the two hosts involved:

**Stage 0 Pre-Migration:** We begin with an active VM on physical host A. To speed any future migration, a target host may be preselected where the resources required to receive migration will be guaranteed.

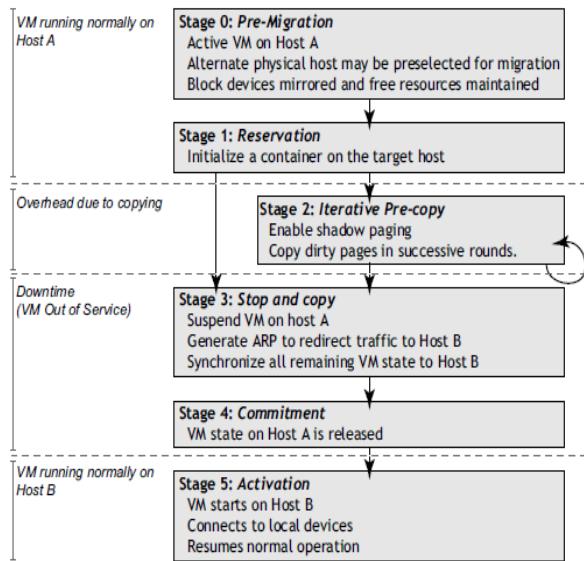
**Stage 1 Reservation:** A request is issued to migrate an OS from host A to host B. We initially confirm that the necessary resources are available on B and reserve a VM container of that size. Failure to secure resources here means that the VM simply continues to run on A unaffected.

**Stage 2 Iterative Pre-Copy:** During the first iteration, all pages are transferred from A to B. Subsequent iterations copy only those pages dirtied during the previous transfer phase.

**Stage 3 Stop-and-Copy:** We suspend the running OS instance at A and redirect its network traffic A to B. As described earlier, CPU state and any remaining inconsistent memory pages are then transferred. At the end of this stage there is a consistent suspended copy of the VM at both A and B. The copy at A is still considered to be primary and is resumed in case of failure.

**Stage 4 Commitment** Host B indicates to A that it has successfully received a consistent OS image. Host A acknowledges this message as commitment of the migration transaction: host A may now discard the original VM, and host B becomes the primary host.

**Stage 5 Activation:** The migrated VM on B is now activated. Post-migration code runs to reattach device drivers to the new machine and advertise moved IP addresses.



**Figure 3: Migration Timeline**

This approach to failure management ensures that at least one host has a consistent VM image at all times during migration. It depends on the assumption that the original host remains stable until the migration commits, and that the VM may be suspended and resumed on that host with no risk of failure. Based on these assumptions, a migration request essentially attempts to move the VM to a new host, and on any sort of failure execution is resumed locally, aborting the migration.

To solve the problem of load balancing amongst the physical hosts in cloud by adaptive live migration of virtual machine, there exists a model in which the load is balanced among the servers according to their processor usage or their IO usage and keeping virtual machines zero downtime. To increase the throughput of the system, it is necessary that the virtual machines are load balanced statically, i.e. the load is distributed to each part of the system in proportion to their computing IO capacity. The System Architecture of such a model is given in next chapter.

#### IV. SYSTEM ARCHITECTURE

By the use of EUCALYPTUS [4] one or more VLAN can be created, each VLAN may be across many physical hosts and may include many virtual machines, so when a virtual machine is migrated from one physical host to another, an inclusion relationship can be kept between VLAN and virtual machine. Following diagram represents the system architecture of an adaptive intra-cloud load balancing model.

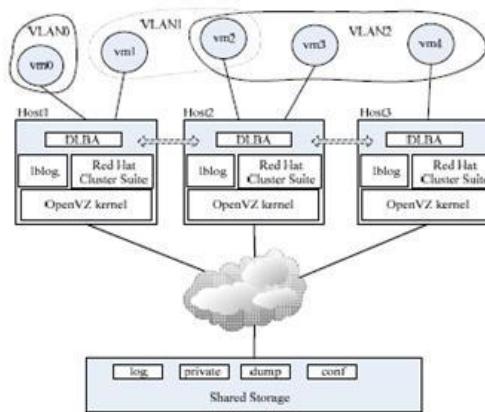
**Lblog:** The program lblog is a small cronjob which runs regularly on each host to monitor predefined resources, such as CPU and IO load, and log their usage to the log directory on the shared storage.

**DLBA (Distributed load balancing algorithm):** It runs on each physical host separately and dynamically migrates virtual machines of local host to other physical hosts according to resource uses.

**Dump:** When virtual machines are migrated, their memory contents are dumped to the dump directory of shared storage.

**Conf:** This directory contains the configuration files.

**Private:** This directory contains the file system.



**Figure 4: System Architecture**

To migrate a virtual machine from one physical host to another, the control of virtual machines is converted to the management of services in **Red Hat Cluster Suite**[5]. This creates a high availability and load balancing cluster services, to migrate a virtual machine from one physical host to another. The cluster suite forcibly terminates a cluster node's access to services or resources to ensure the node and data is in a known state. The node is terminated by removing power or access to the shared storage. The detailed explanation of the working of Red Hat Cluster Suite is given in next chapter.

#### V. RED HAT CLUSTER SUITE

##### Introduction

##### High-availability Service Management

High-availability[7] service management provides the ability to create and manage high-availability *cluster services* in a Red Hat cluster. The key component for high-availability service management in a Red Hat cluster, rgmanager, implements cold failover for off-the-shelf applications. In a Red Hat cluster, an application is configured with other cluster resources to form a high-availability cluster service. A high-availability cluster service can fail over from one cluster node to another with no apparent interruption to cluster clients. Cluster-service failover can occur if a cluster node fails or if a cluster system administrator moves the service from one cluster node to another (for example, for a planned outage of a cluster node). A cluster

service can run on only one cluster node at a time to maintain data integrity. Failover priority can be specified in a failover domain. Specifying failover priority consists of assigning a priority level to each node in failover domain. The priority level determines the failover order — determining which node that a cluster service should fail over to. If you do not specify failover priority, a cluster service can fail over to any node in its failover domain. Also, if a cluster service is restricted to run only on nodes of its associated failover domain. (When associated with an unrestricted failover domain, a cluster service can start on any cluster node in the event no member of the failover domain is available.)

## VDE (Virtual Distributed Ethernet)

### Introduction

The acronym VDE [6] is self explaining: it is a *Virtual* network because it is built completely in software, it is *Distributed* as parts of the same network can run on different physical (real) computers and it is an *Ethernet* as the entire virtual software structure is able to forward, dispatch and route plain Ethernet packets. The main features of VDE are the following:

- VDE is Ethernet compliant.
- VDE is general; it is a virtual infrastructure that gives connectivity to several kinds of software components: emulators/virtual machines, real operating systems and other connectivity tools.
- VDE is distributed.
- VDE does not need specifically administration privileges to run.

With VDE it is also possible to integrate real computers in the emulated network. When a real computer is connected to a VDE a virtual interface (based on tuntap) is visible from the operating system. This virtual interface appears exactly as it were a hardware interface and behaves as a physical ethernet interface. This operation however changes the network behavior of the host computer and thus need administrative privileges to be completed.

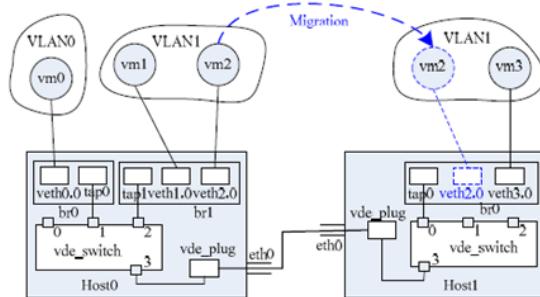
Currently VDE supports User-Mode Linux virtual machines, qemu, Bochs and MPS.

### Network Connection for Zero-Downtime Migrated Virtual Machine

EUCALYPTUS provides support for VLAN across multi physical hosts via VDE virtual switches. There is a VDE switch on each physical host. Ports of VDE switch are classified into two categories: tagged port and untagged port. Tagged ports transmit packages belonging to multi VLAN, for example the port 3 on Host0 and untagged ports transmit packages belonging to a VLAN, for example the port 1 and 2 on Host0. On single physical host there may be multi VLAN, and each VLAN binds to an untagged port of VDE switch on the host dynamically, or example VLAN0 and VLAN1 on Host0 respectively bind to untagged port 1 and 2. Bridge device connects to untagged port of VDE switch via virtual network card tap. VDE switches on different physical hosts connect to each other via vde plug which

is part of VDE switch software suite and implements package transmission, encryption and authentication functions.

In figure 10 below, when vm2 belonging to VLAN1 migrates from Host0 to Host1, we need to delete virtual network device veth2.0 from bridge br1 on Host0. If there were no other network devices in bridge br1 except tap1 and veth2.0, we also would needed to unbind VLAN1 to untagged port 2, thus port 2 could be used by other VLAN in the future. On the target Host1, because VLAN1 has existed, before Red Hat Cluster Suite starts vm2, we only need create a pair of virtual network devices on Host1 for vm2: eth0 and veth2.0 (eth0 in virtual machine vm2 and veth2.0 on the Host1), then add veth2.0 into the bridge br0. If there were not VLAN1 on Host1, we would need to find an untagged port in VDE switch and bind it to VLAN1, also create the corresponding bridge br0.



**Figure 8: Virtual Machine Migration**

### Advantages of Live Migration

1. Reduce IT costs and improve flexibility with server consolidation.
2. Decrease downtime and improve reliability with business continuity and disaster recovery.
3. Increase energy efficiency by running fewer servers and dynamically powering down unused servers with our green IT solutions
4. Accessing more processing power (in the sense of load balancing),
5. Exploitation of resource locality (for performance), resource sharing (meaning sharing of expensive or rare).
6. Resources - such as telescopes or medical equipment – or large amounts o free memory by processes over a network), fault resilience.
7. Simplified system administration and mobile computing (for instance as used by commuters from office to home).

### VI. CONCLUSION

Thus, Live Migration is the movement of a virtual machine from one physical host to another while continuously powered-up. This helps in decreasing downtime and improves

reliability with business continuity and disaster recovery. To migrate a virtual machine from one physical host to another, the control of virtual machines is configured to the management of services in **Red Hat Cluster Suite**. VDE (virtual distributed Ethernet) virtual switches are used which are connected between the physical machines. Using VDE, users can create VLAN for each virtual machine instance. Hence, the motive of zero downtime of virtual machines is achieved by live migration.

#### REFERENCES

- [1] <http://www.vmware.com/virtualization/virtual-machine.html>
- [2] <http://searchservervirtualization.techtarget.com/definition/virtual-machine>
- [3] C Clark, K Fraser, S Hand, J Hansen, and E Jul., "Live migration of virtual machines", Proceedings of the 2nd ACM/USENIX Symposium on Networked Systems Design and Implementation (NSDI), pages: 273-286, 2005
- [4] Daniel Nurmi, Rich Wolski, Chris Grzegorczyk., "The EUCLYPTUS Open-source Cloud computing System", in Proceedings of Cloud Computing and Its Applications [online], Chicago, Illinois, October 2008
- [5] Yi Zhao, Wenlong Huang " Adaptive Distributed Load Balancing Algorithm based on Live Migration of Virtual Machines in Cloud"
- [6] Davoli,R., "VDE: virtual distributed Ethernet", Test beds and Research Infrastructures for the Development of Networks and Communities, pages 213-220, February 2005
- [7] [http://www.centos.org/docs/5/html/5.1/Cluster\\_Suite\\_Overview/s1-service-management-overview-CSO.html](http://www.centos.org/docs/5/html/5.1/Cluster_Suite_Overview/s1-service-management-overview-CSO.html)

#### AUTHORS

**First Author** – Ashima Agarwal, Third year, computer engineering, MIT College of Engineering, Pune, Email: ashimaagarwal30@gmail.com

**Second Author** – Shangruff Raina, Third year, computer engineering, MIT College of Engineering, Pune, Email: shangruff@gmail.com

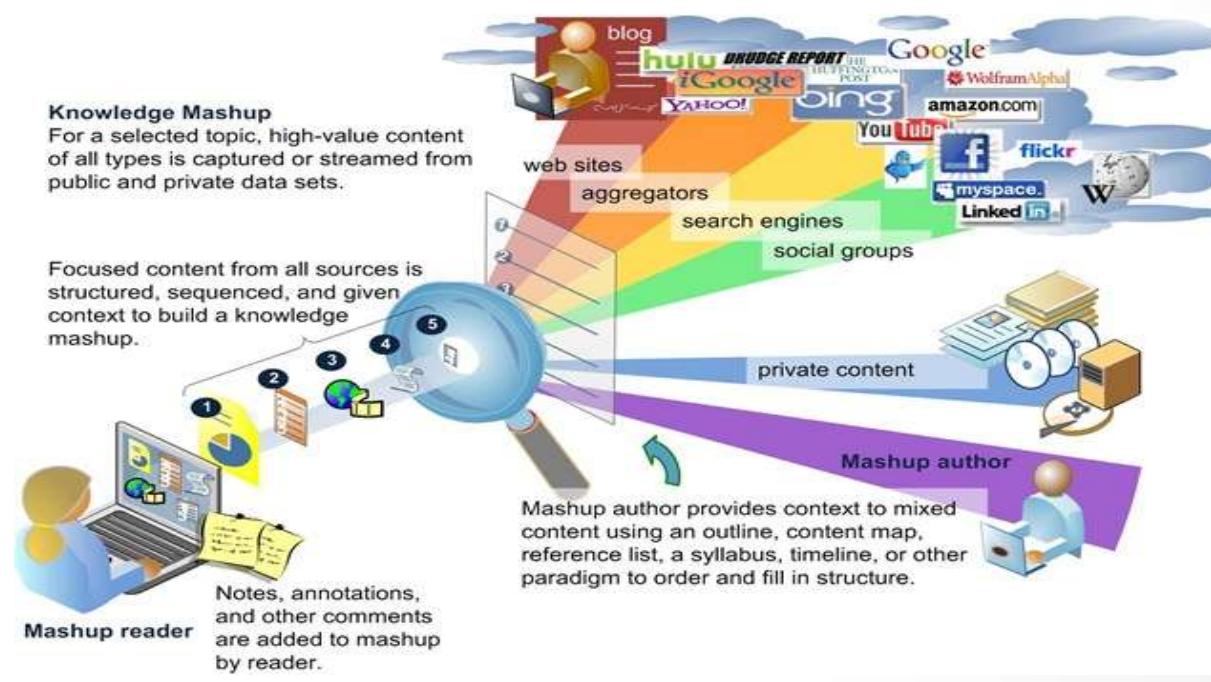
# *Cloud Technology*

## Mashup



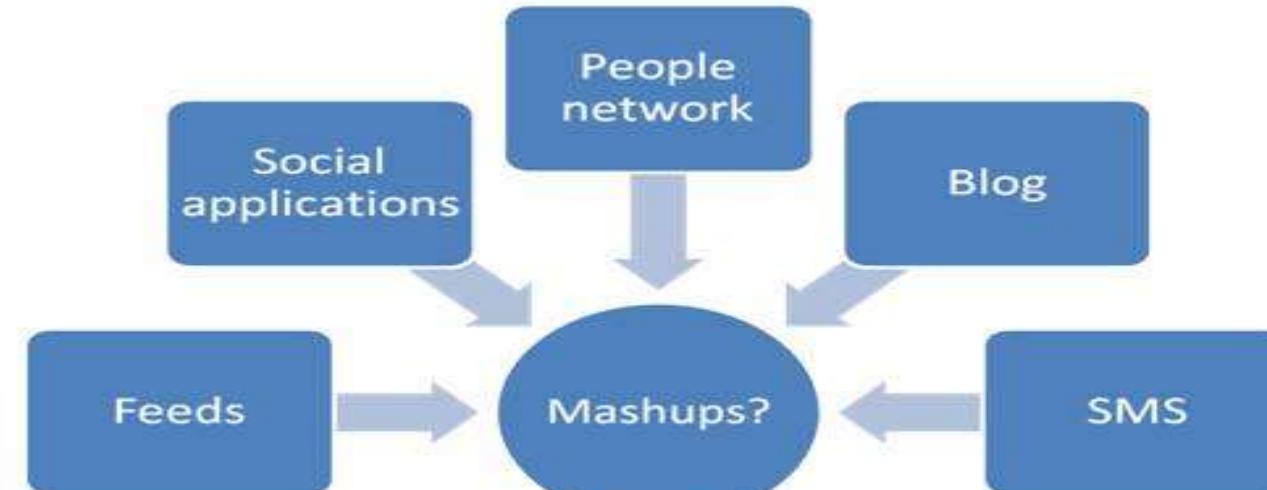
# Mashup

A mashup is a *Web page* or *application* that uses and combines data, presentation or functionality from two or more sources to create new services. The term implies easy, fast integration, frequently using open *API* (Application Programming Interface) and *data sources* to produce enriched results that were not necessarily the original reason for producing the raw source data.



# Good ideas behind Mashup

- Allow information to be viewed from **different perspectives** (e.g., view real estate data on a map)
- Combine data from multiple sources into a **single unified view** (e.g., compare gas prices in the neighborhood).
- Enrich raw data with new information (e.g., view eBay real estate auction along with Amazon)





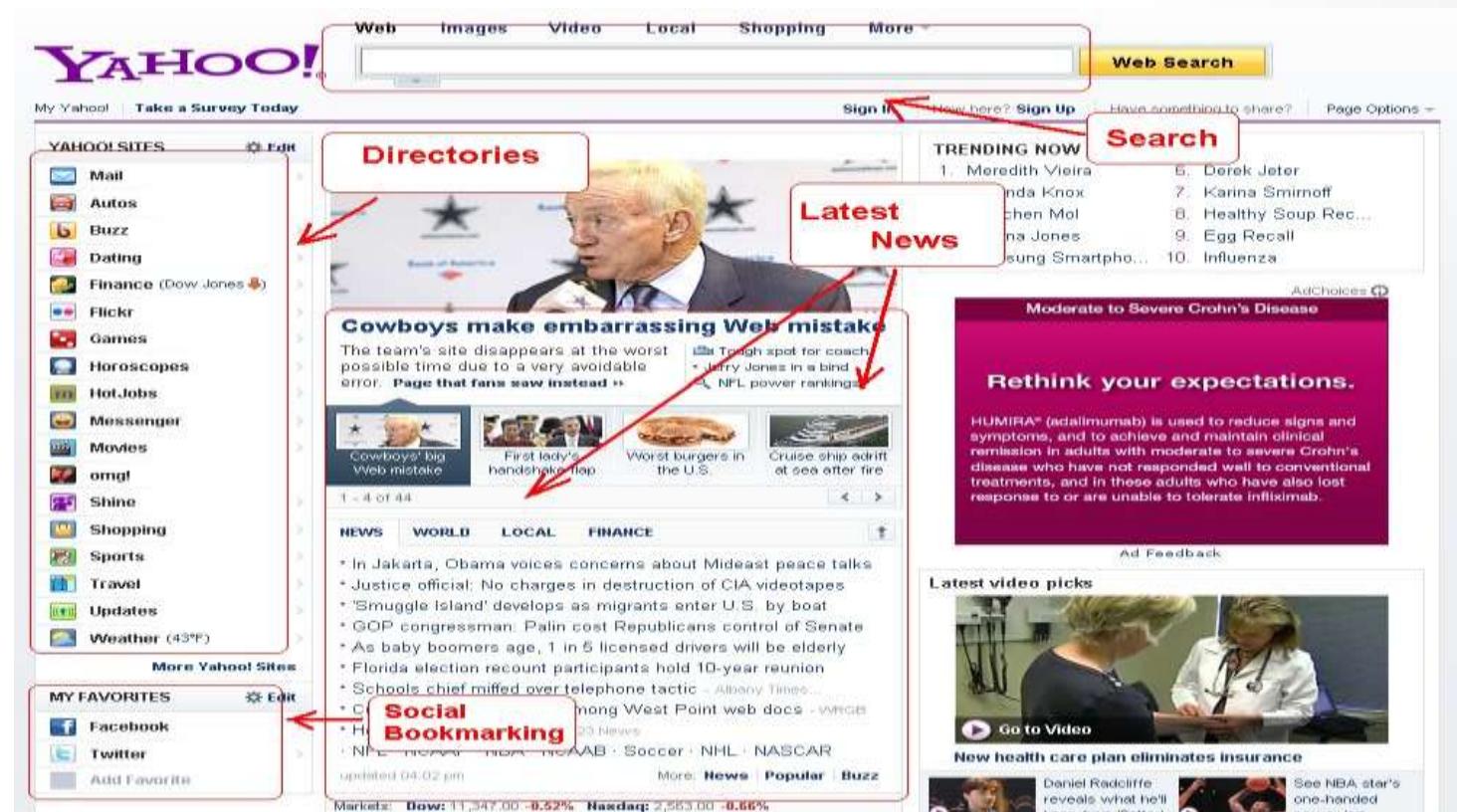
Mashups have recently exploded on the web, for two main reasons.

- First, many of the major internet companies, such as **Yahoo!**, **Google**, and **Amazon**, have opened up their data to be used with other data sources **without a lengthy licensing negotiation**. In just a minute or two, you can set up and use the data resources they make available.
- The **other reason** for this **rapid growth** is the advent of new tools that make creating mashups easy for anyone, regardless of their technical know-how.



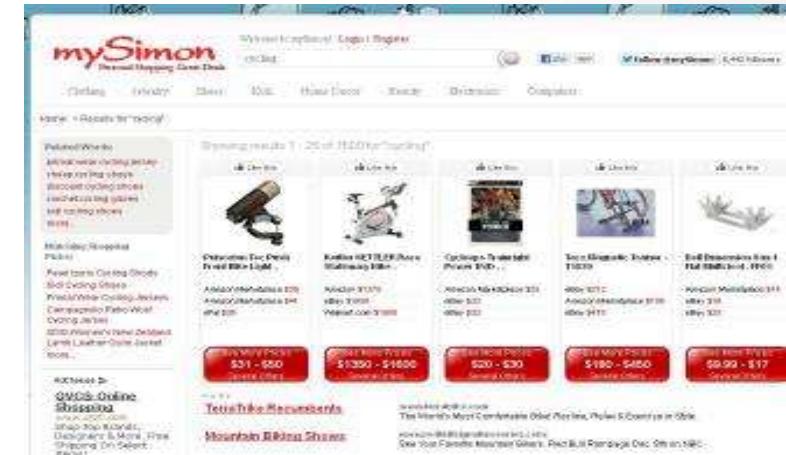
# Recent Common Services! (Similar to mashup)

- a) Web Portals: Yahoo! provide information of different kinds under a single unified theme.



# Recent Common Services! (Similar to mashup)

b) Web Information Aggregators: (MySimon) etc. provide price comparison services for many products.



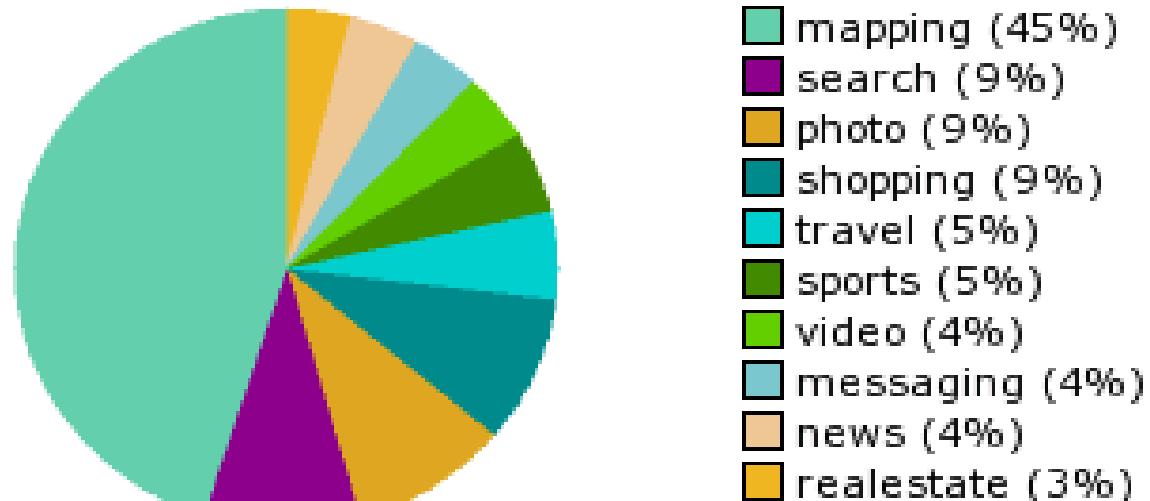
c) RSS (Really Simple Syndication): collect feeds from different news sites to create a news channel.





# Mashup Types

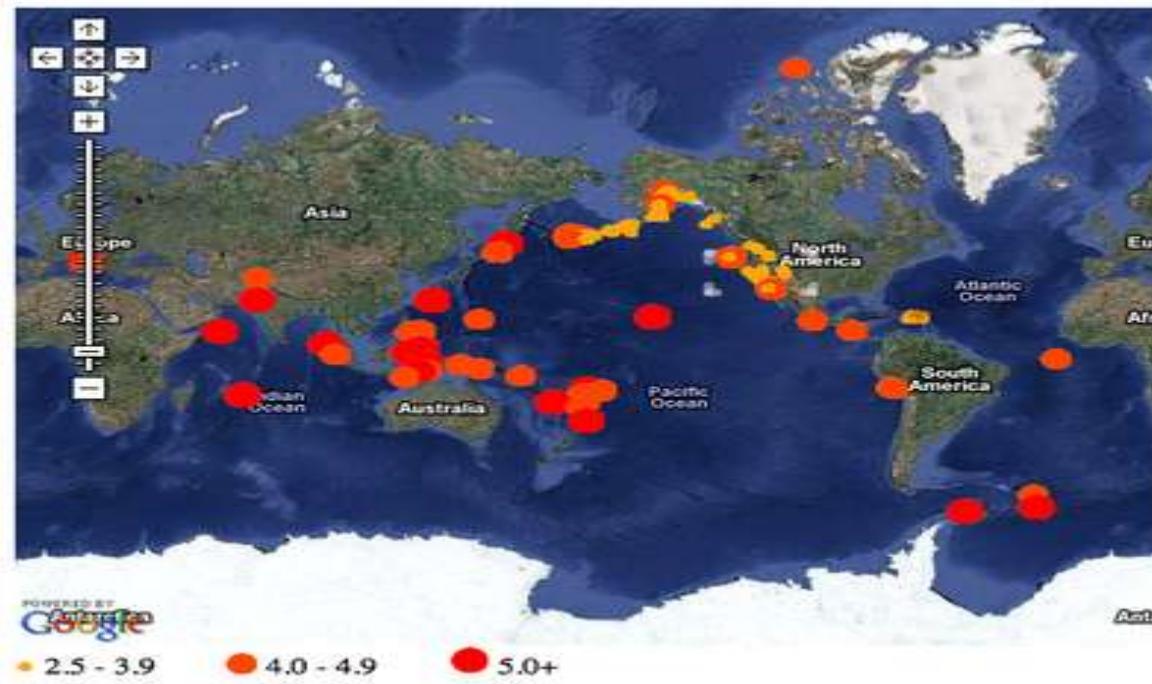
Mashups have several different colloquial interpretations, which has resulted in some confusion regarding the term and its use.



ProgrammableWeb.com 12/31/06

## 1. Mapping mashups

Mapping mashup allow users to navigate most of the globe through a Web interface, viewing varying levels of resolution through maps, satellite imagery, or a combination. One of the big catalysts for the advent of mashups was Google's introduction of its **Google Maps API**. This opened the floodgates, allowing Web developers to mash all sorts of data onto maps.



## 2. Video and photo mashups

The emergence of photo hosting and social networking sites like **Flickr** with APIs that expose photo sharing has led to a variety of interesting mashups. Because these content providers have **metadata** associated with the images they host (*such as who took the picture, what it is a picture of, where and when it was taken, and more*), mashup designers can mash photos with other information that can be associated with the metadata.



### 3. Search and Shopping mashups

To facilitate mashups and other interesting Web applications, consumer marketplaces such as **eBay** and **Amazon** have released APIs for programmatically accessing their content. Hundreds of shopping mashups exist ranging from commercial desktop applications to comparison shopping web sites to whimsical hacks.



## 4. News mashups

News sources (such as the **New York Times**, the **BBC**, or **Reuters**) have used syndication technologies like **RSS** since 2002 to disseminate news feeds related to various topics. Syndication feed mashups can aggregate a user's feeds and present them over the Web, **creating a personalized newspaper** that caters to the reader's particular interests.

**Overview**

**Associated Press**

**Description**  
A simple mashup that plots stories from the AP National News RSS feed on a Google Map. The Yahoo! Geocode API is used to convert the location of each story to a longitude/latitude point. Currently, only the U.S./National news stories are plotted.

**APIs** Google Maps + Yahoo Geocoding  
**Tags** mapping, news

**Added** 11 Dec 2005  
**Who** Michael Young  
**URL** <http://www.81nassau.com/apnew...>

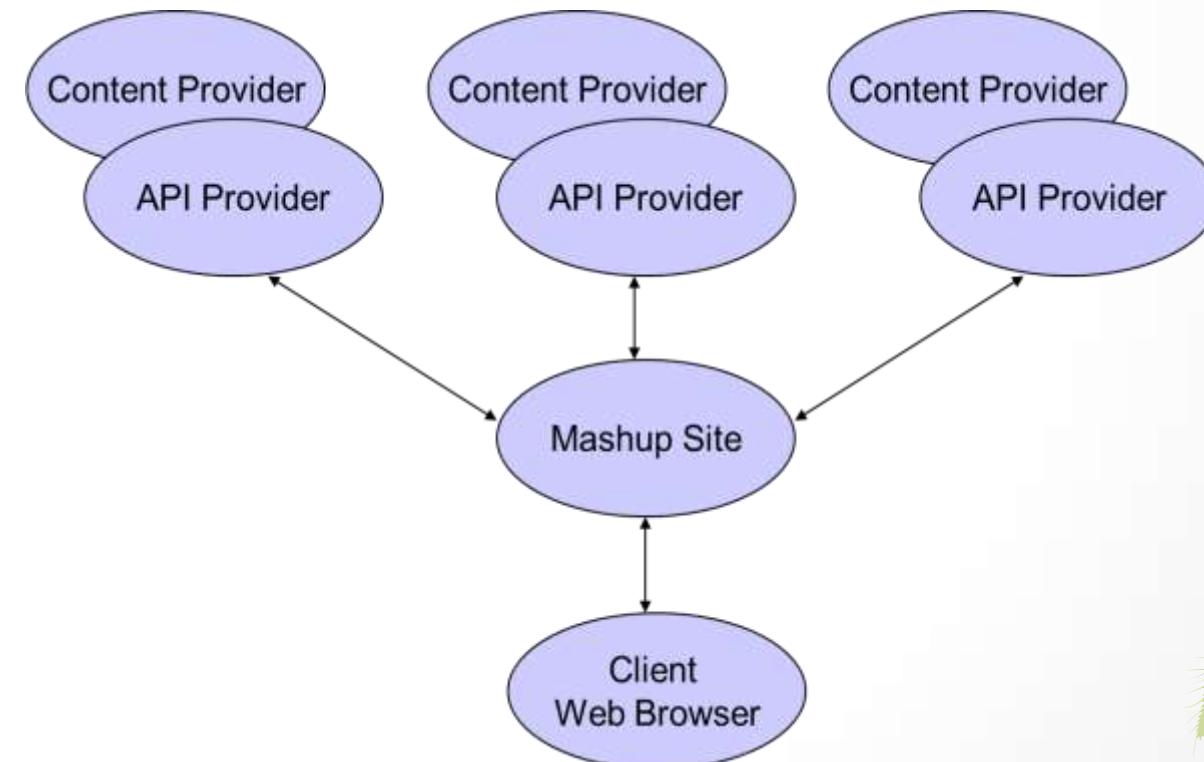
**See Related Mashups**

Click stars to vote

**BlueOkapi**

# Mashup architecture

A mashup application is architecturally comprised of three different participants that are logically and physically disjoint (they are likely separated by both network and organizational boundaries): **API/content providers**, the **mashup site**, and the **client's Web browser**.



- **The API/content providers.** These are the providers of the content being mashed. It denotes the process by which a tool attempts to extract information from the content provider by attempting to parse the provider's Web pages.
- **The mashup site.** This is where the mashup is hosted. Interestingly enough, just because this is where the mashup logic resides, it is not necessarily where it is executed. On one hand, mashups can be implemented similarly to traditional Web applications using **server-side** dynamic content generation technologies like **Java servlets**, **PHP** or **ASP**. The benefits of **client-side** mashing include less overhead on behalf of the mashup server (data can be retrieved directly from the content provider).
- **The client's Web browser.** This is where the application is rendered graphically and where **user interaction** takes place.

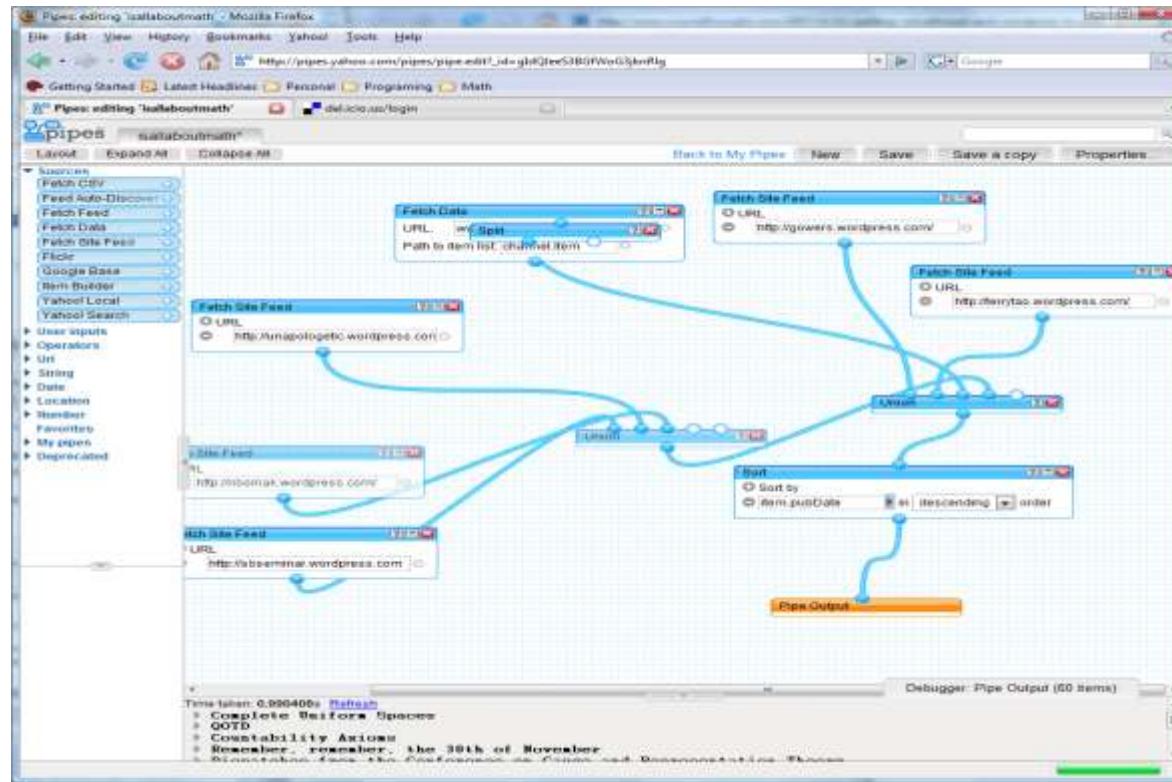
# Mashup tools

- **Microsoft Popfly:** Microsoft Popfly is Microsoft's Mashup Editor. Its very useful to an end-user and requires little technical understanding. Popfly is built on **Microsoft Silverlight**.



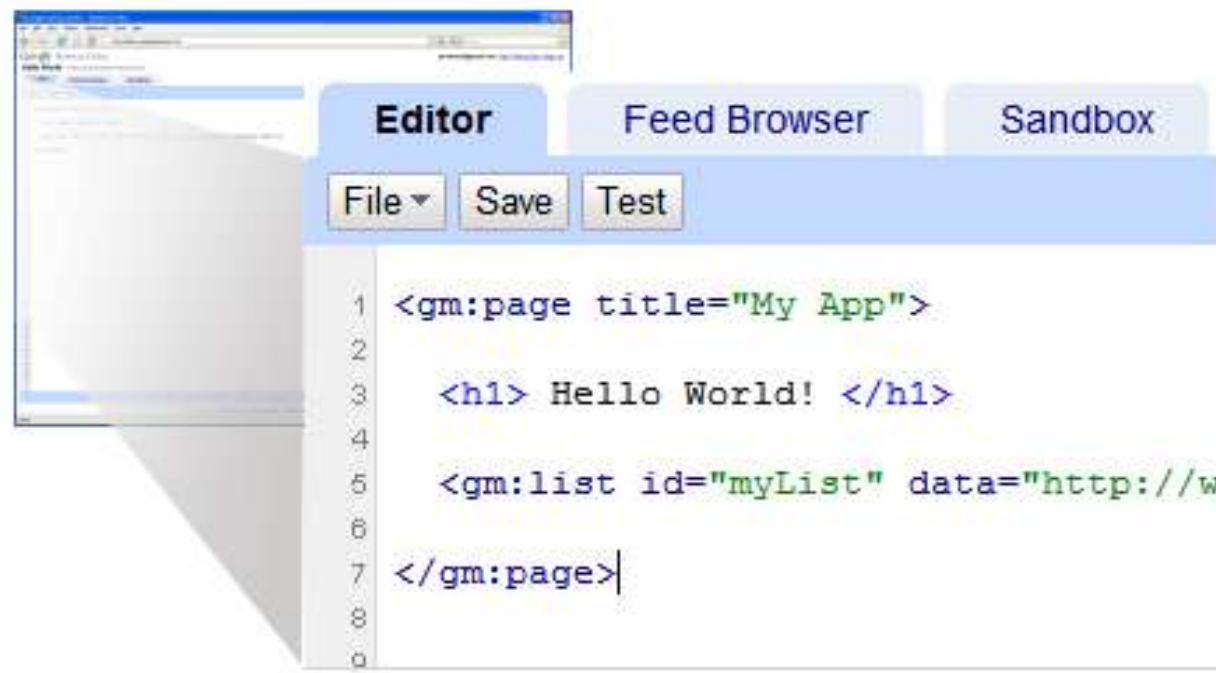
# Mashup tools

- **Yahoo Pipes:** Yahoo Pipes is **Yahoo's flash-based** tool to aggregate, manipulate, and mashup content from around the web. Yahoo Pipes was one of the first mashup editing tools to come out. It appears to be targeted to the slightly more technical people. However it has a **drag-drop interface**. Its quite easy to use.



# Mashup tools

- **Google Mashup Editor:** The Google Mashup Editor (**GME**), is defiantly the most advanced out of all the ones. Most powerful, and It's basically got a tag based **markup language**, that lets you also embed **HTML** into your results.



The screenshot shows the Google Mashup Editor interface. At the top, there are three tabs: "Editor" (which is selected), "Feed Browser", and "Sandbox". Below the tabs is a toolbar with "File ▾", "Save", and "Test" buttons. The main area is a code editor displaying the following XML-like code:

```
1 <gm:page title="My App">
2
3   <h1> Hello World! </h1>
4
5   <gm:list id="myList" data="http://w
6
7 </gm:page>
```

# Pros and Cons

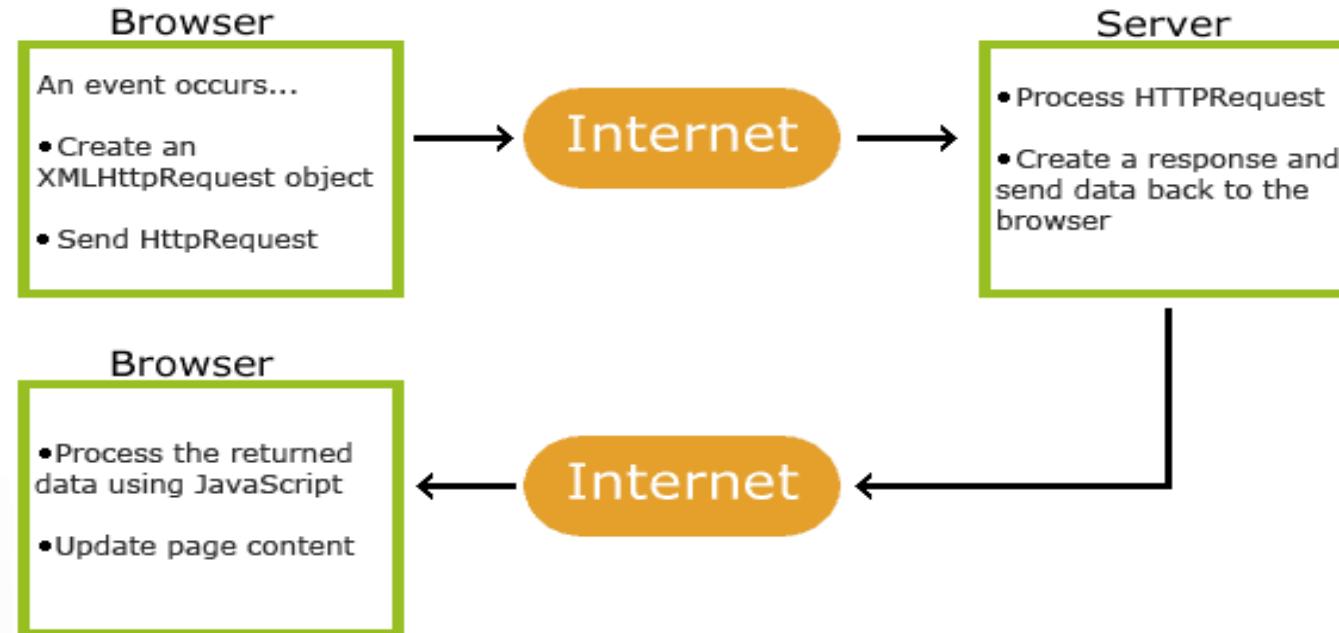


- **Innovation potential:** fusion of multiple services. More services at a *low cost with reusable components*.
- **Use of Open APIs:** allow to diffuse content of service adapted to the *needs of users*.
- **Security problems:** aggregation of own site with application coming from *unknown source* using APIs not fully understood by “developer”.



# Ajax (Asynchronous JavaScript + XML)

AJAX stands for Asynchronous JavaScript and XML. It is a Web programming technique that exchanges small amounts of information behind the scenes to speed up applications that run on the Web. It is the use of the [XMLHttpRequest](#) object to communicate with **server-side** scripts. It can send as well as receive information in a variety of formats, including XML, HTML, and even text files. AJAX's most appealing characteristic, however, is its "**asynchronous**" nature, which means it can do all of this **without having to refresh the page**. This lets you update portions of a page based upon *user events*.



# **Itinerary Planner: A Mashup Case Study**

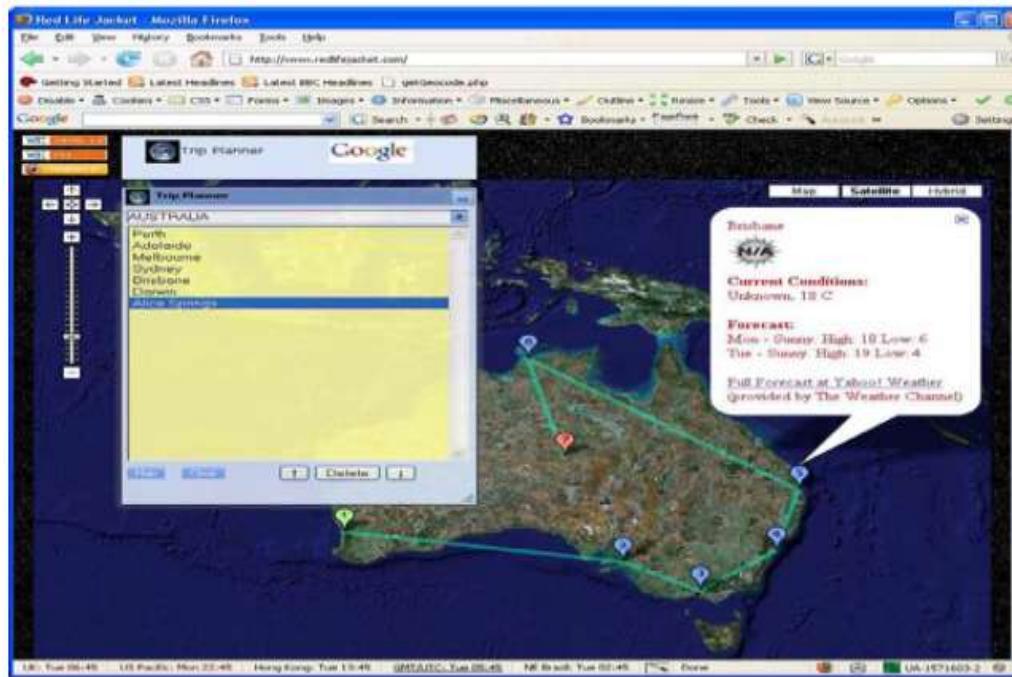
**Itinerary Planner Mashup (IPM).** IPM is a Single-page application that allows users to create an itinerary of the destinations (cities) that they plan to visit and display these destinations on a map. Users can display additional information about each destination including weather data (derived from **Yahoo Weather RSS feeds**) and other local information (derived using the **Google Ajax Search API**).



# Itinerary Planner: A Mashup Case Study

## 1. User Interface Design

The Trip Planner floating pane allows users to type the country name into a **combo box** widget; the Trip Planner then populates the **list-box** with the list of major cities and charts the route on the map as illustrated in the Figure below. When users select (click on) a place marker on the map, the application displays the current weather information using the Yahoo weather RSS feeds.



# Itinerary Planner: A Mashup Case Study

Selecting a destination from the list-box and clicking the “Google” icon causes the application to launch a new floating pane containing the search results that relate to the selected destination. The search floating pane includes **Video, Blog, News, Book, Local and Web** results as shown in the Figure below. The place markers on the map are numbered according to the order specified in the Trip Planner list and the colors indicate the start (green) and end (red) of the trip.



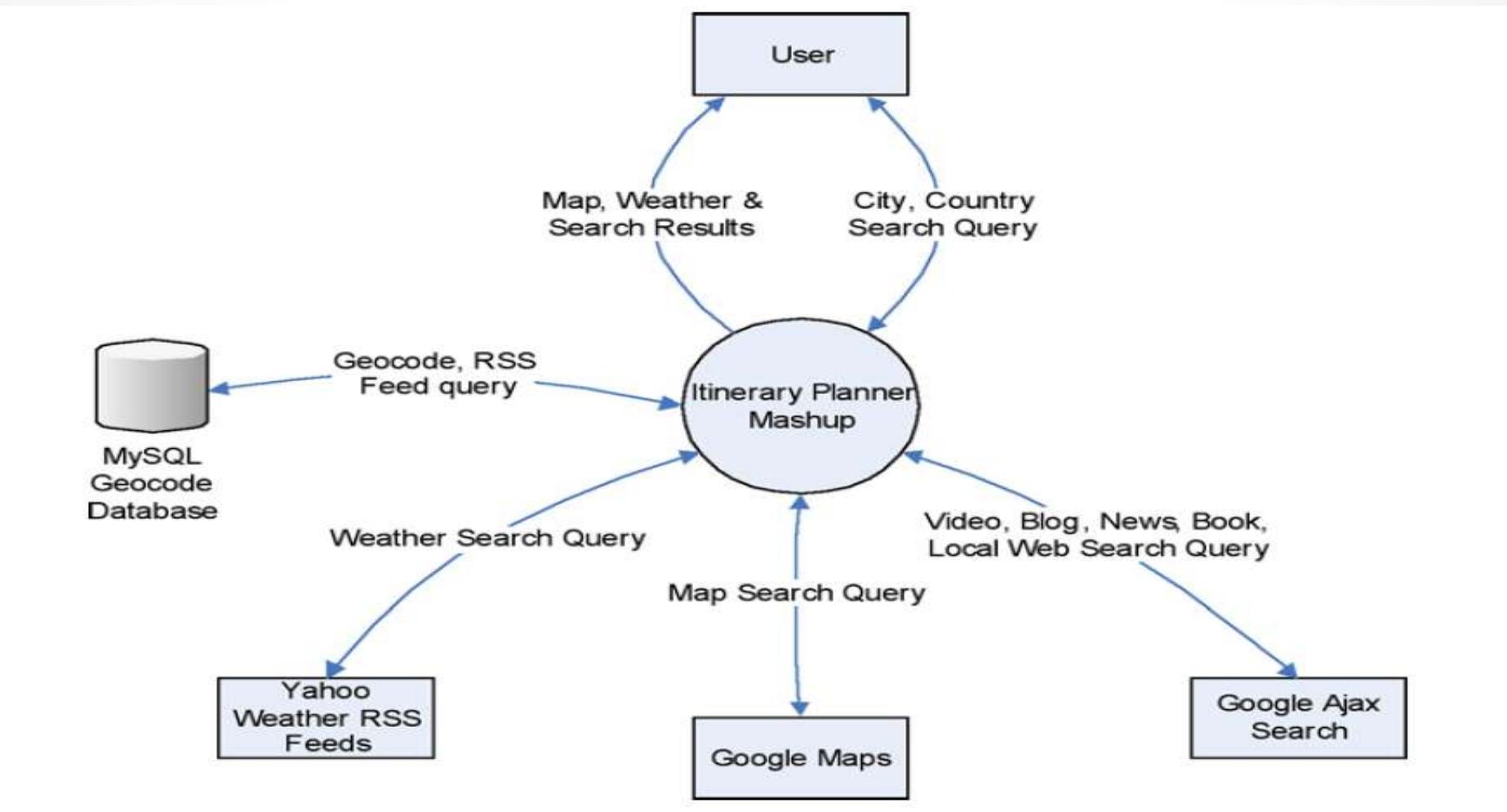
# **Itinerary Planner: A Mashup Case Study**

## **2. Data Integration Design**

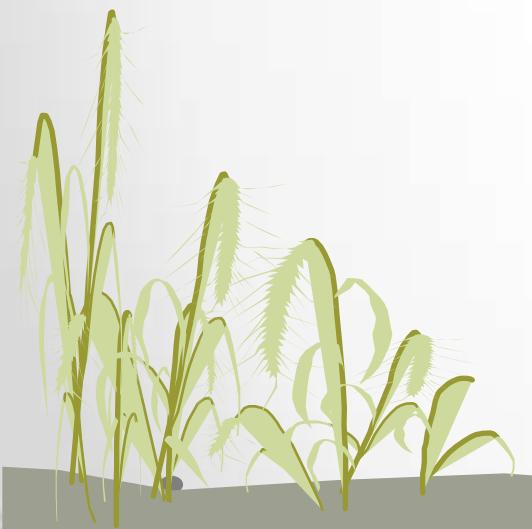
The **Google Maps API** retrieves Google Maps, the **Google Ajax Search API** allows customized display of Google Search results and the **Yahoo RSS** Weather feeds provide the current weather and forecast information. While the Google Maps API and the Google Ajax Search API are accessed directly by the **client-side Javascripts**, the Yahoo RSS feeds are retrieved by a **server-side PHP script**. The PHP XML\_RSS Parser was used to process the RSS feeds



# Itinerary Planner: A Mashup Case Study



# *Thank You*





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

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

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

# 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)
  - ✓ 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
  - ✓ PCs
  - ✓ 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.

# 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



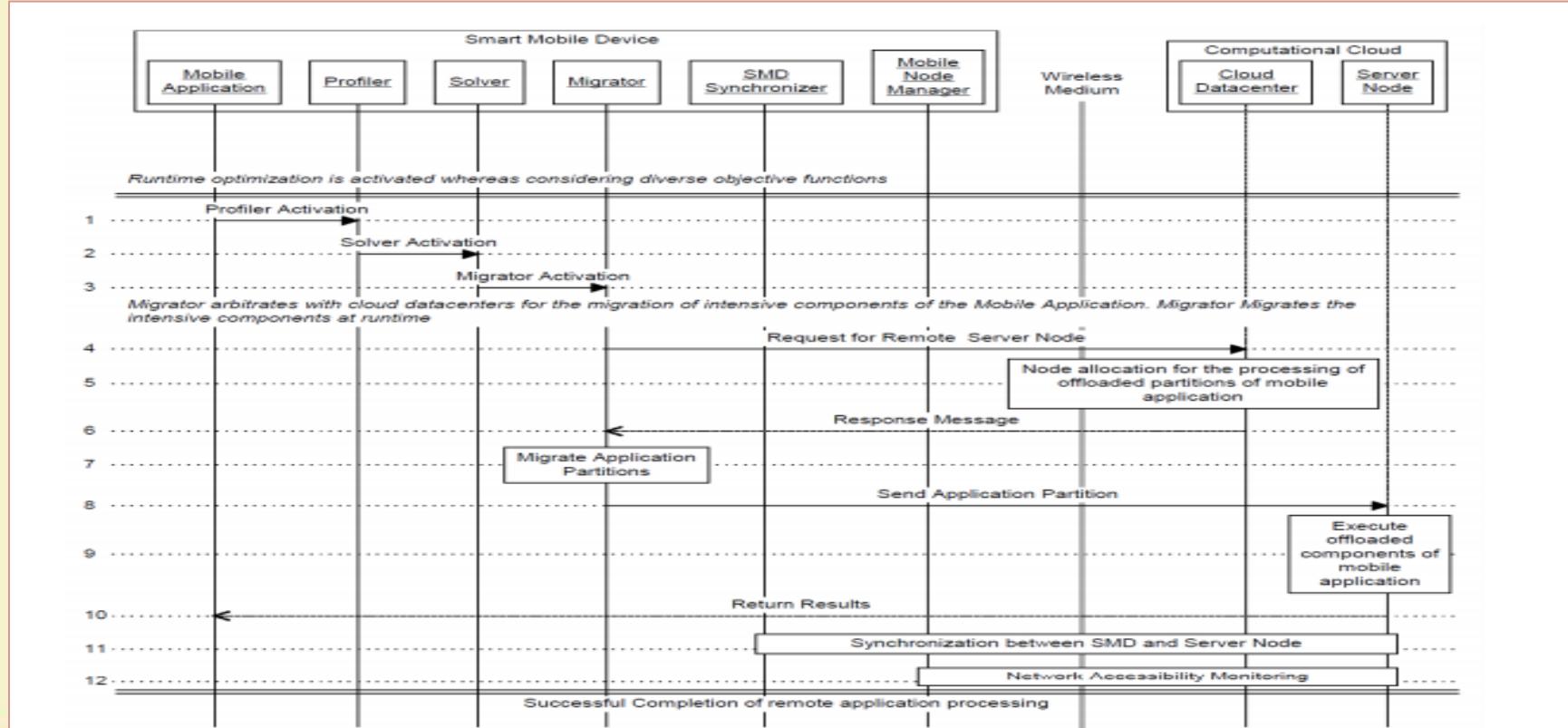
IIT KHARAGPUR



NPTEL  
NPTEL ONLINE  
CERTIFICATION COURSES

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

# 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

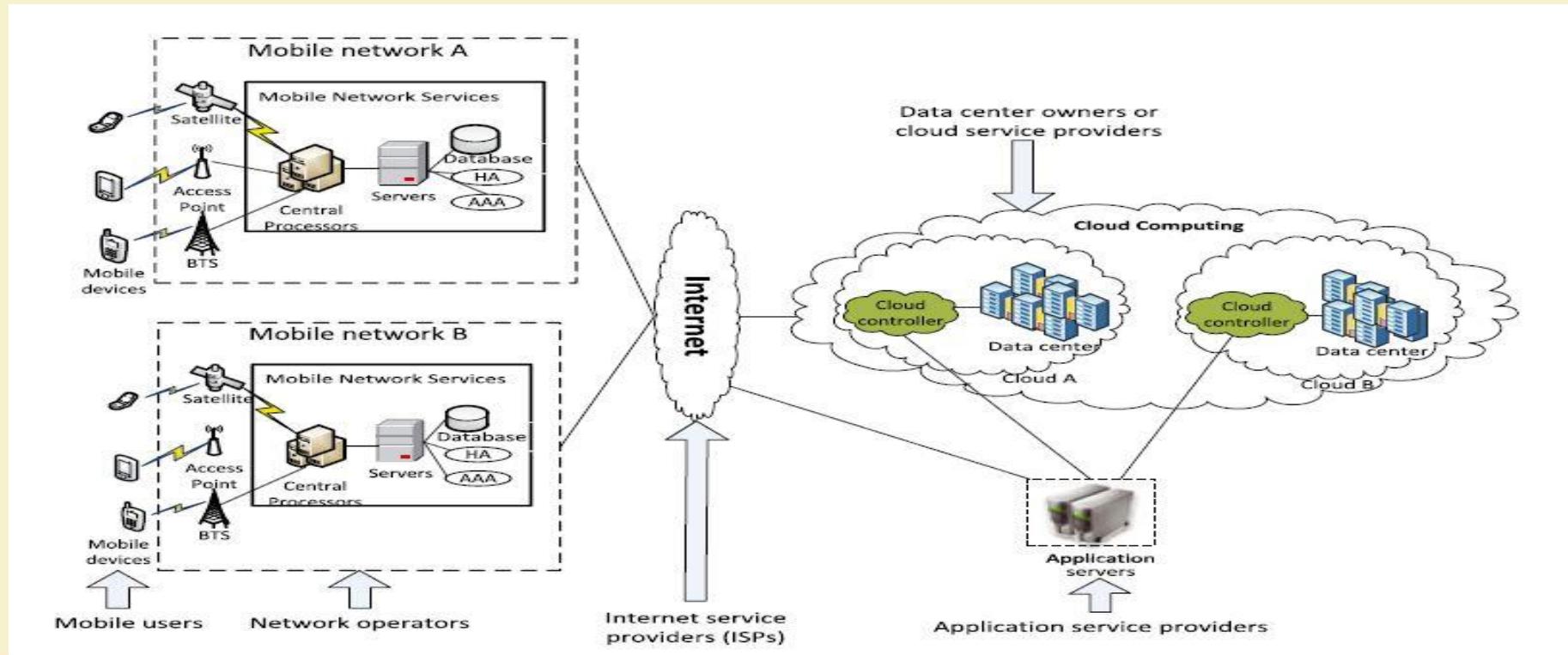


IIT KHARAGPUR



NPTEL ONLINE  
CERTIFICATION COURSES

# Mobile Cloud Computing – Typical Architecture



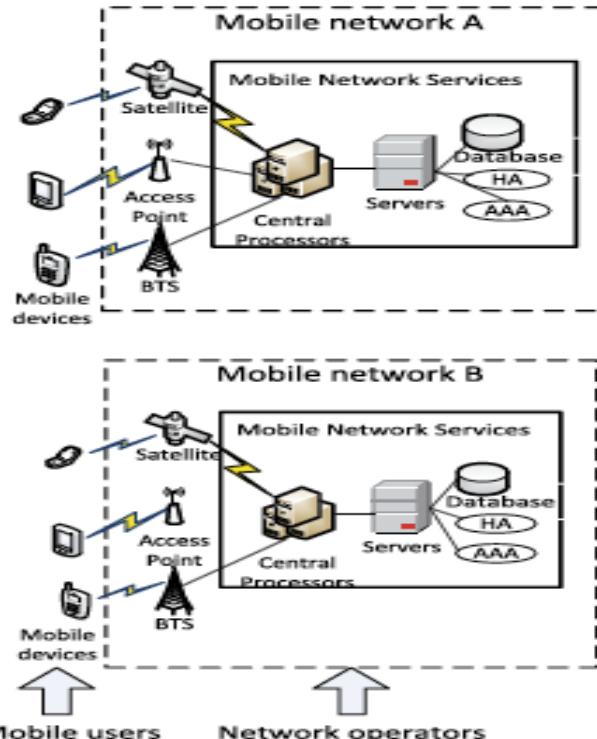
IIT KHARAGPUR



NPTEL  
ONLINE  
CERTIFICATION COURSES

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.

# 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

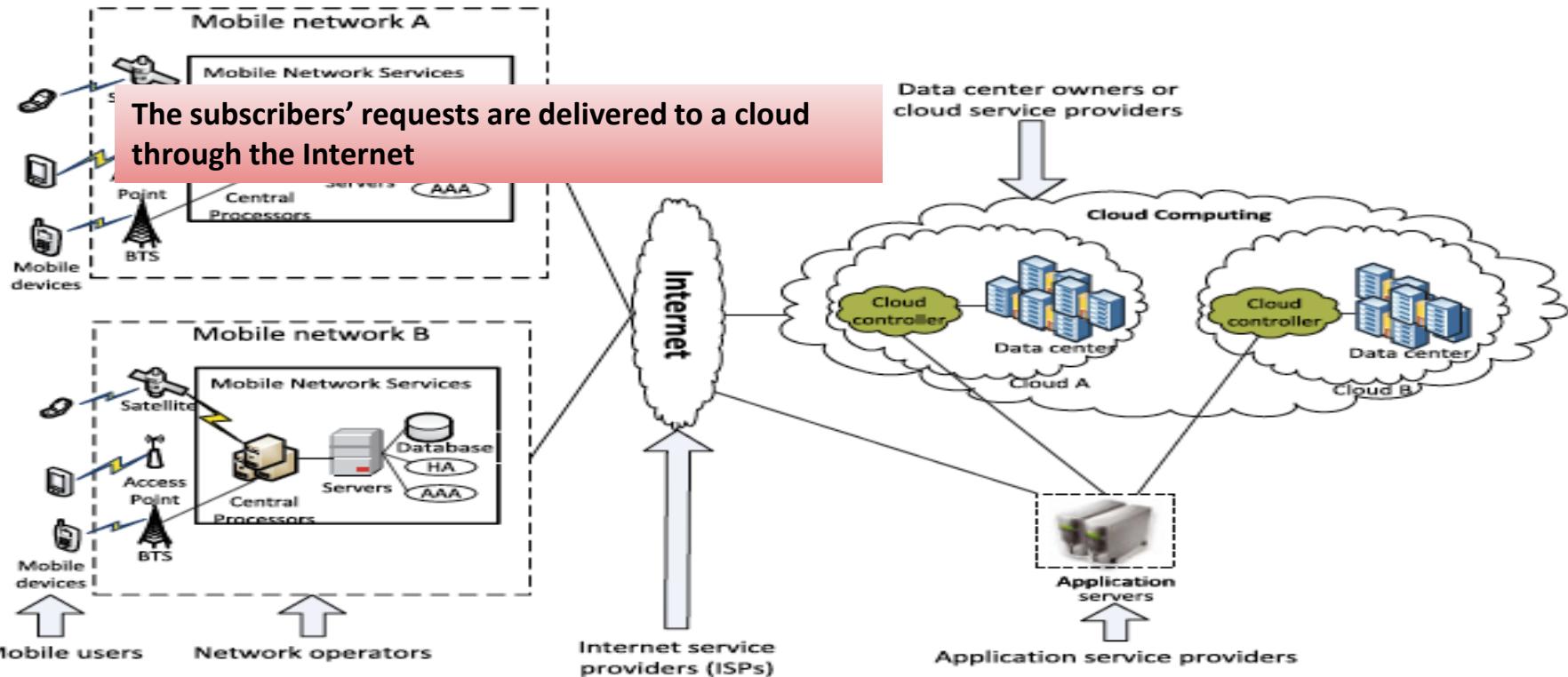


IIT KHARAGPUR



NPTEL  
NPTEL ONLINE  
CERTIFICATION COURSES

# Mobile Cloud Computing - Architecture

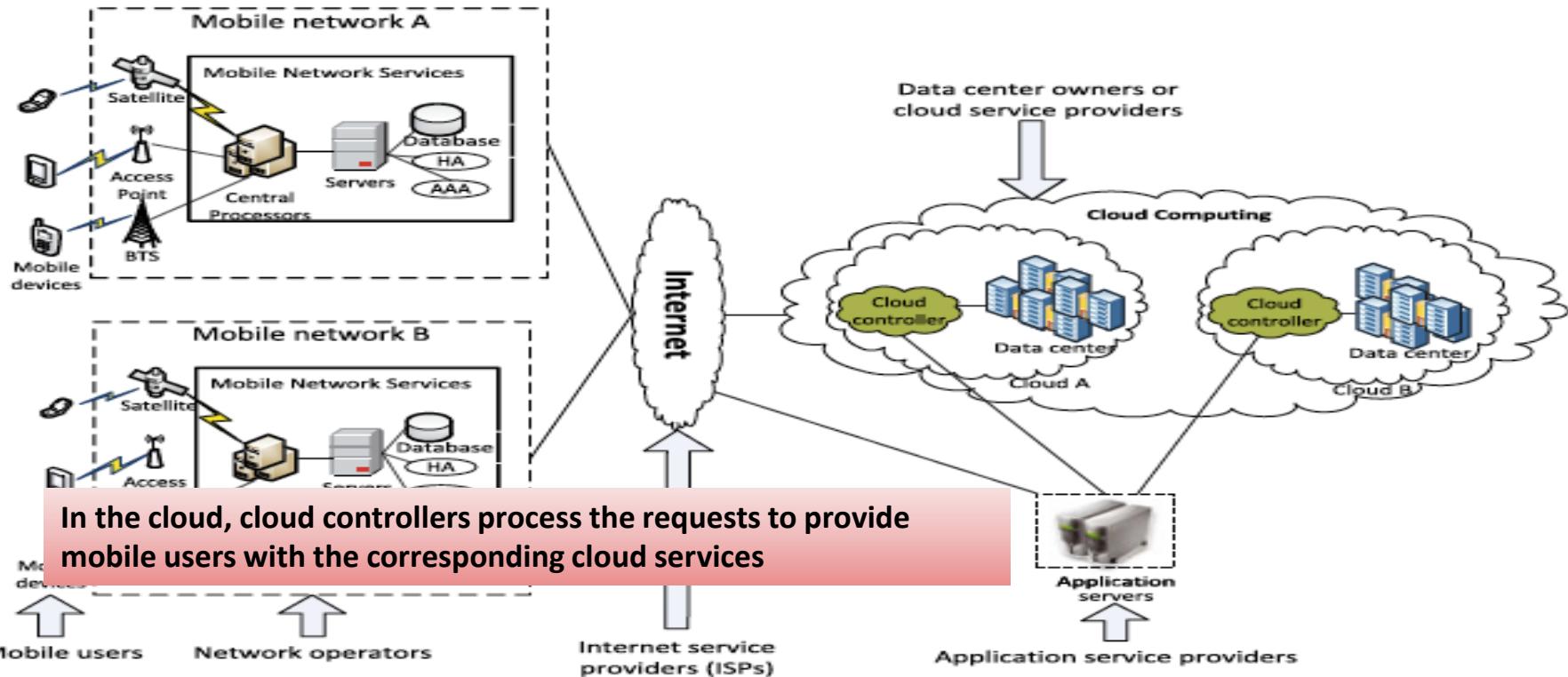


IIT KHARAGPUR



NPTEL  
NPTEL ONLINE  
CERTIFICATION COURSES

# Mobile Cloud Computing - Architecture



IIT KHARAGPUR



NPTEL  
NPTEL ONLINE  
CERTIFICATION COURSES

# 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.



IIT KHARAGPUR



NPTEL  
NPTEL ONLINE  
CERTIFICATION COURSES

# 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.

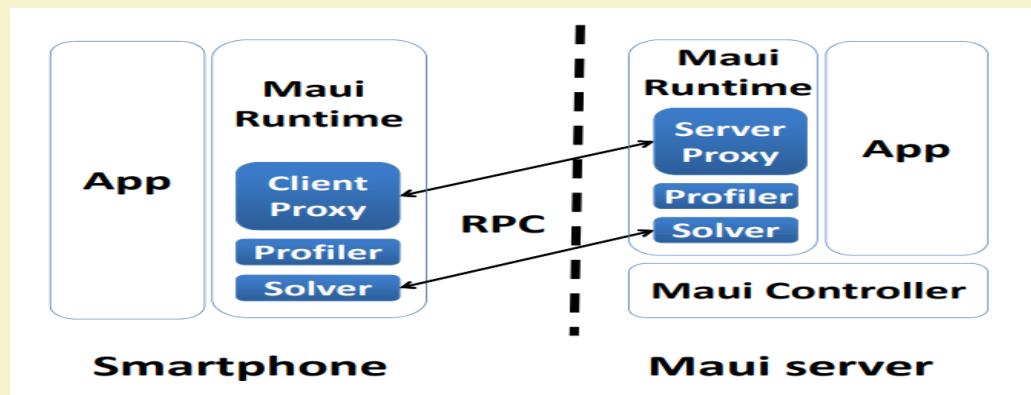
## ***Service Convergence:***

- 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

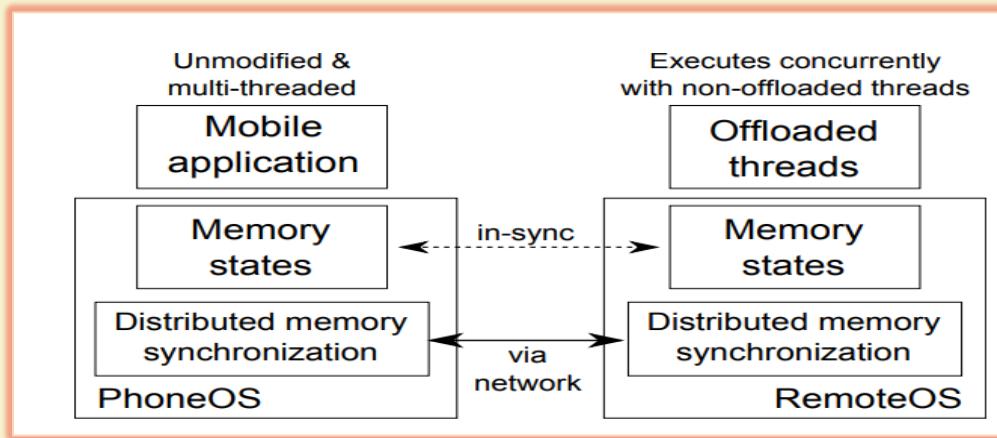
# MCC Systems: MAUI (Mobile Assistance Using Infrastructure)



- **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 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.

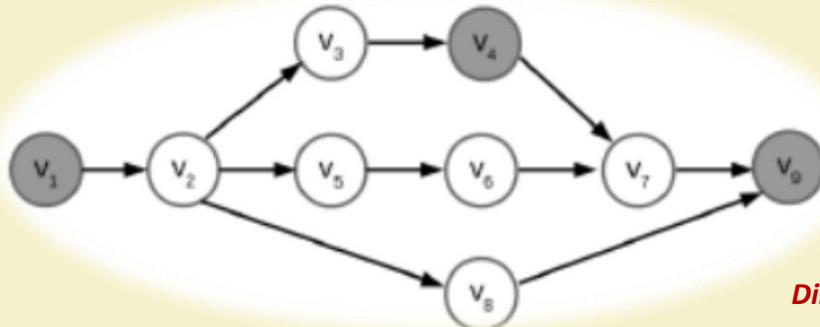


IIT KHARAGPUR



NPTEL ONLINE  
CERTIFICATION COURSES

# Mathematical Formulation



*Directed Acyclic Graph represents an application Call Graph*

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

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

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

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

- 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

## CODE OFFLOADING USING CLOUDLET

- **CLOUDLET:**

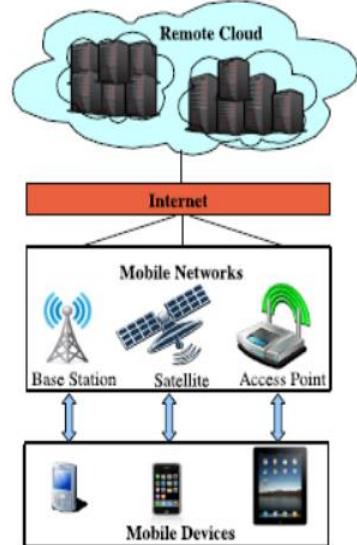
- ✓ “*a trusted, resource-rich computer or cluster of computers that is well-connected to the Internet and is available for use by nearby mobile devices.*”

- **Code Offloading :**

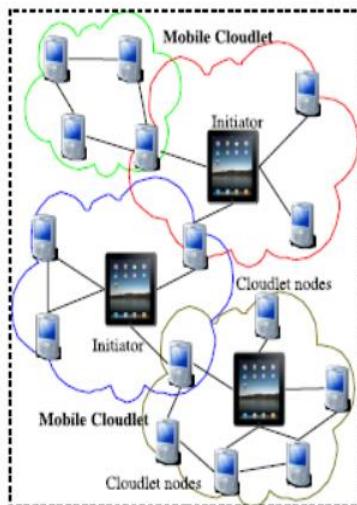
- ✓ Offloading the code to the remote server and executing it.
  - ✓ This architecture decreases latency by using a single-hop network and potentially lowers battery consumption by using Wi-Fi or short-range radio instead of broadband wireless which typically consumes more energy.

# CODE OFFLOADING USING CLOUDLET

## Cloudlet

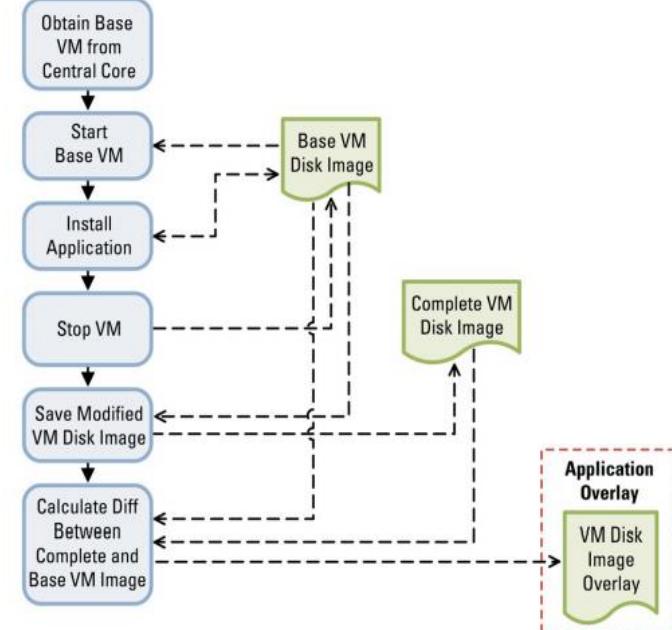


Use remote cloud



Use cloudlet

## Application Overlay Creation Process



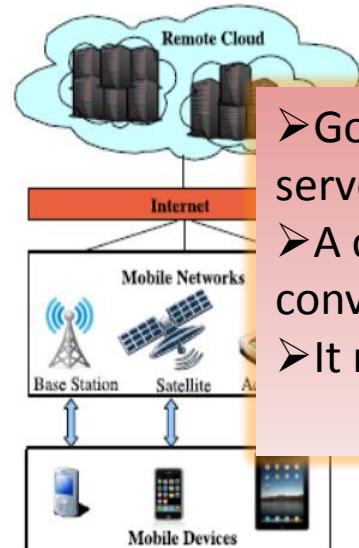
IIT KHARAGPUR



NPTEL  
ONLINE  
CERTIFICATION COURSES

# CODE OFFLOADING USING CLOUDLET

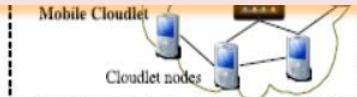
## Cloudlet



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***

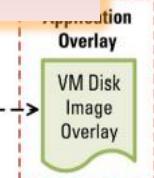


Use cloudlet

## *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<sub>c</sub>: the energy cost per second when the mobile phone is doing computing*

*P<sub>i</sub>: the energy cost per second when the mobile phone is idle.*

*P<sub>tr</sub>: 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.

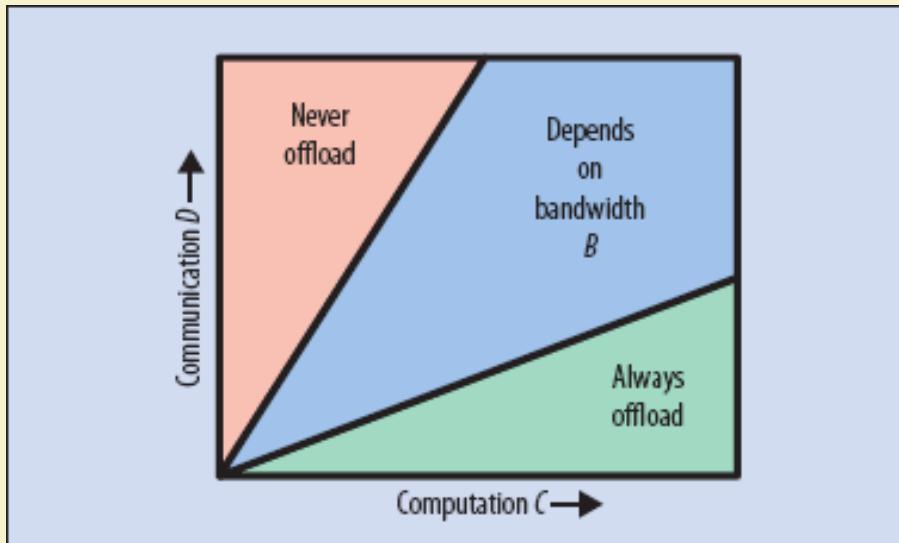
The amount of energy saved is :

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

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?

# 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*

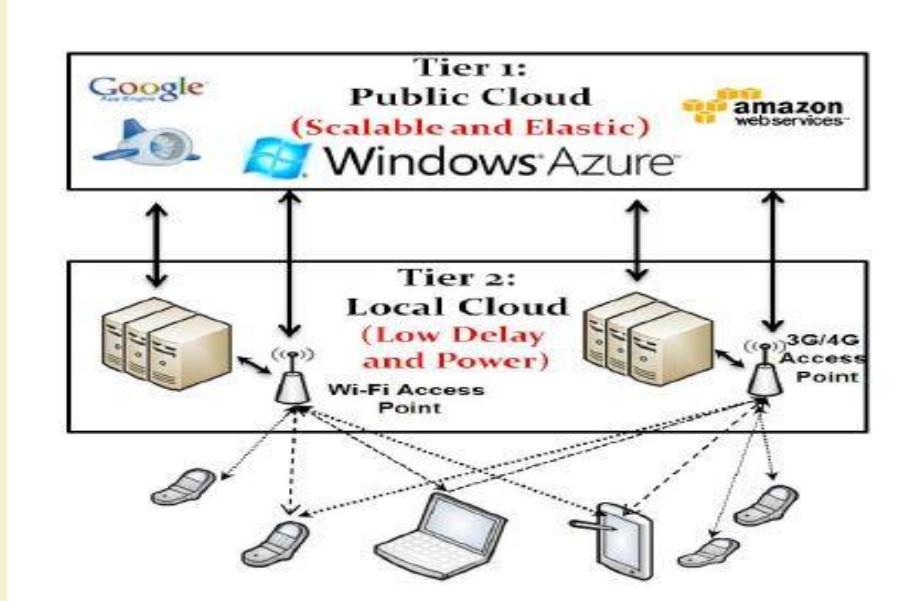
# MuSIC: Mobility-Aware Optimal Service Allocation in Mobile Cloud Computing

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  $\{(1; l_1); \dots; (n; l_m)\}$  where  $(i; l_i)$  implies that the mobile user is in location  $l_i$  for time duration  $i$

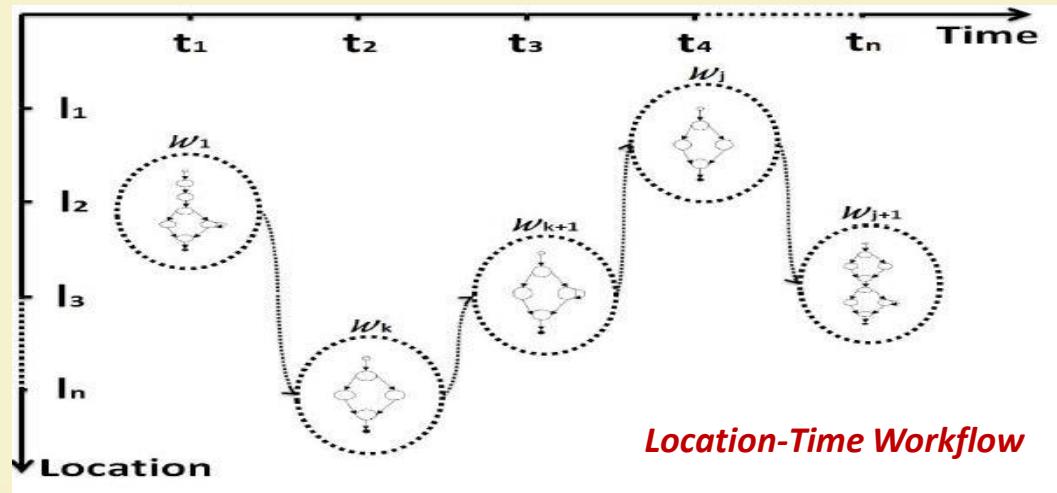
## Center of Mobility:

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.



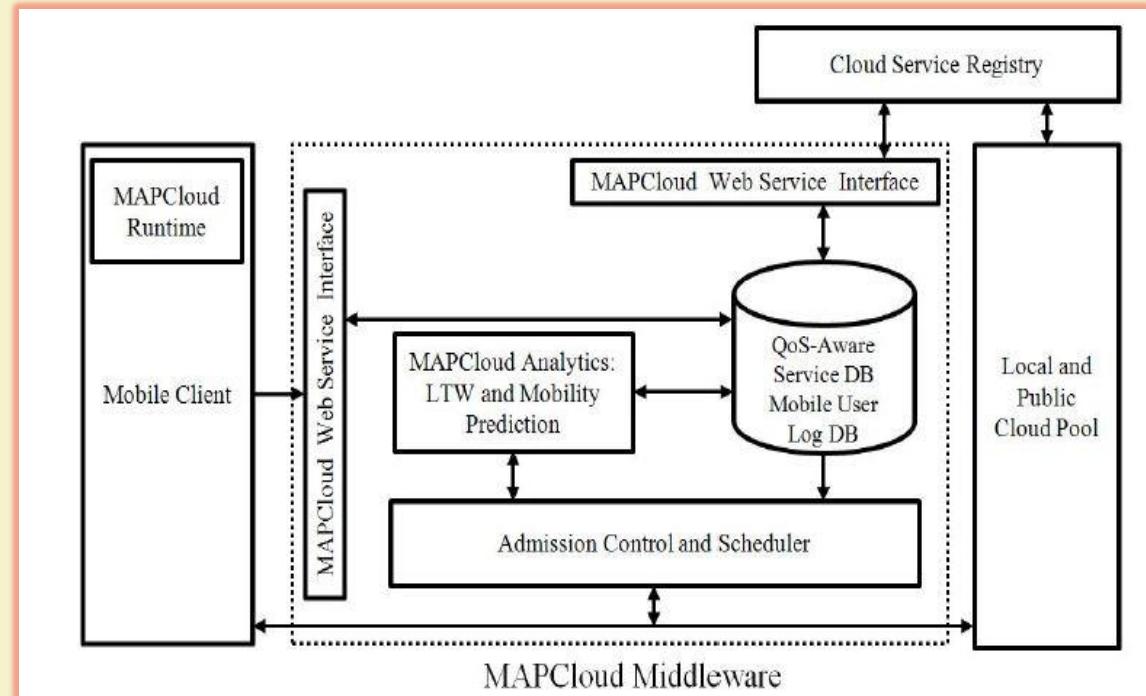
# Mobile Application Modelling

## 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.

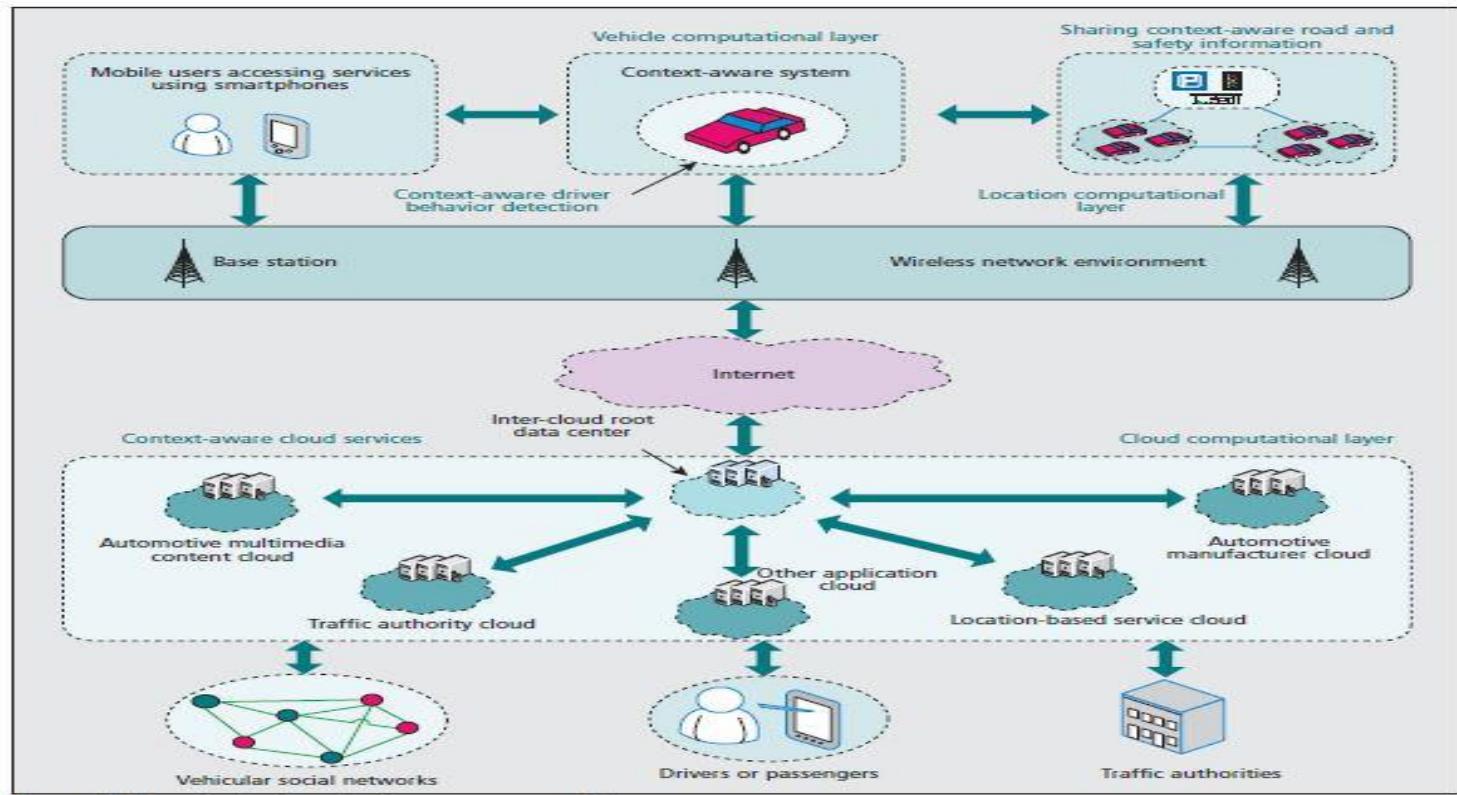
- As the number of vehicles increases, there is an increasing trend of insufficient parking spaces in many large cities, and this problem is gradually getting worse
- With the proliferation of wireless sensor networks (WSNs) and cloud computing, there exists strong potential to alleviate this problem 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

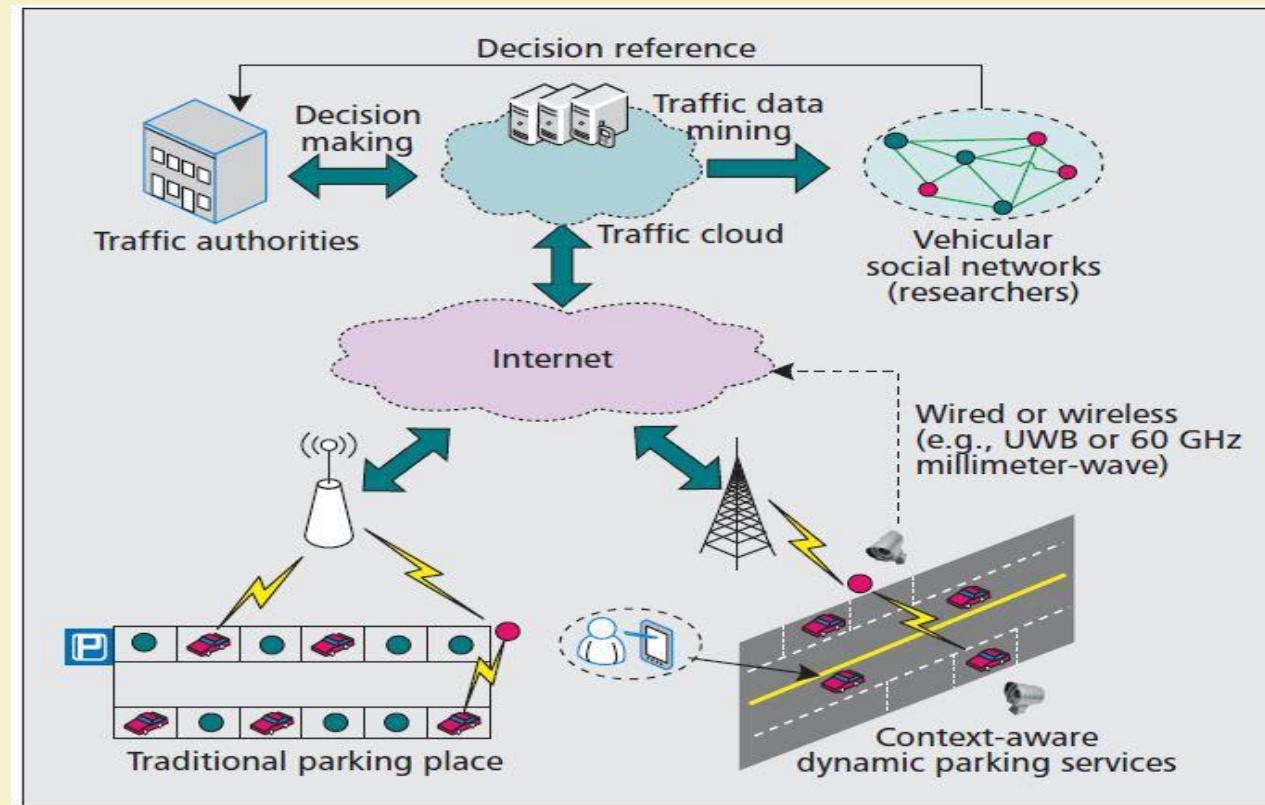


NPTEL  
ONLINE  
CERTIFICATION COURSES

Example cloud-assisted context-aware architecture

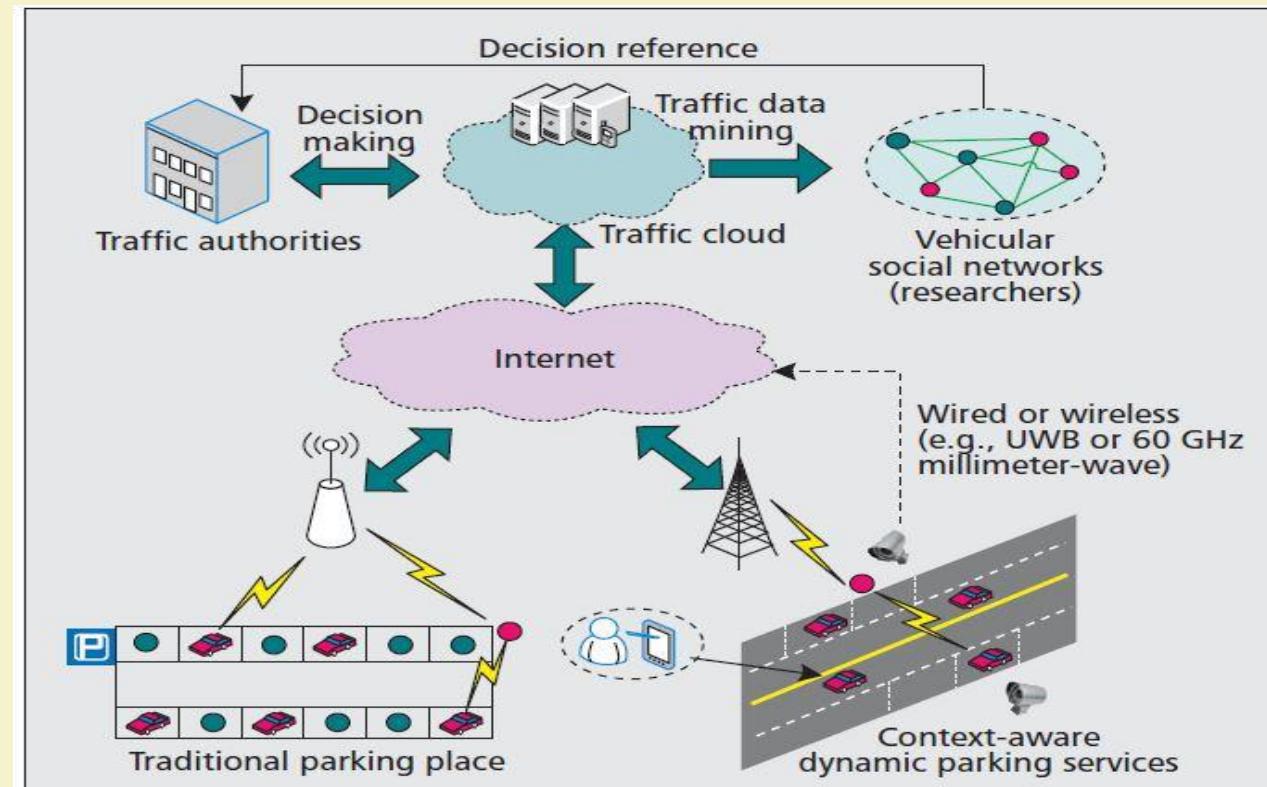
### Traditional parking garages:

- 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**

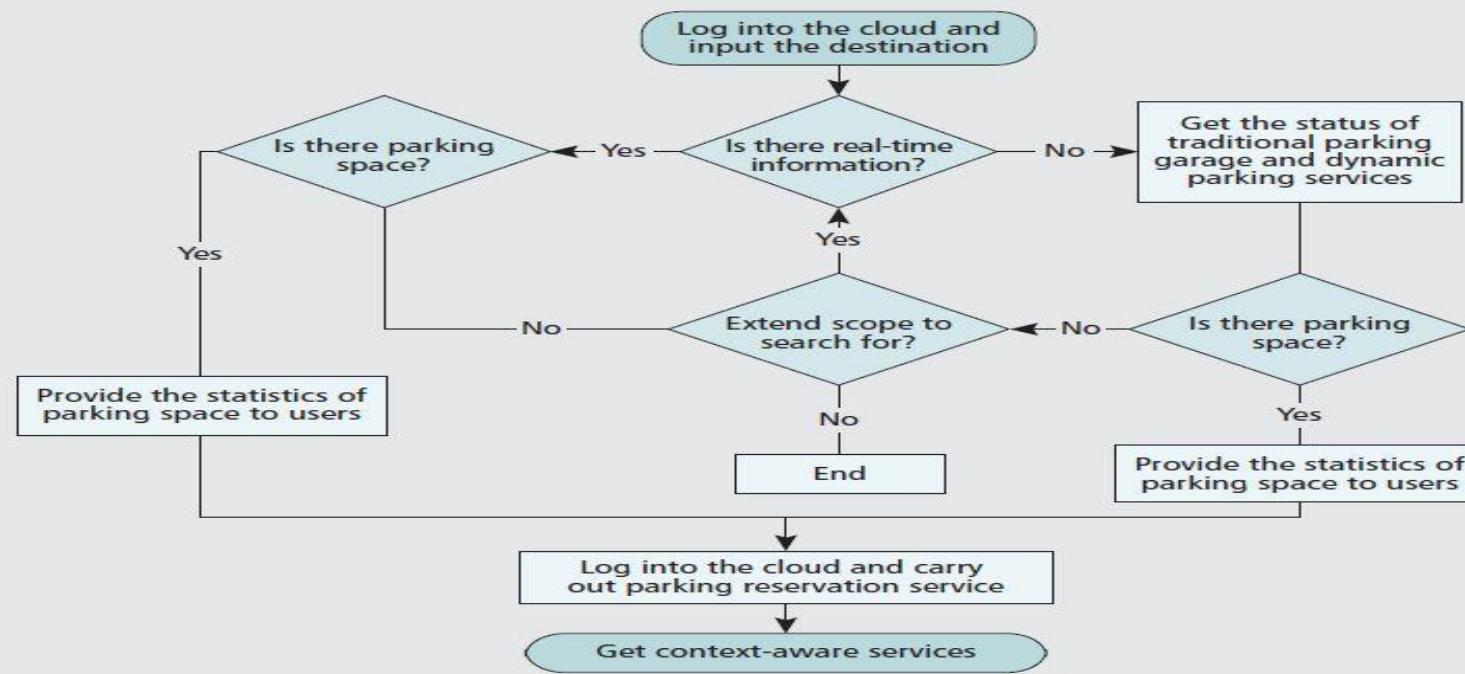
- The decision-making process of the proposed scheme heavily depends on many factors, such as historical traffic flow capacity, road conditions, weather conditions, and traffic flow forecasting
- In order to make an effective prediction, researchers on vehicular social networks carry out traffic data 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



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!!



IIT KHARAGPUR



NPTEL  
NPTEL ONLINE  
CERTIFICATION COURSES



IIT KHARAGPUR



NPTEL ONLINE  
CERTIFICATION COURSES

# CLOUD COMPUTING

## Mobile Cloud Computing - II

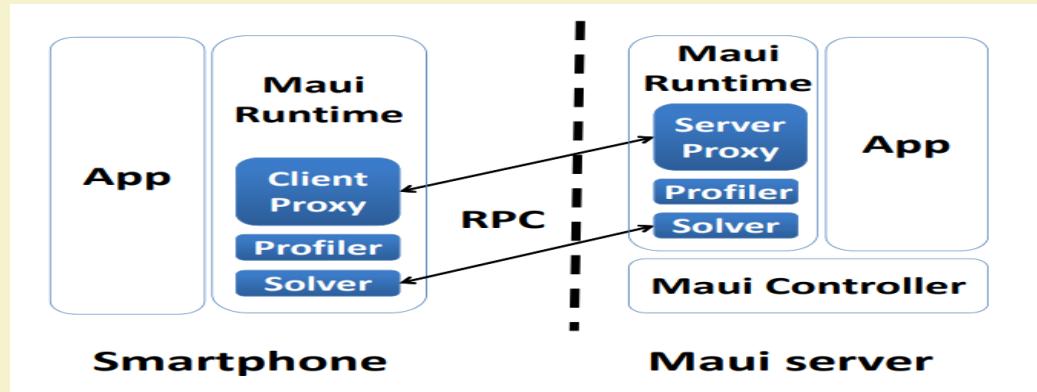
Prof. Soumya K Ghosh

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

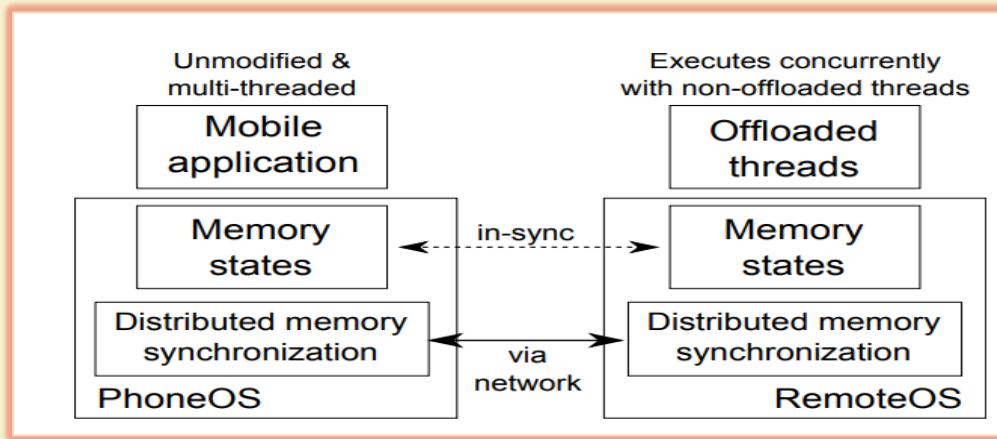
# MCC Systems: MAUI (Mobile Assistance Using Infrastructure)



- **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 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.

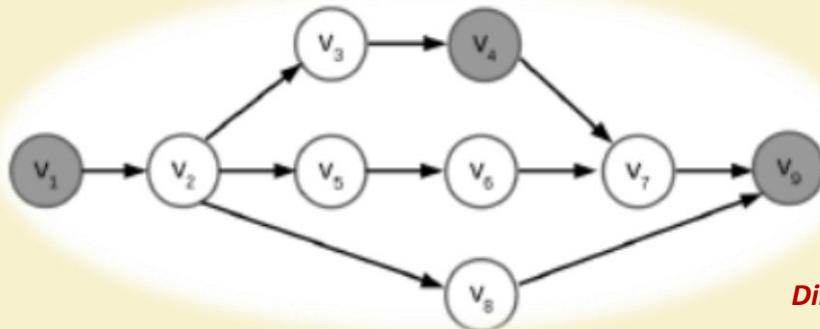


IIT KHARAGPUR



NPTEL ONLINE  
CERTIFICATION COURSES

# Mathematical Formulation



*Directed Acyclic Graph represents an application Call Graph*

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

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

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

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

- 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<sub>c</sub>: Energy cost per second when the mobile phone is doing computing*

*P<sub>i</sub>: Energy cost per second when the mobile phone is idle.*

*P<sub>tr</sub>: 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.

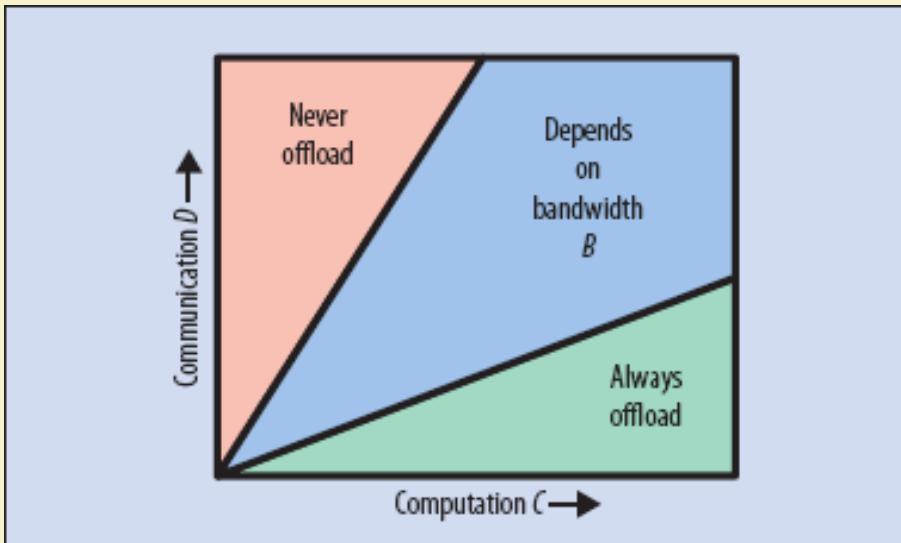
The amount of energy saved is :

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

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

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

# 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*

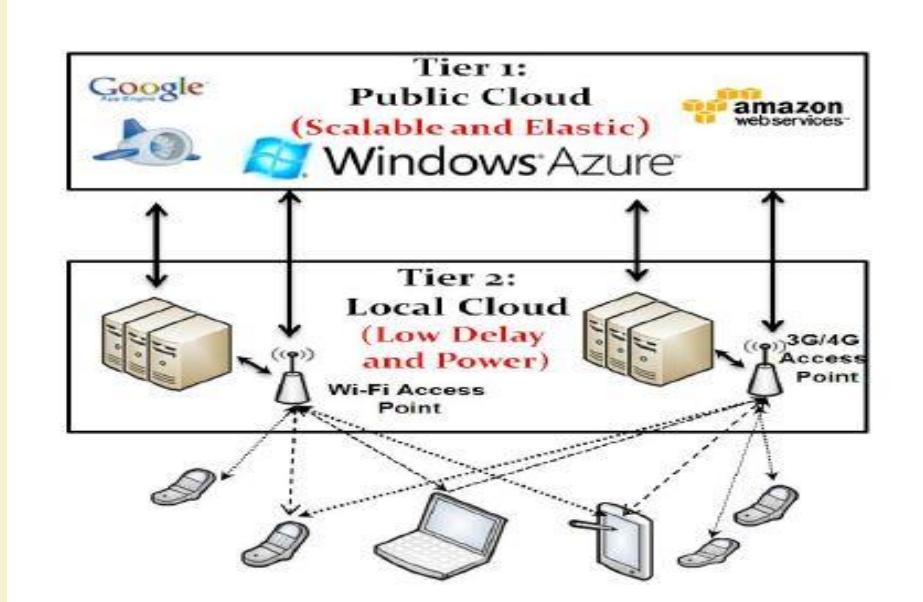
# MuSIC: Mobility-Aware Optimal Service Allocation in Mobile Cloud Computing

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  $\{(1; l_1); \dots; (n; l_m)\}$  where  $(i; l_i)$  implies that the mobile user is in location  $l_i$  for time duration  $i$

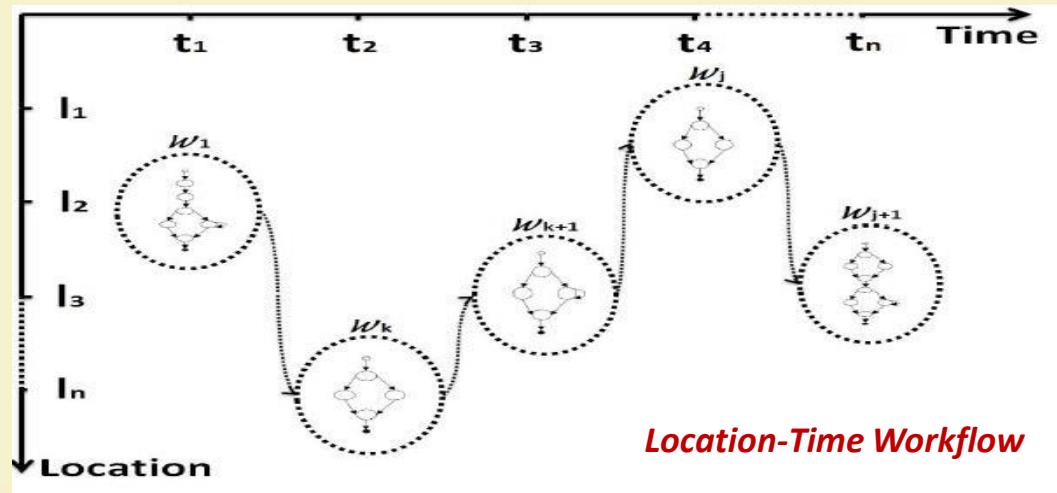
## Center of Mobility:

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.



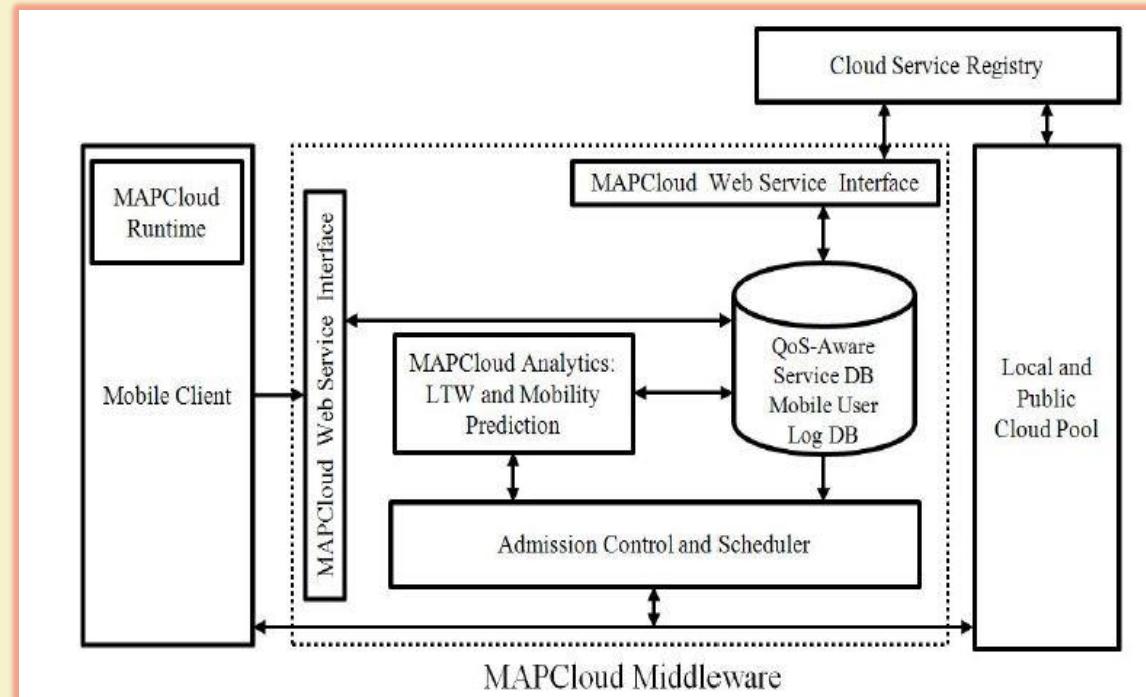
# Mobile Application Modelling

## 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.

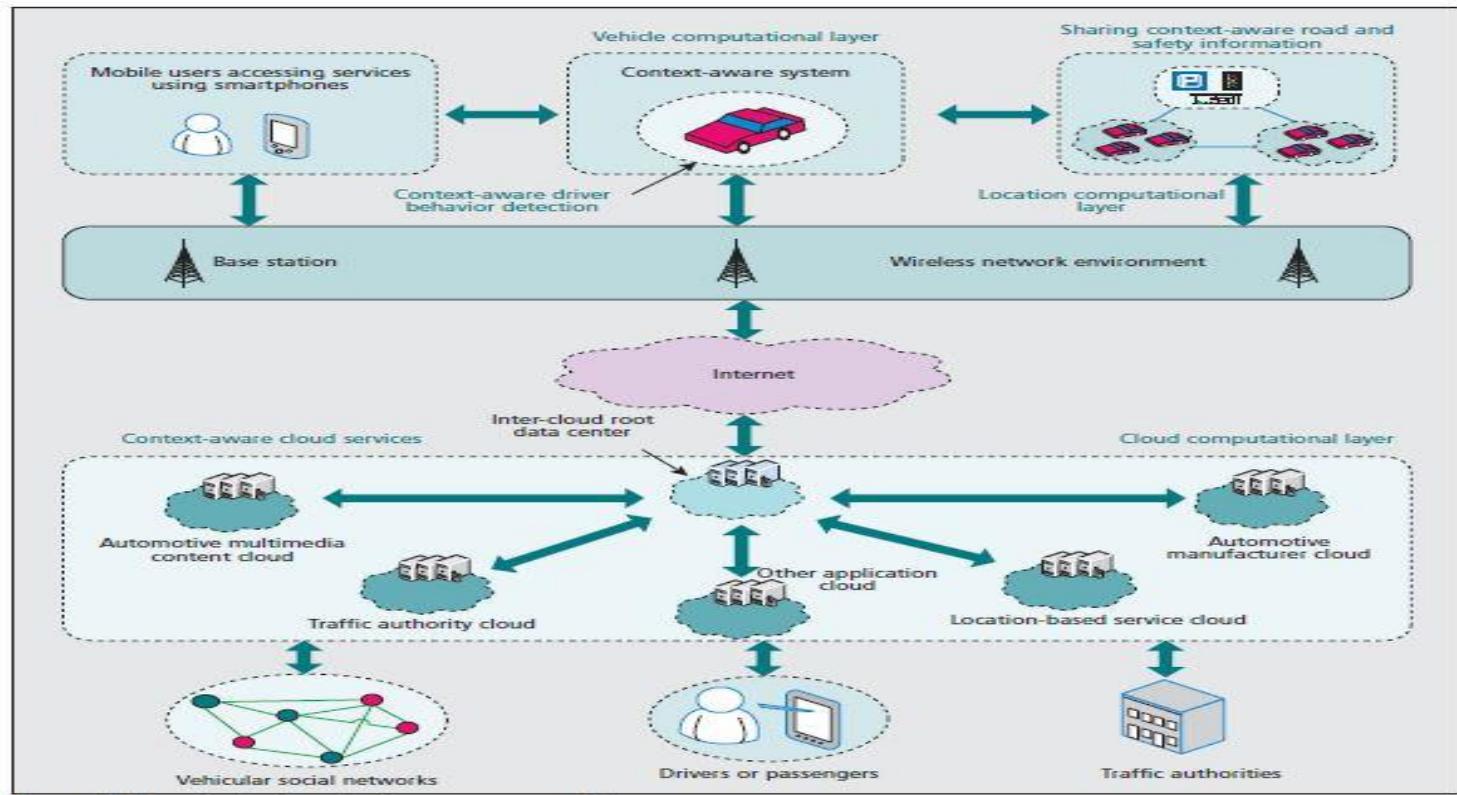
- As the number of vehicles increases, there is an increasing trend of insufficient parking spaces in many large cities, and this problem is gradually getting worse
- With the proliferation of wireless sensor networks (WSNs) and cloud computing, there exists strong potential to alleviate this problem 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

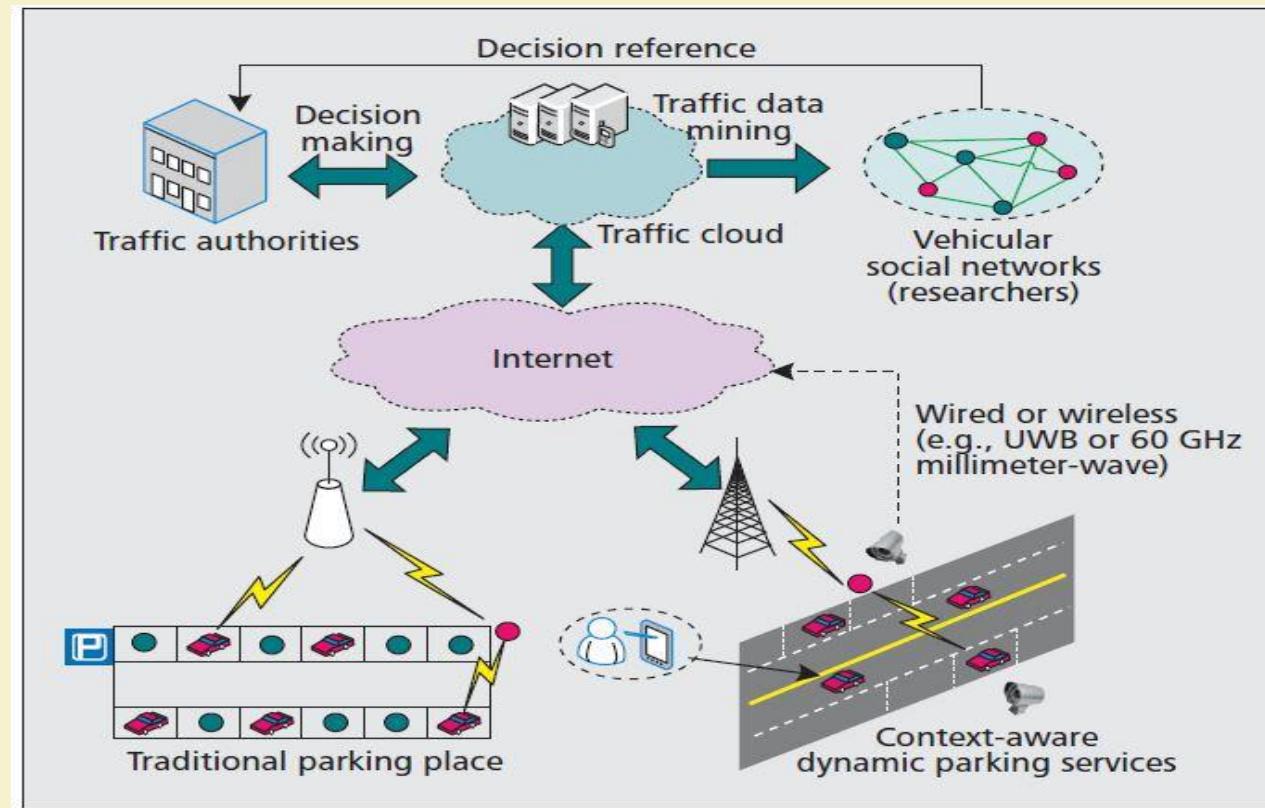


NPTEL  
ONLINE  
CERTIFICATION COURSES

Example cloud-assisted context-aware architecture

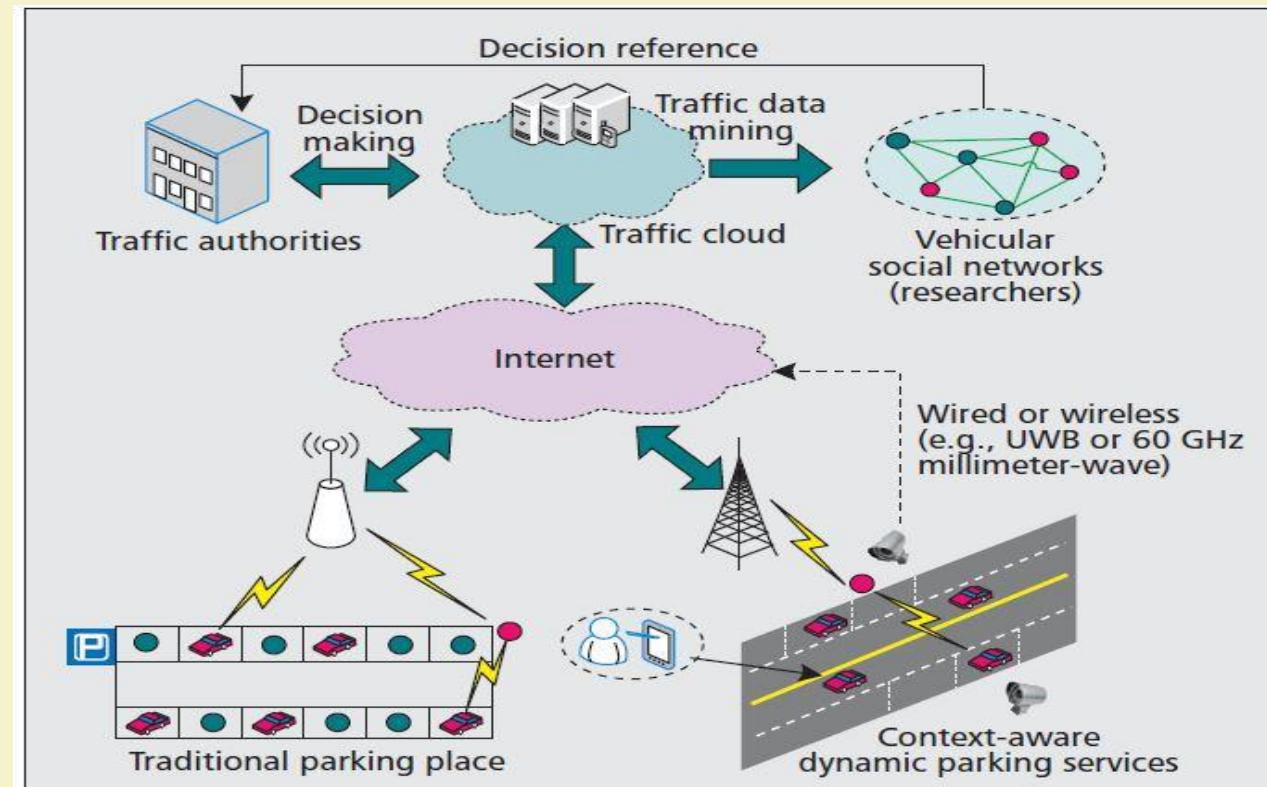
### Traditional parking garages:

- 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**

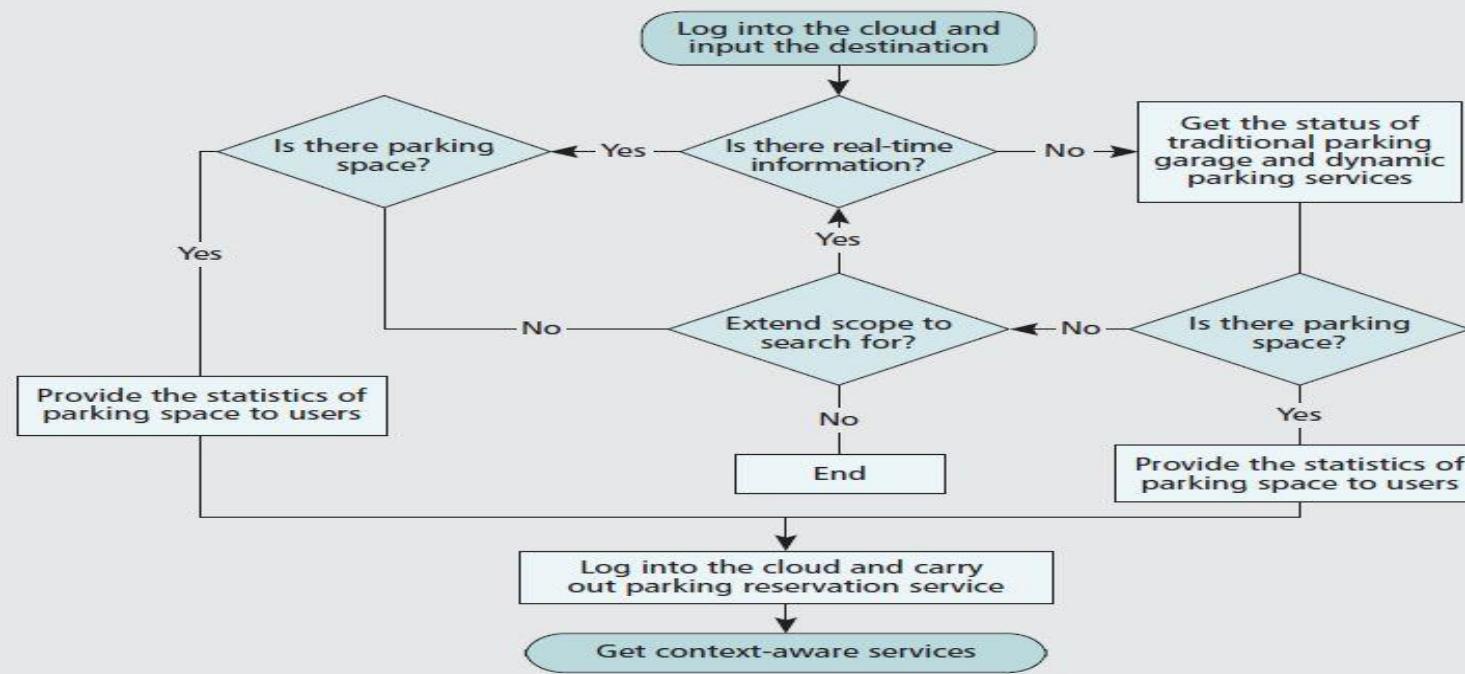
- The decision-making process of the proposed scheme heavily depends on many factors, such as historical traffic flow capacity, road conditions, weather conditions, and traffic flow forecasting
- In order to make an effective prediction, researchers on vehicular social networks carry out traffic data 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



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!!



IIT KHARAGPUR



NPTEL  
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  
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



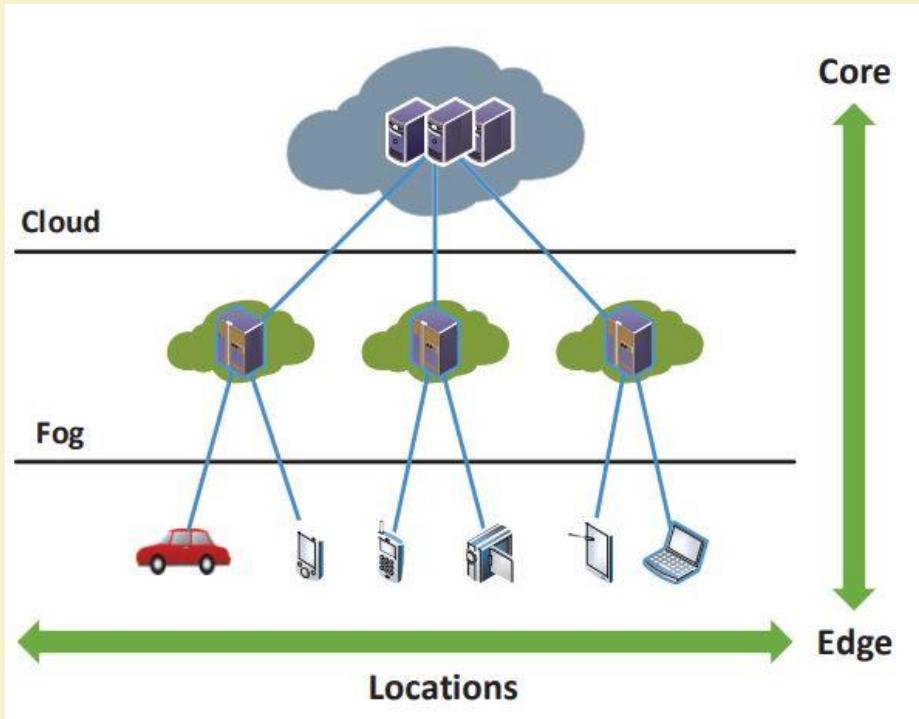
IIT KHARAGPUR



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  
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

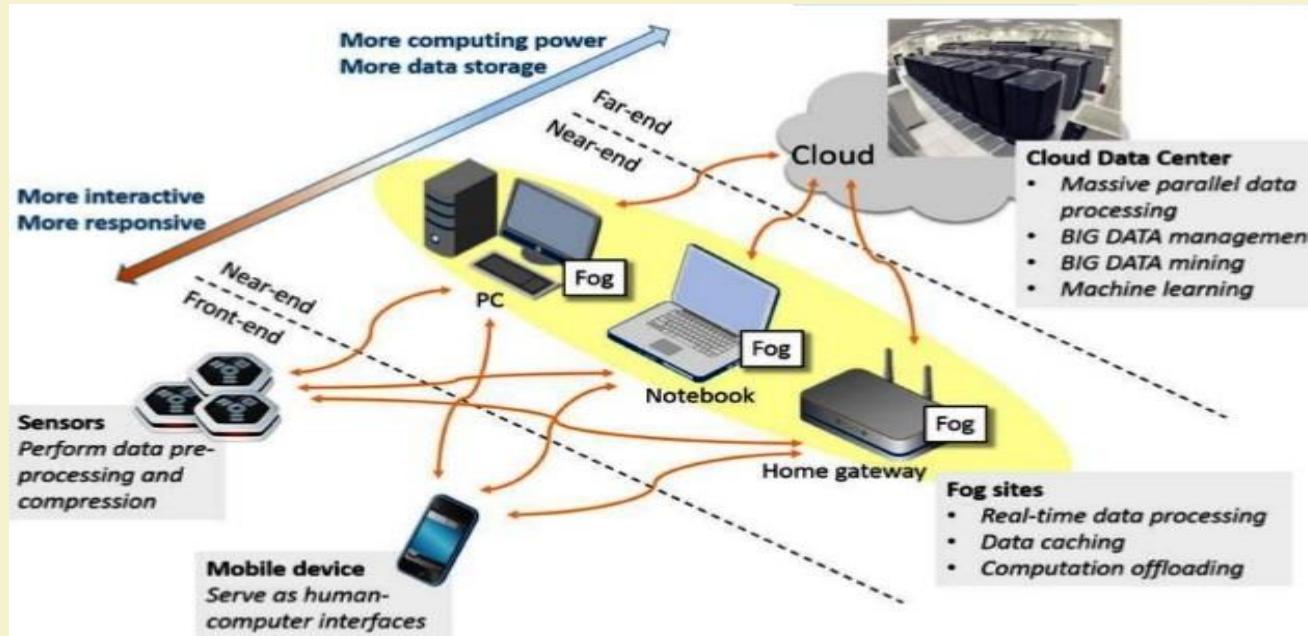


IIT KHARAGPUR



NPTEL  
NPTEL ONLINE  
CERTIFICATION COURSES

# Fog Computing



Source: Internet



IIT KHARAGPUR



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

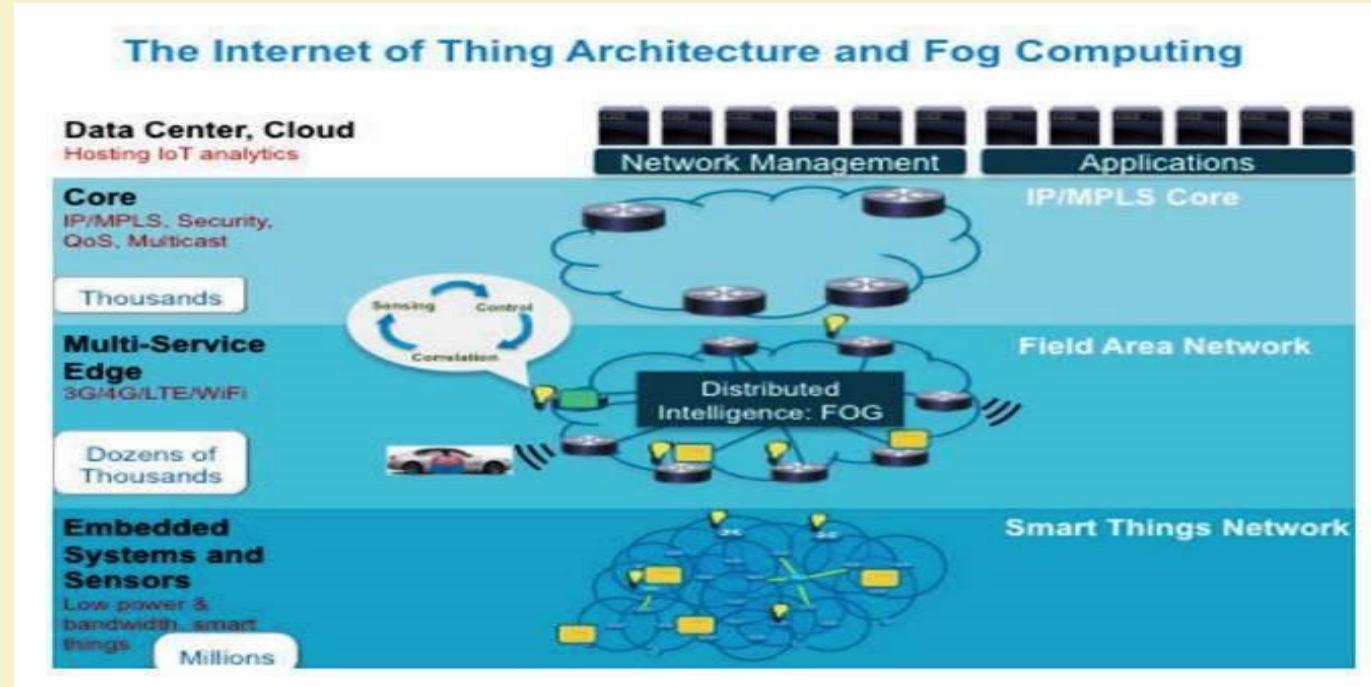


IIT KHARAGPUR



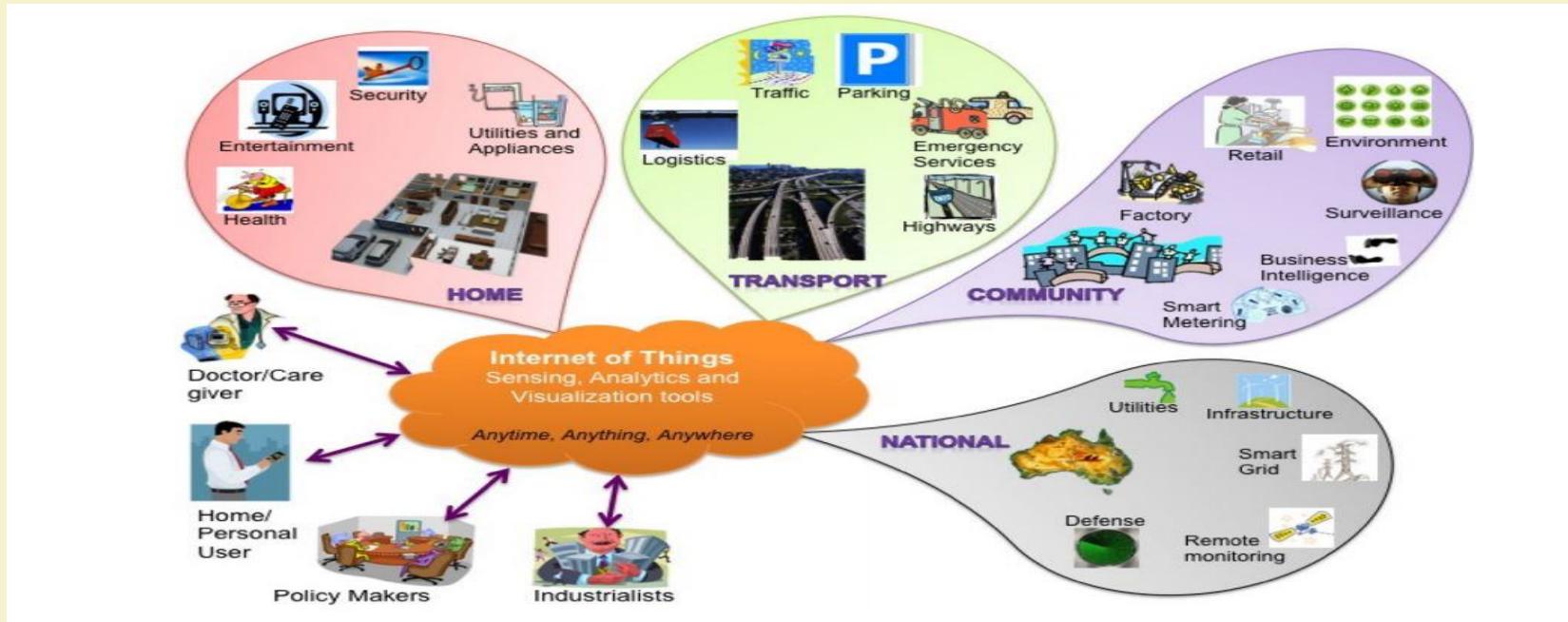
NPTEL  
NPTEL ONLINE  
CERTIFICATION COURSES

# 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



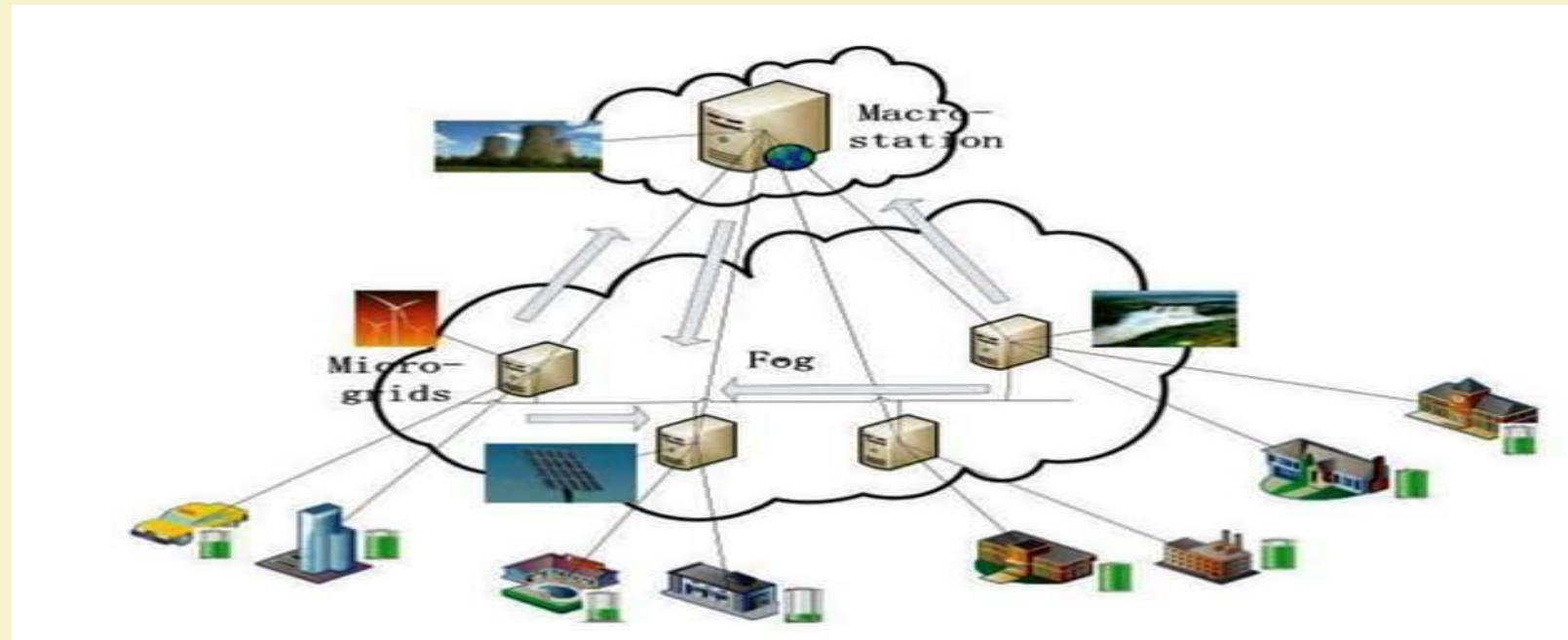
Source: Internet of Things (IoT): A vision, architectural elements, and future directions, Jayavardhana Gubbi, Rajkumar Buyya, Slaven Marusic, Marimuthu Palaniswami

# 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

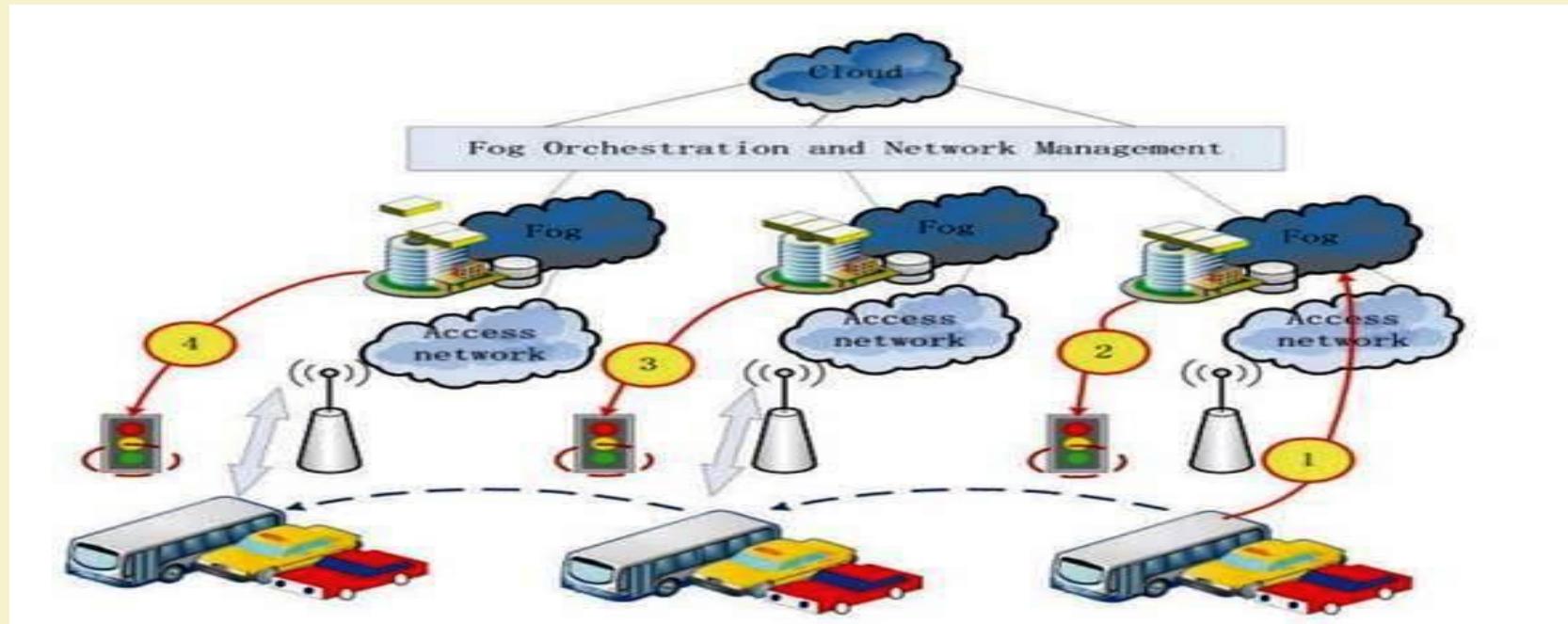


IIT KHARAGPUR



NPTEL ONLINE  
CERTIFICATION COURSES

# 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!!



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.

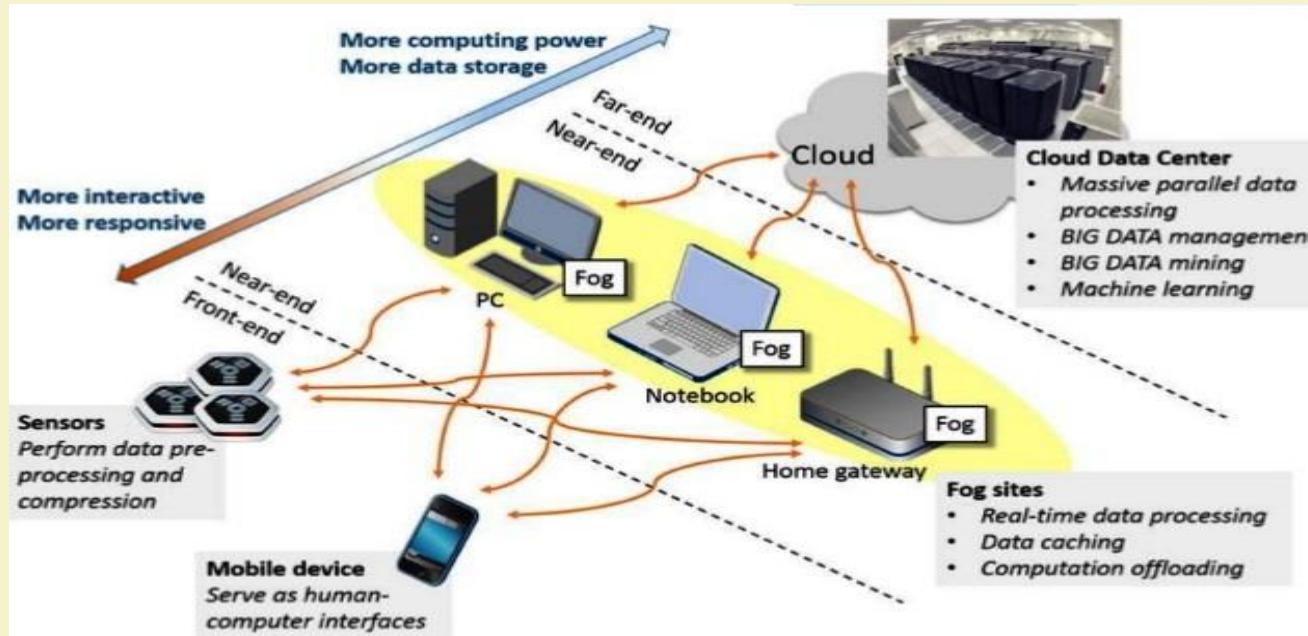


IIT KHARAGPUR



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



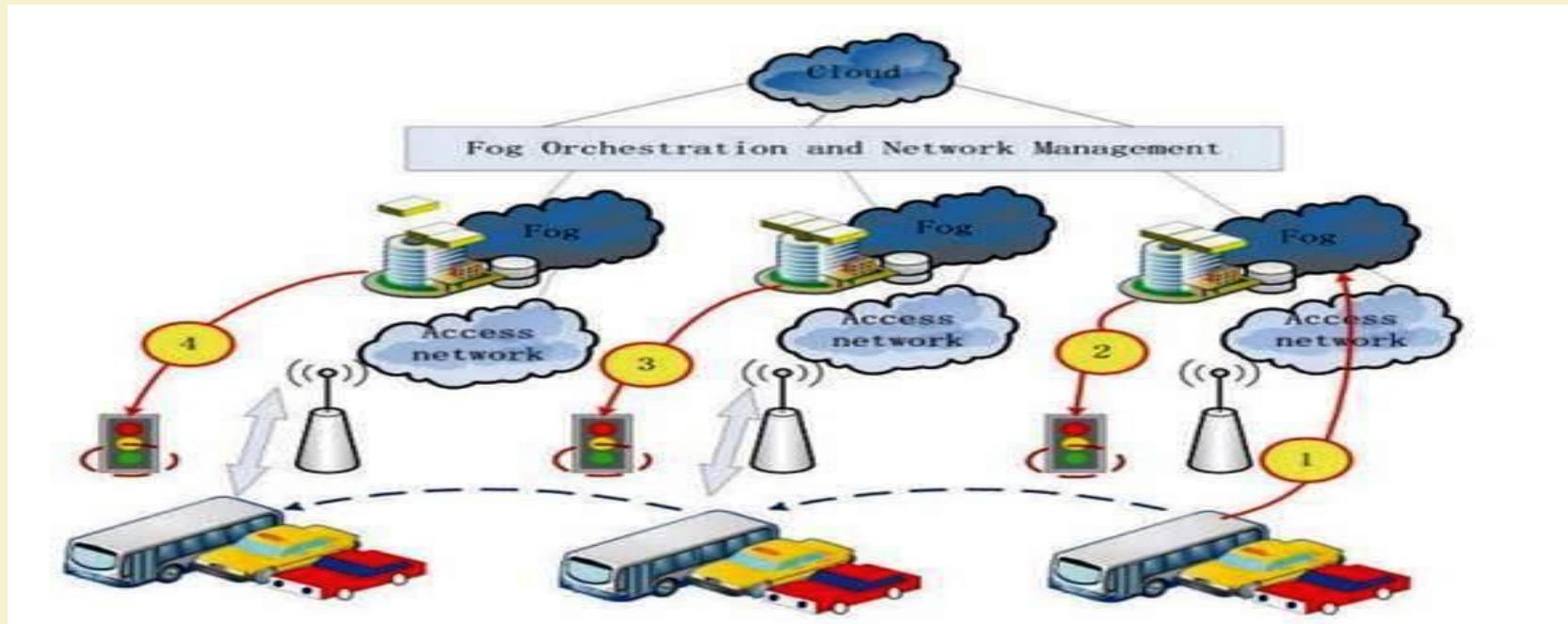
NPTEL  
NPTEL ONLINE  
CERTIFICATION COURSES

# 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



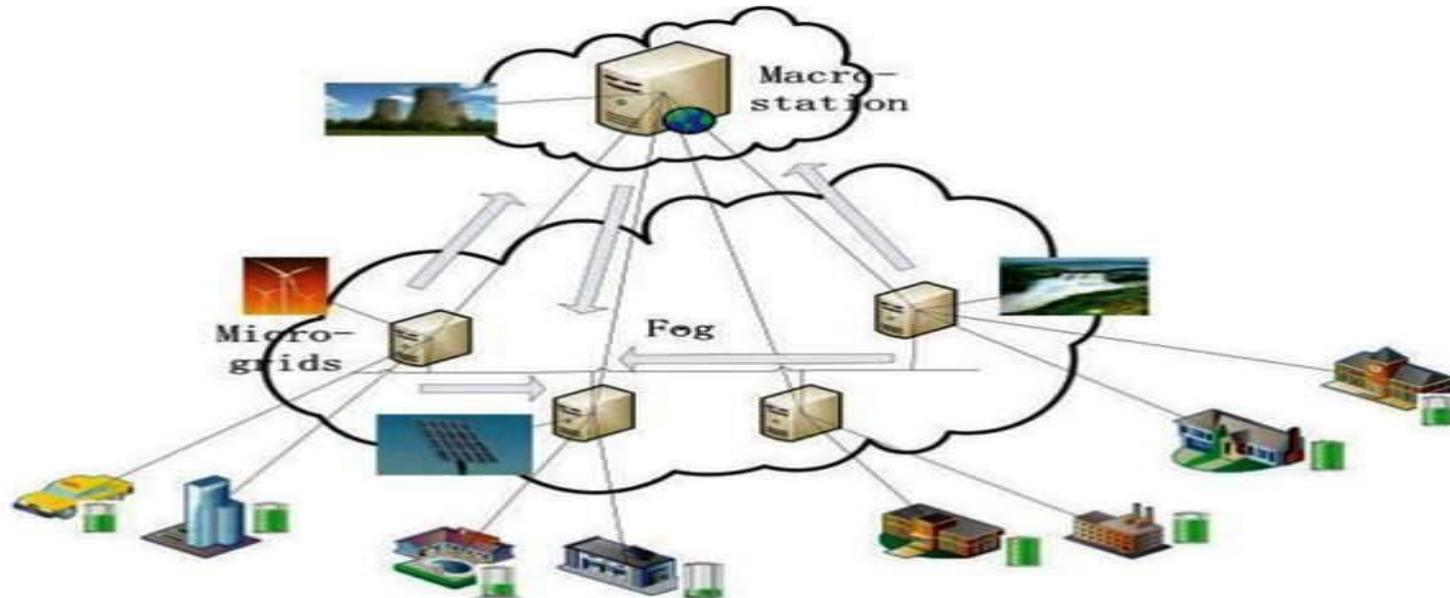
NPTEL  
ONLINE  
CERTIFICATION COURSES

# 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

# 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  
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!!



IIT KHARAGPUR



NPTEL  
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](mailto: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

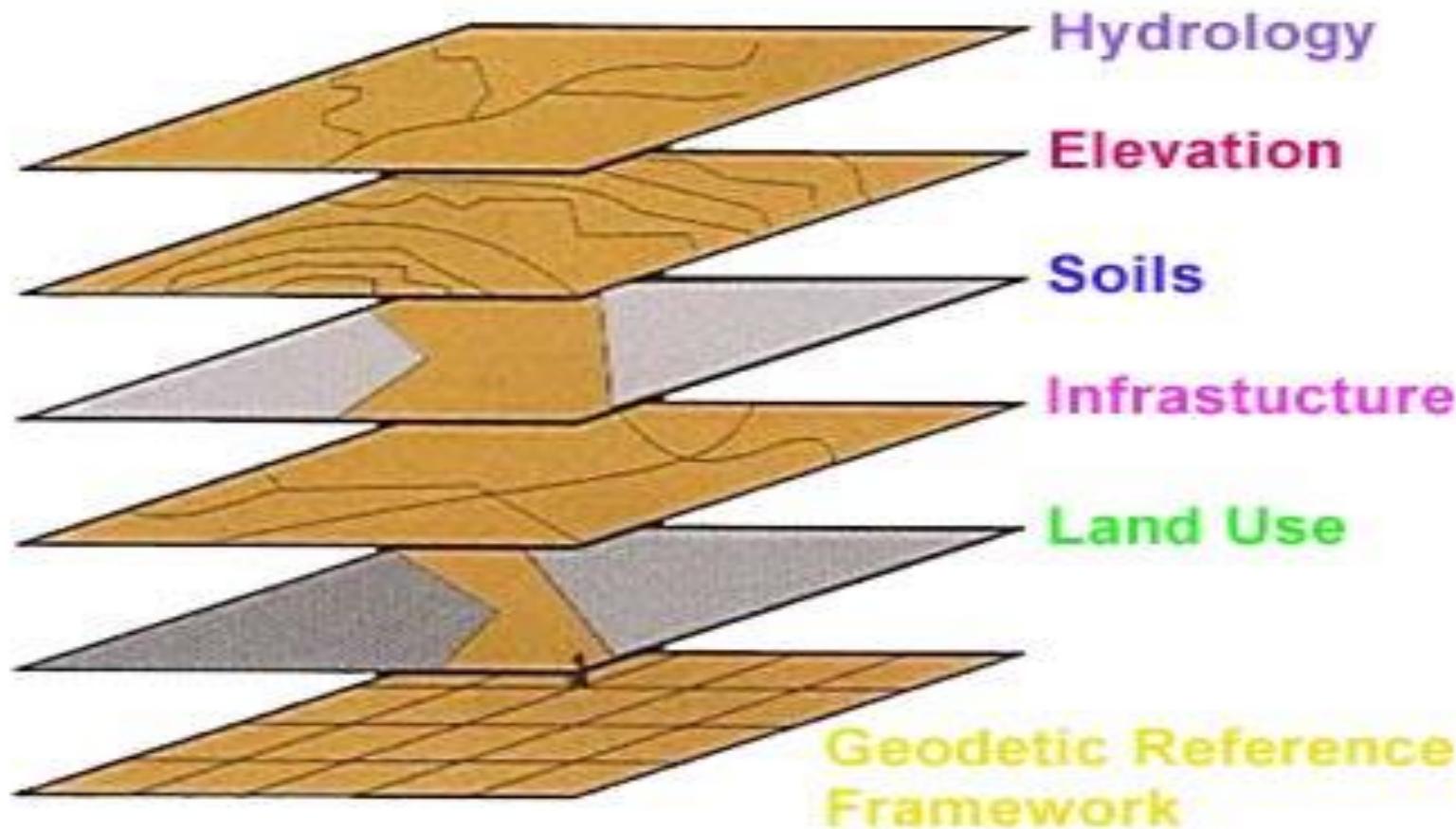


# Geospatial Information System - Challenges

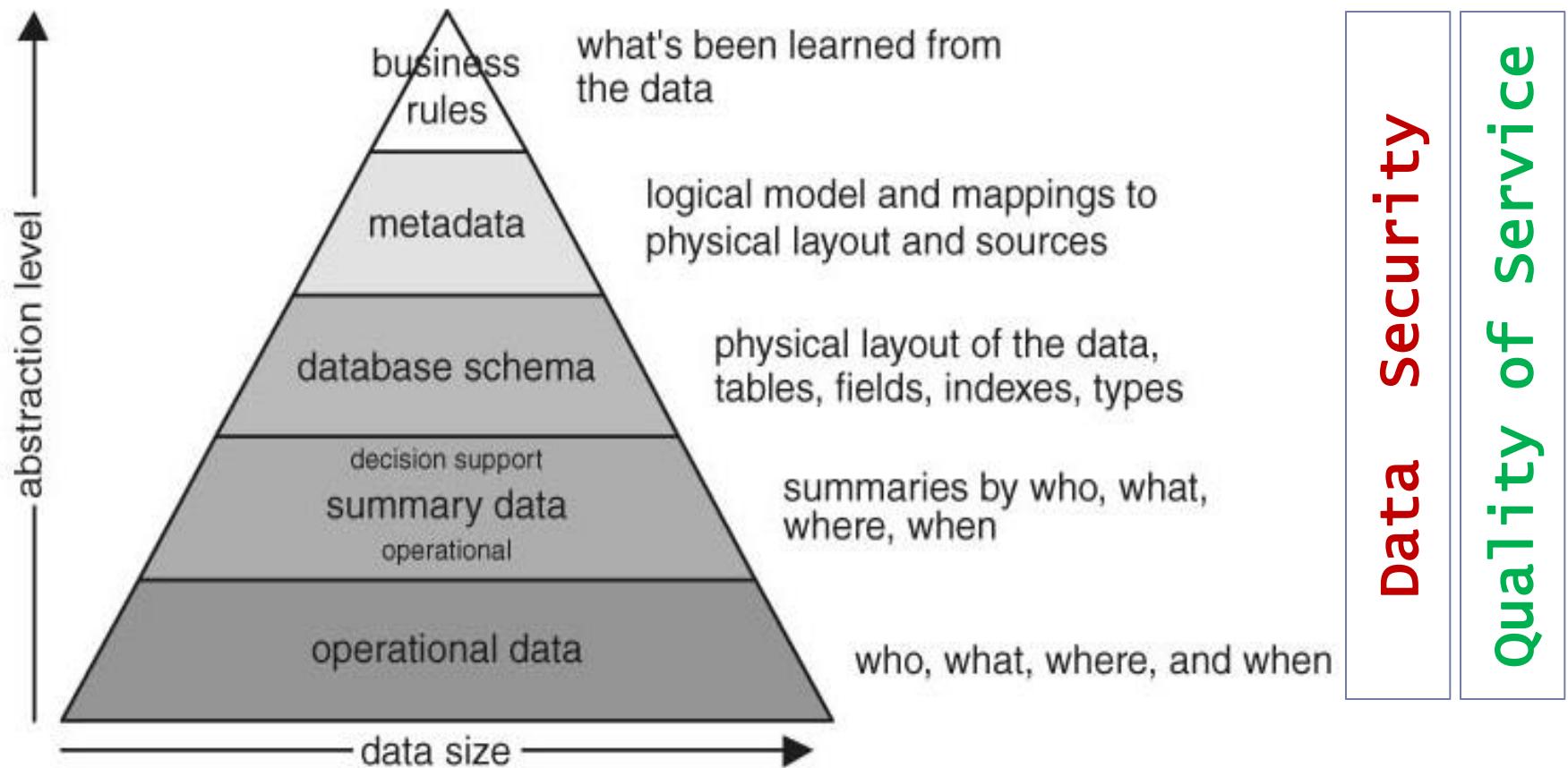
---

- ▶ 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

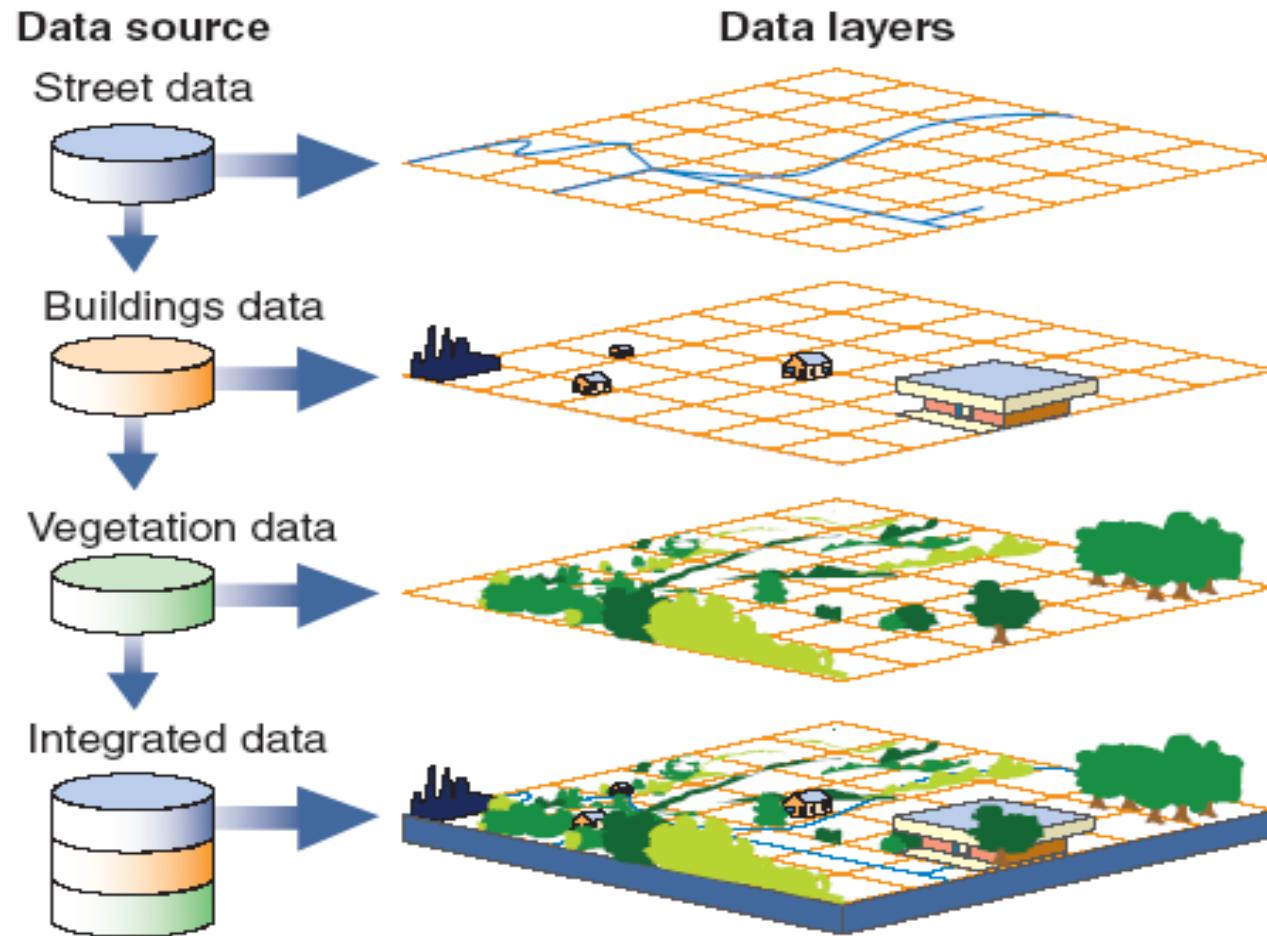


# Heterogeneity Issue

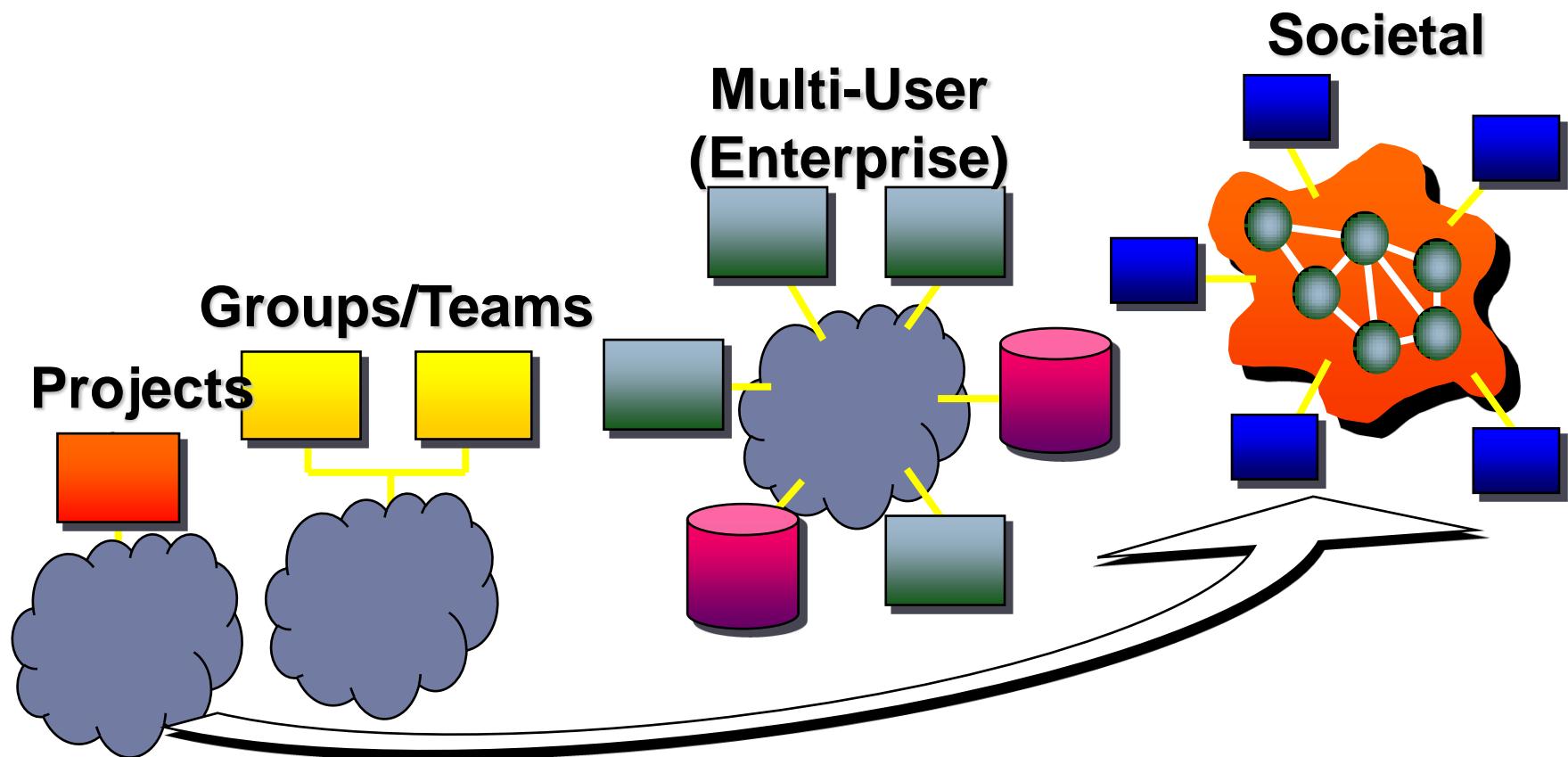
---

- ▶ **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



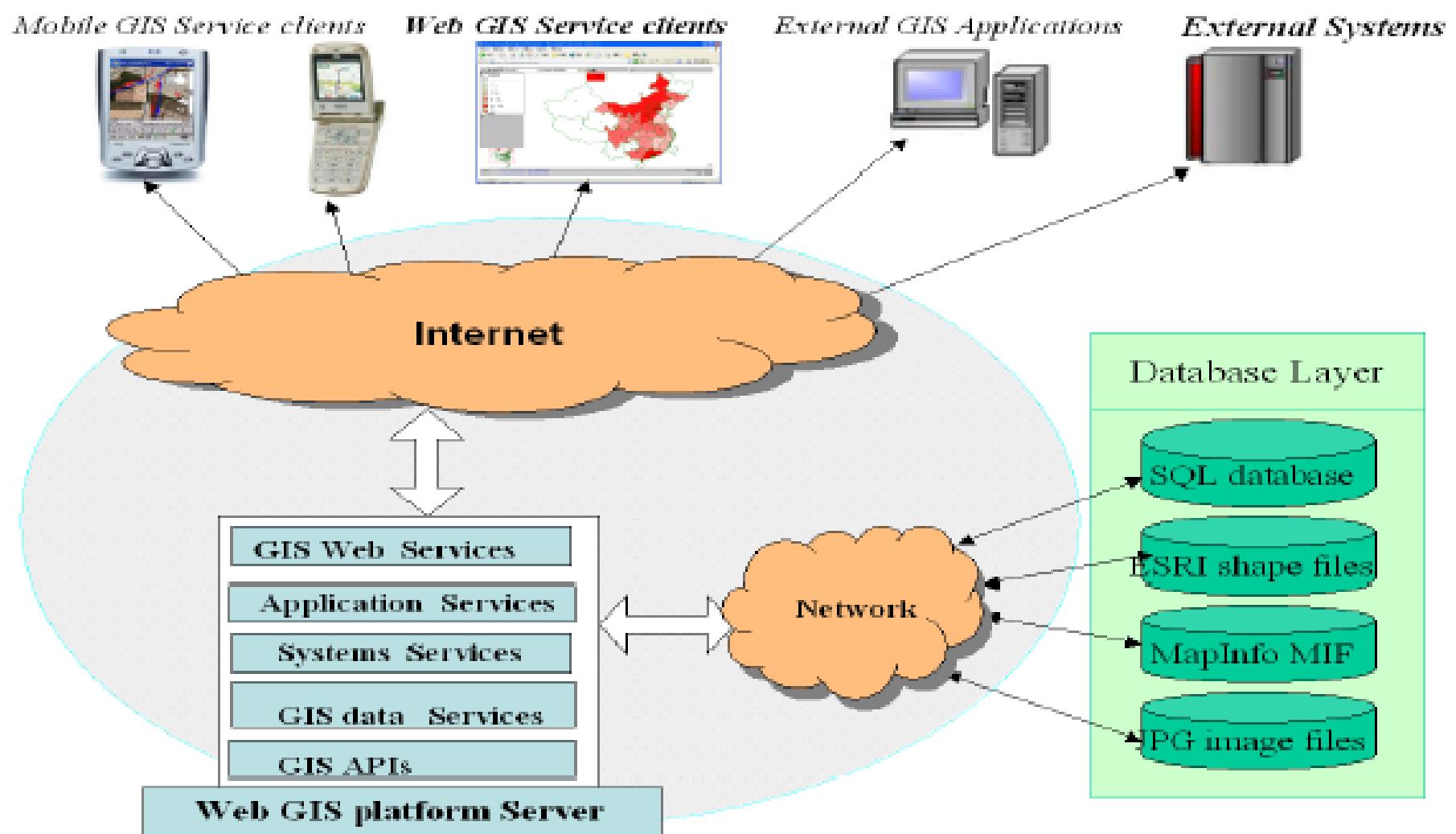
# Spatial Data Infrastructure (SDI)

---

- ▶ “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



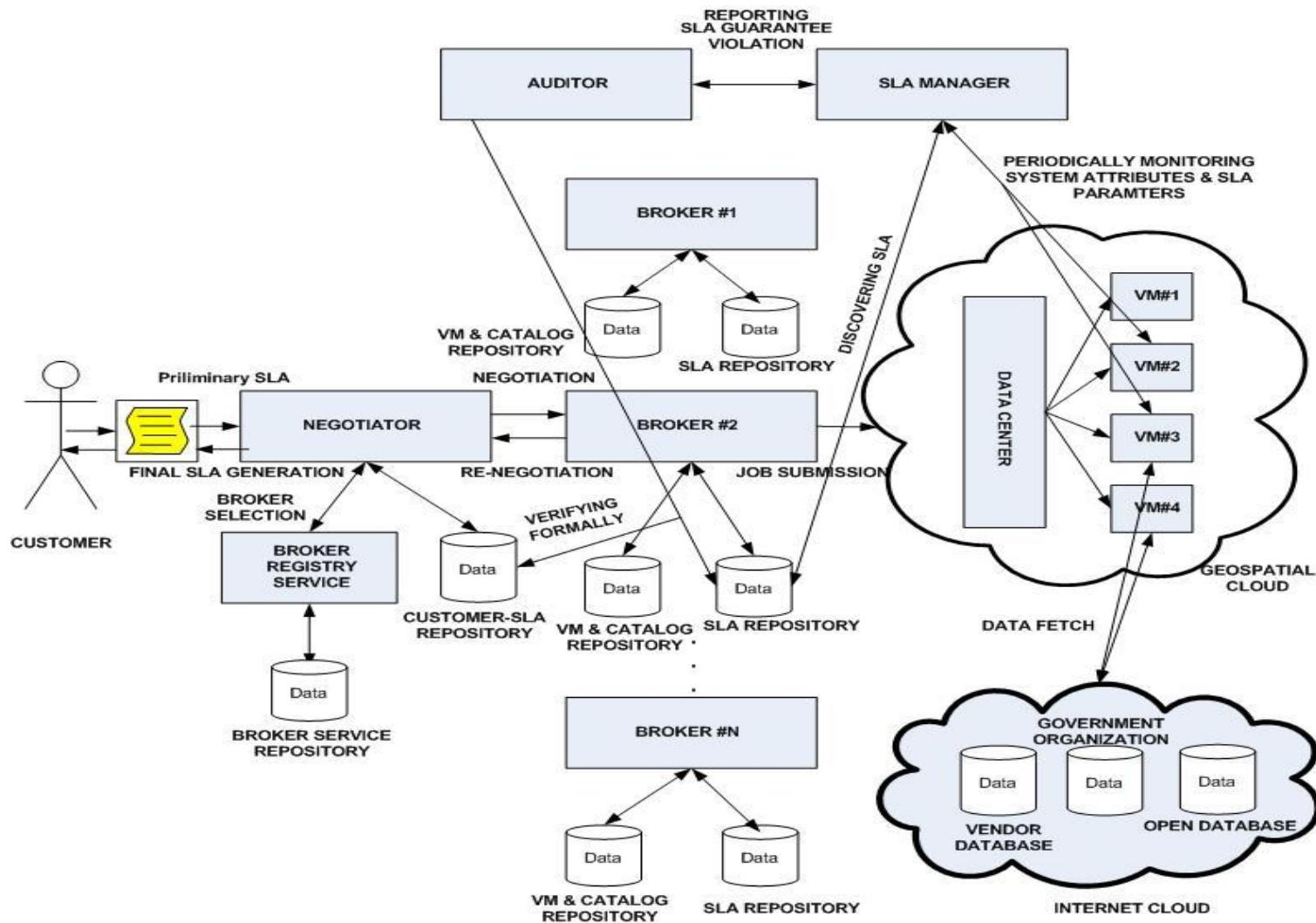
# **Cloud Actors**

---

- ▶ **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



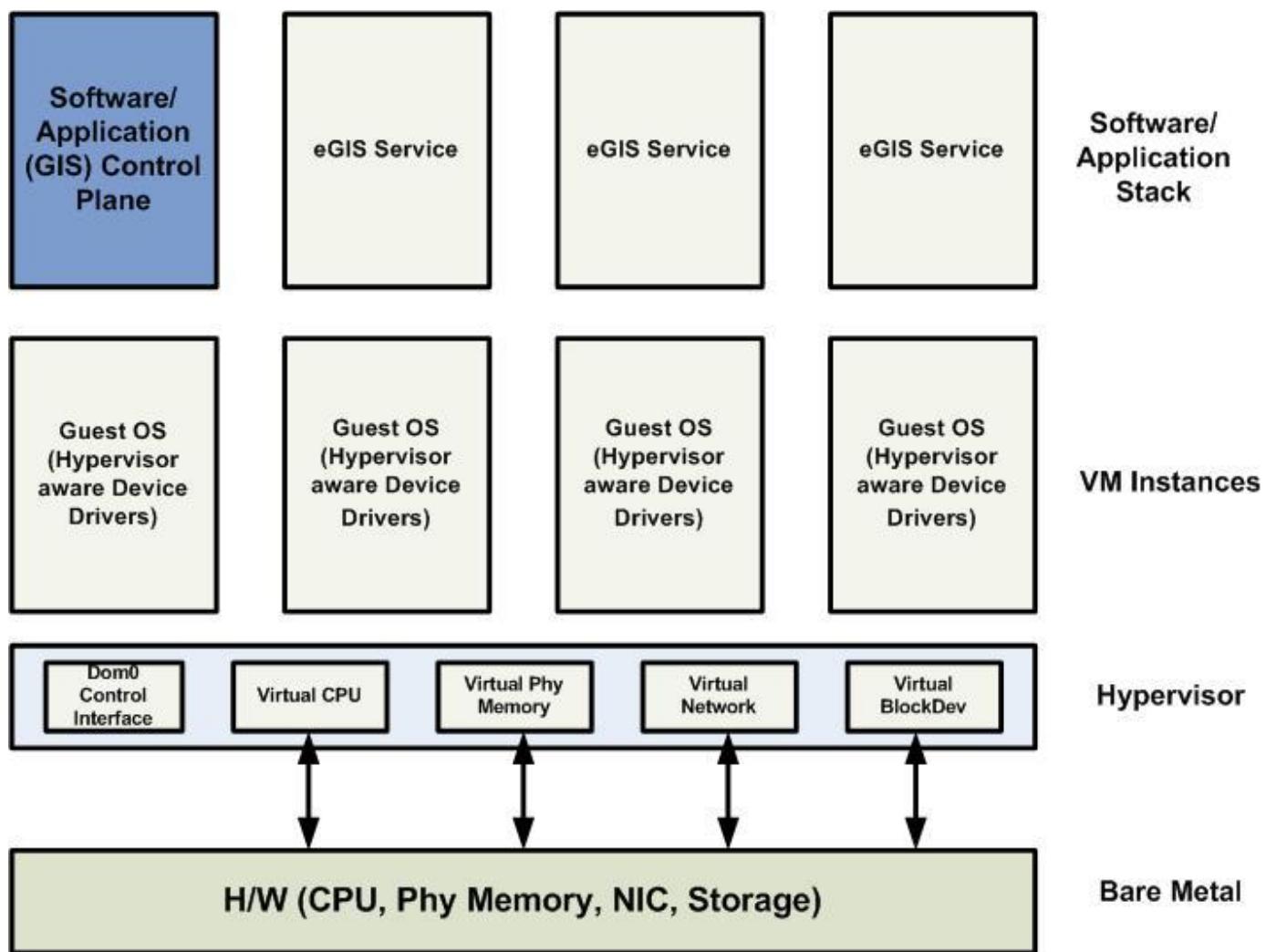
# Cloud as Service Provider

---

- 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

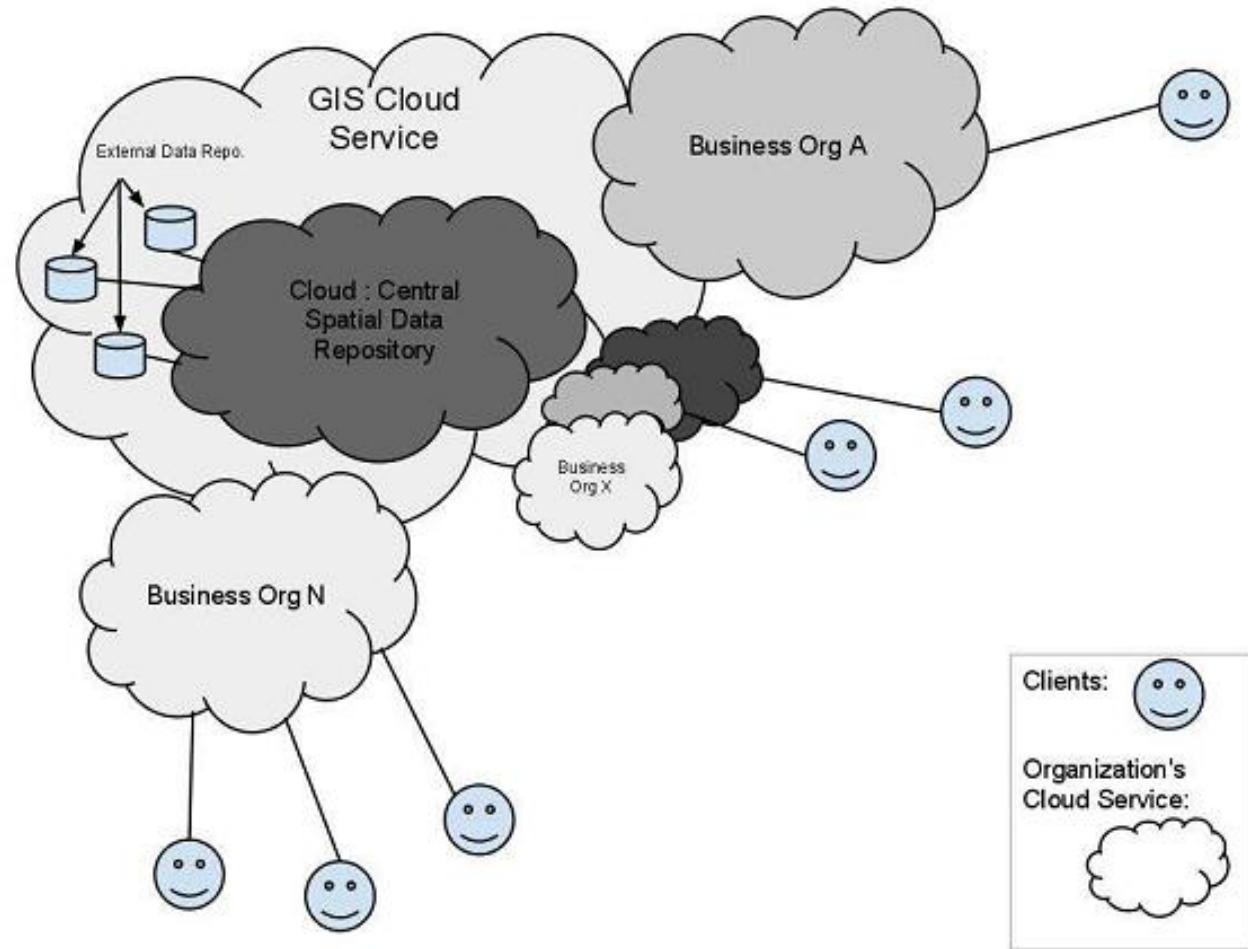


# **Geospatial Cloud Model**

---

- ▶ 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 Scenerio



# Geospatial Cloud

---

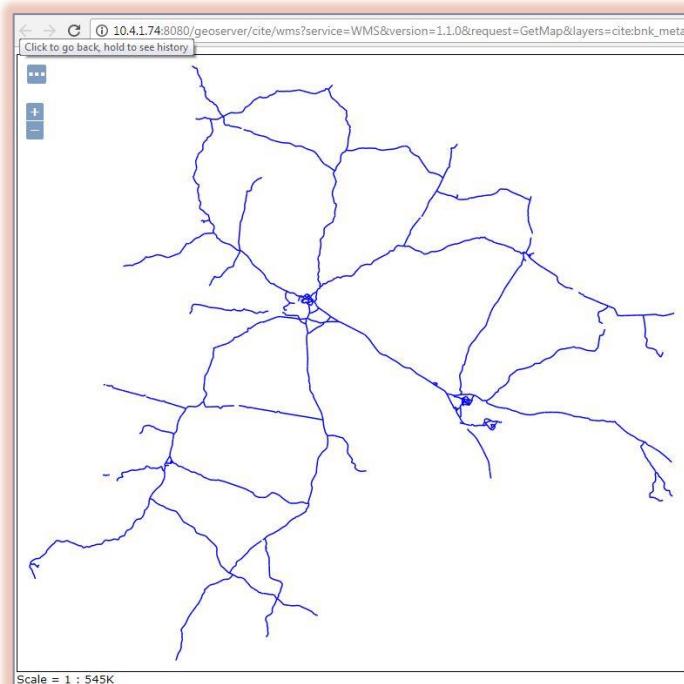
- ▶ Need to integrate data in an unified format.
- ▶ Performance Metrics: computation power, network bandwidth.
- ▶ Data sources:
  - Central Data Repository within the cloud.
  - External Data Repository providing data as WFS,WMS services.

---

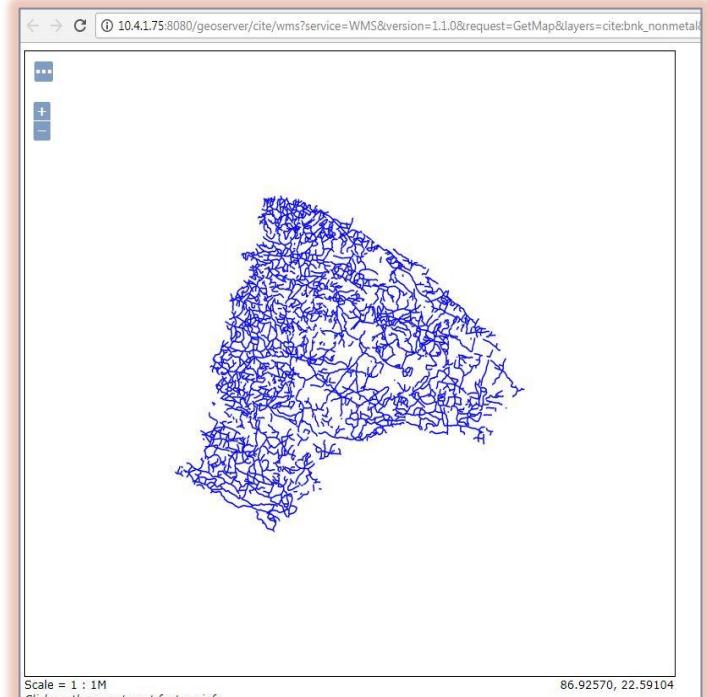
# Experimental GeoSpatial-Cloud @IITKgp



# Service Integration for Query in Cloud (Case Study 1)



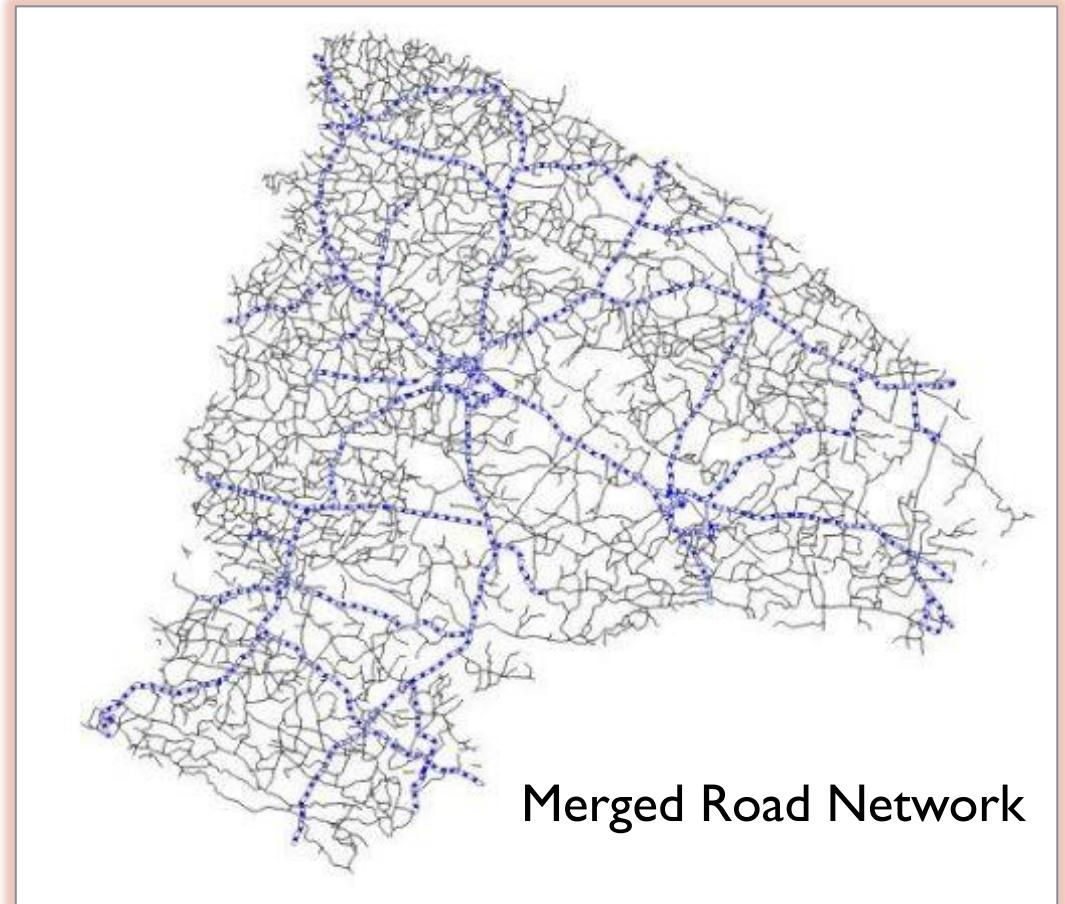
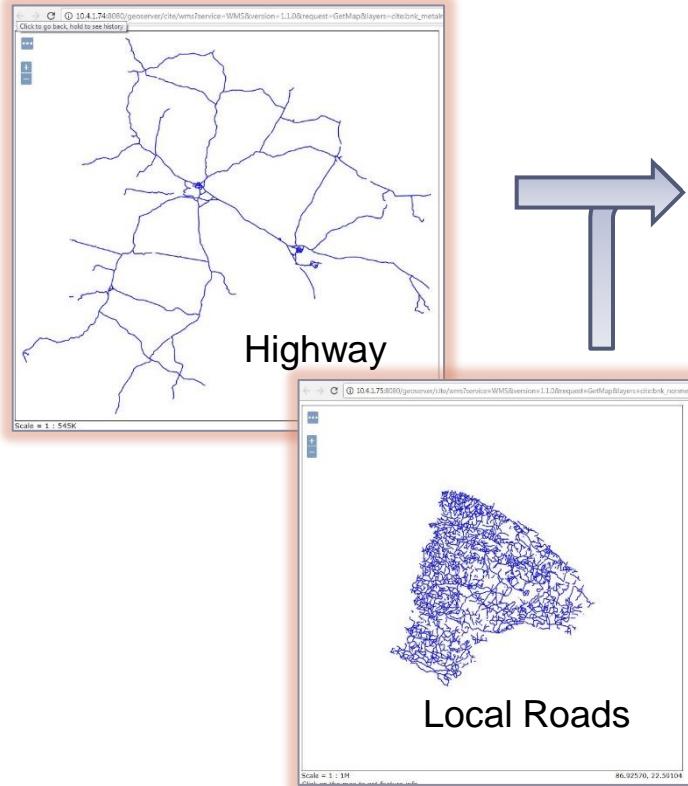
Highway



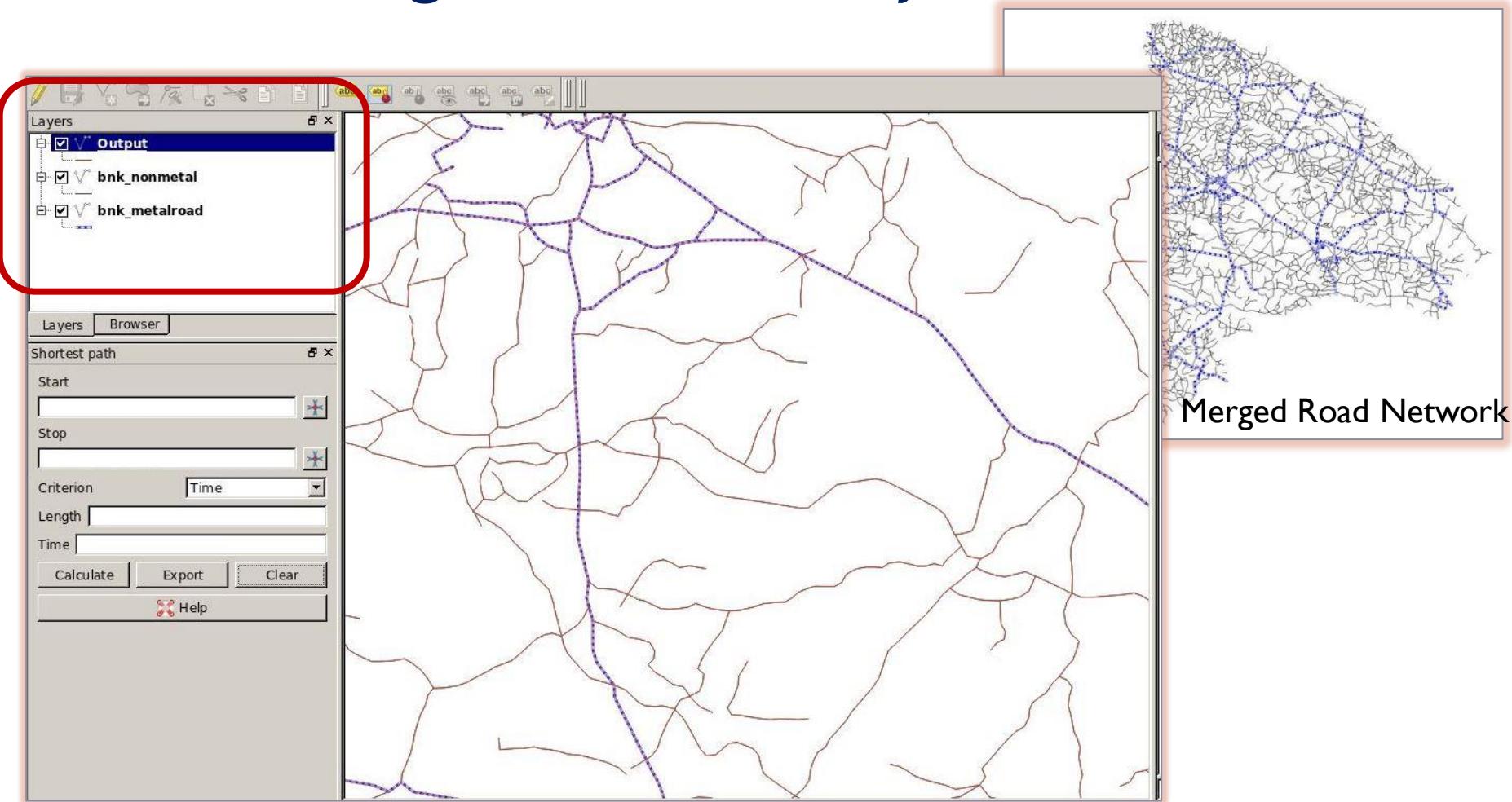
Local Roads



# Service Integration for Query in Cloud

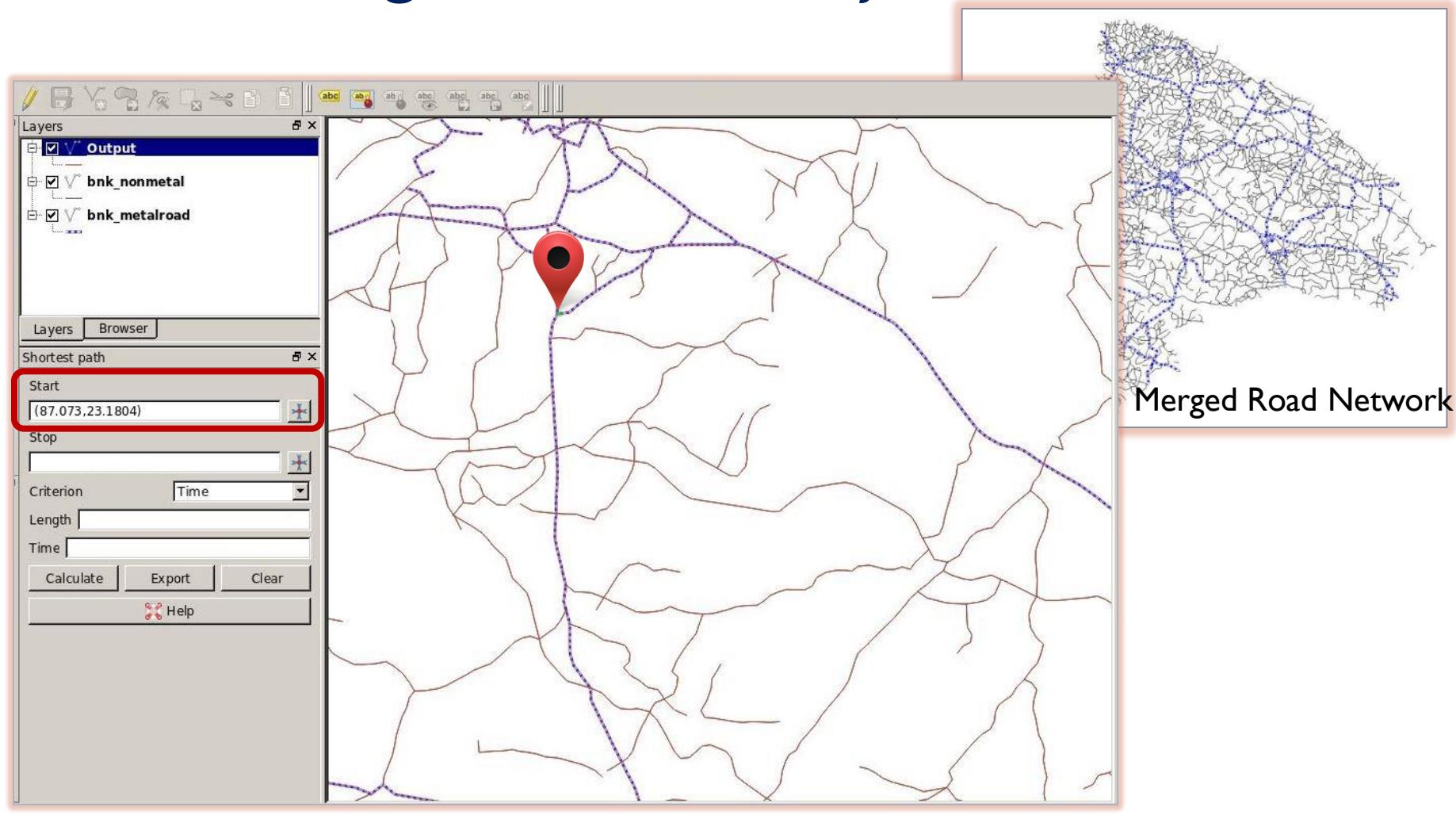


# Service Integration for Query in Cloud



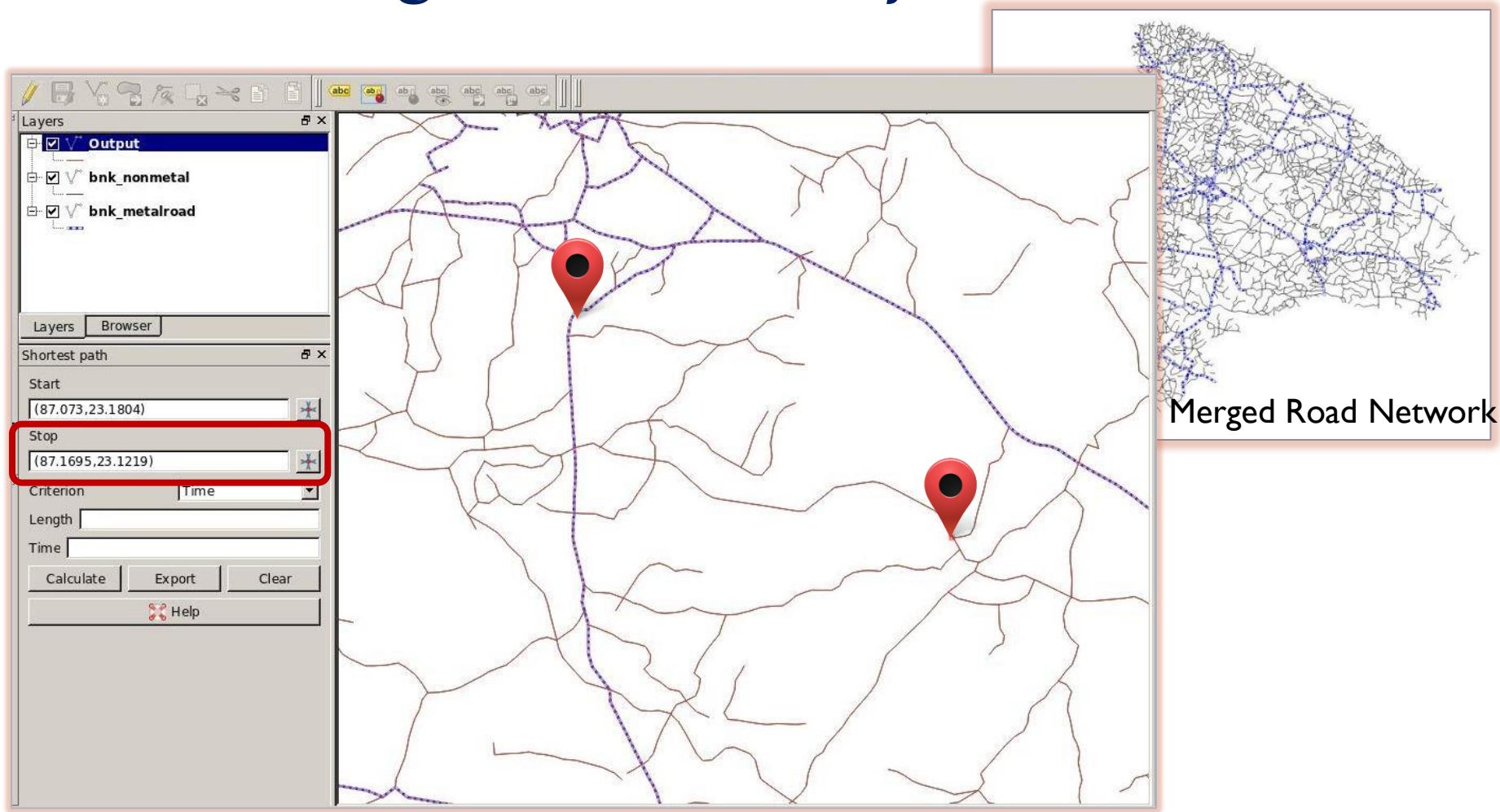
► **Shortest Path Calculation**  
CSE, IIT Kharagpur

# Service Integration for Query in Cloud



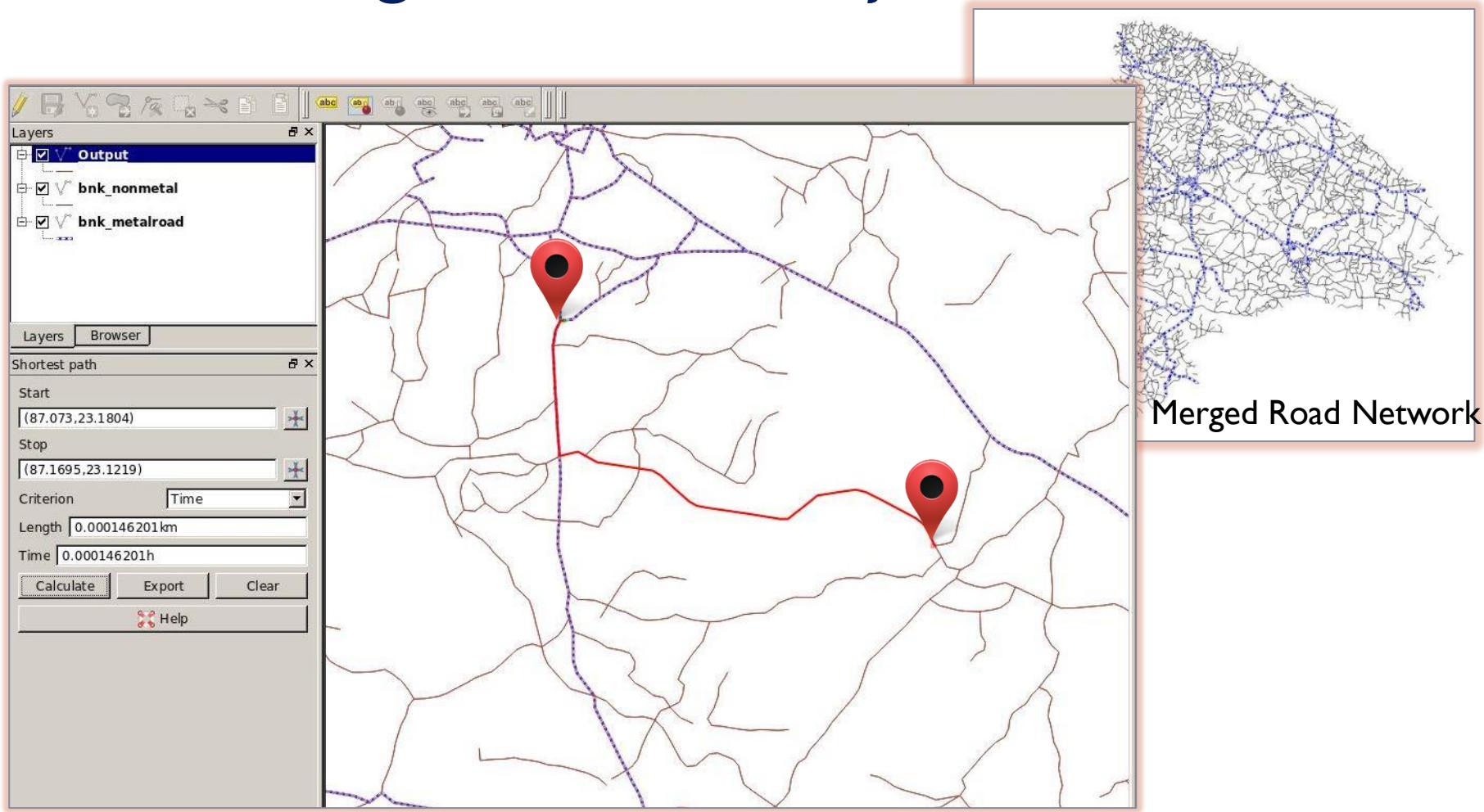
Shortest Path Calculation  
CSE, IIT Kharagpur

# Service Integration for Query in Cloud



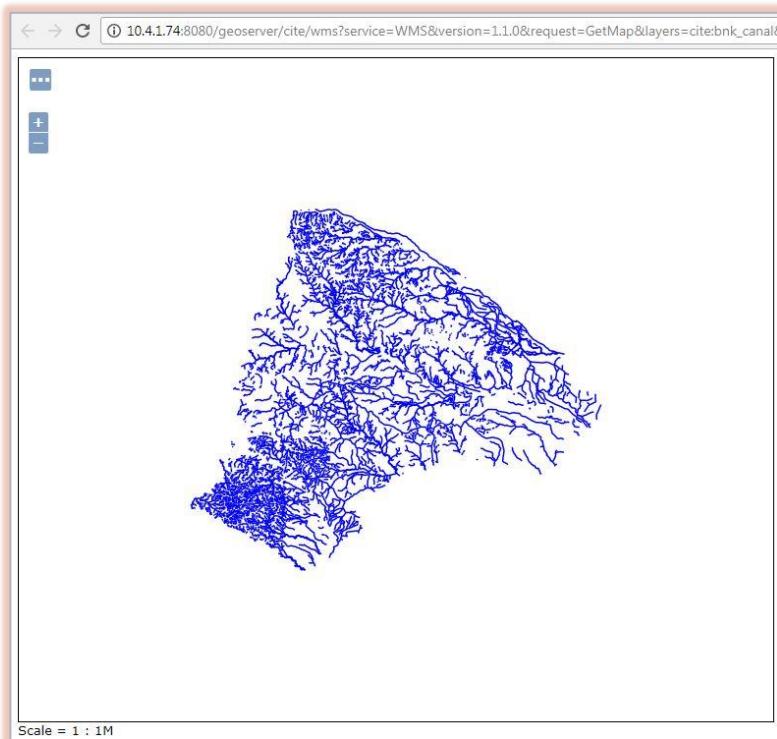
► **Shortest Path Calculation**  
CSE, IIT Kharagpur

# Service Integration for Query in Cloud

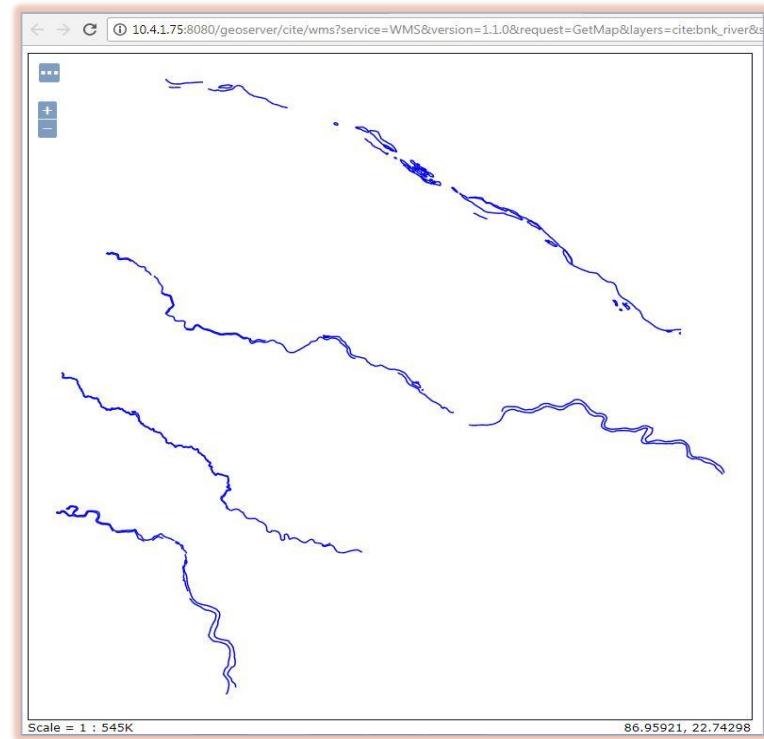


Shortest Path Calculation  
CSE, IIT Kharagpur

# Service Integration for Query in Cloud (Case Study 2)



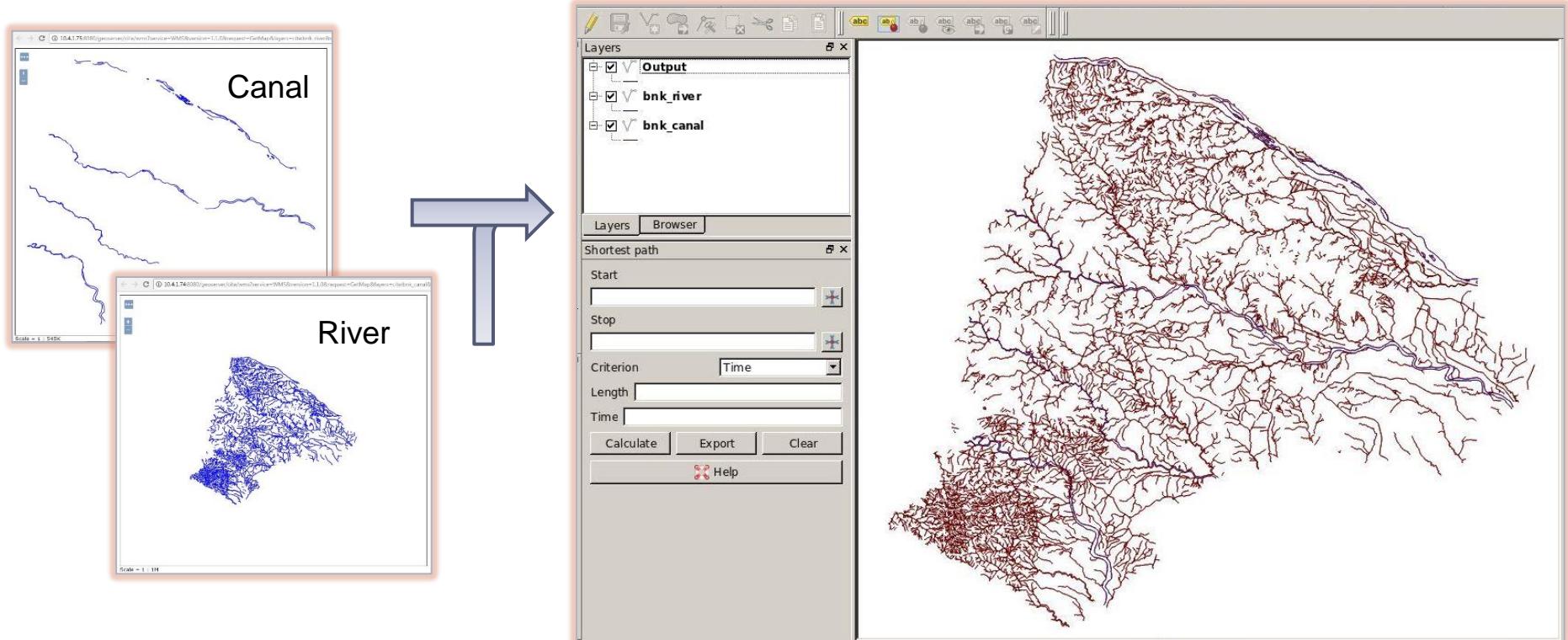
Canal



River



# Service Integration for Query in Cloud

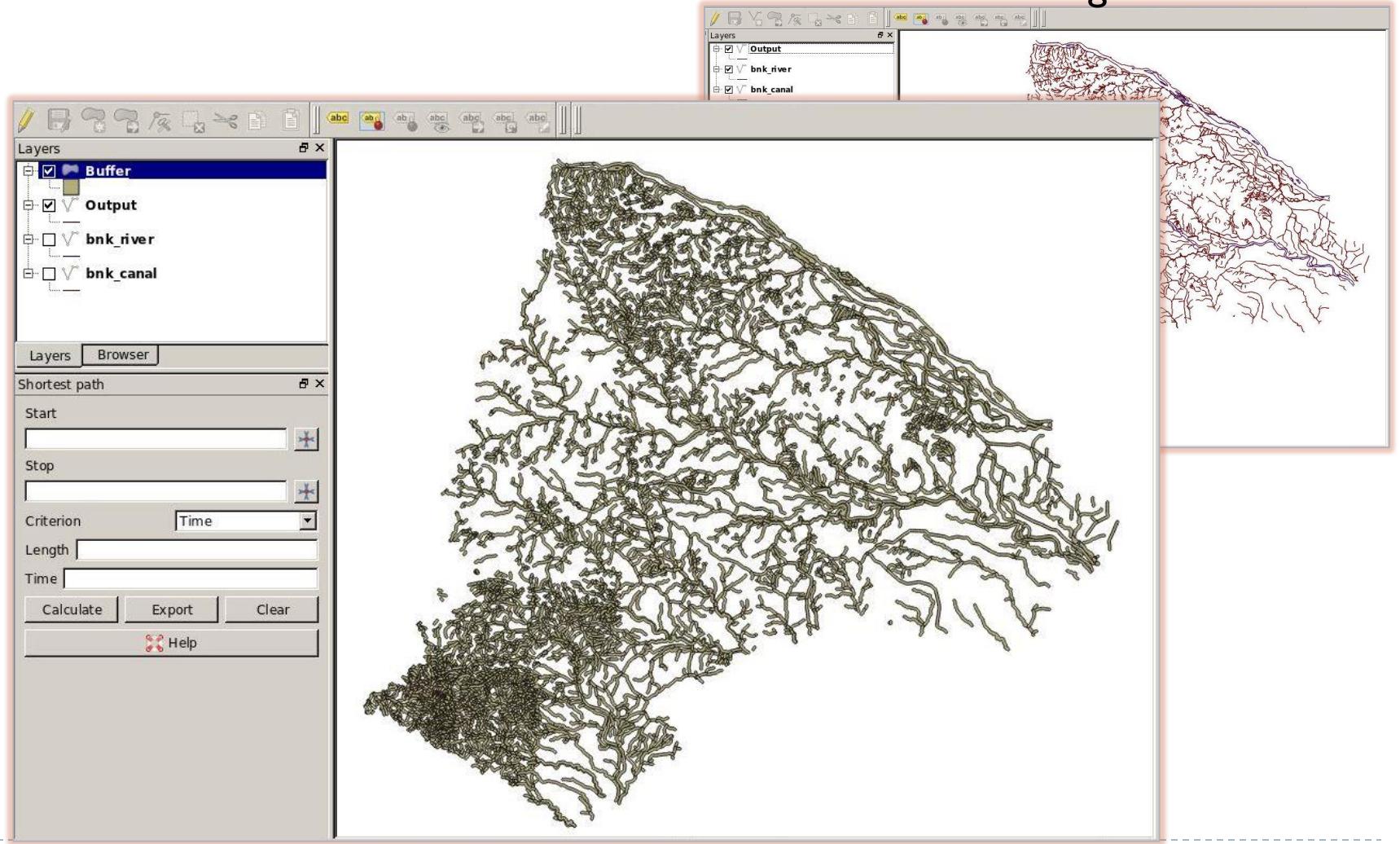


Merged Water Network



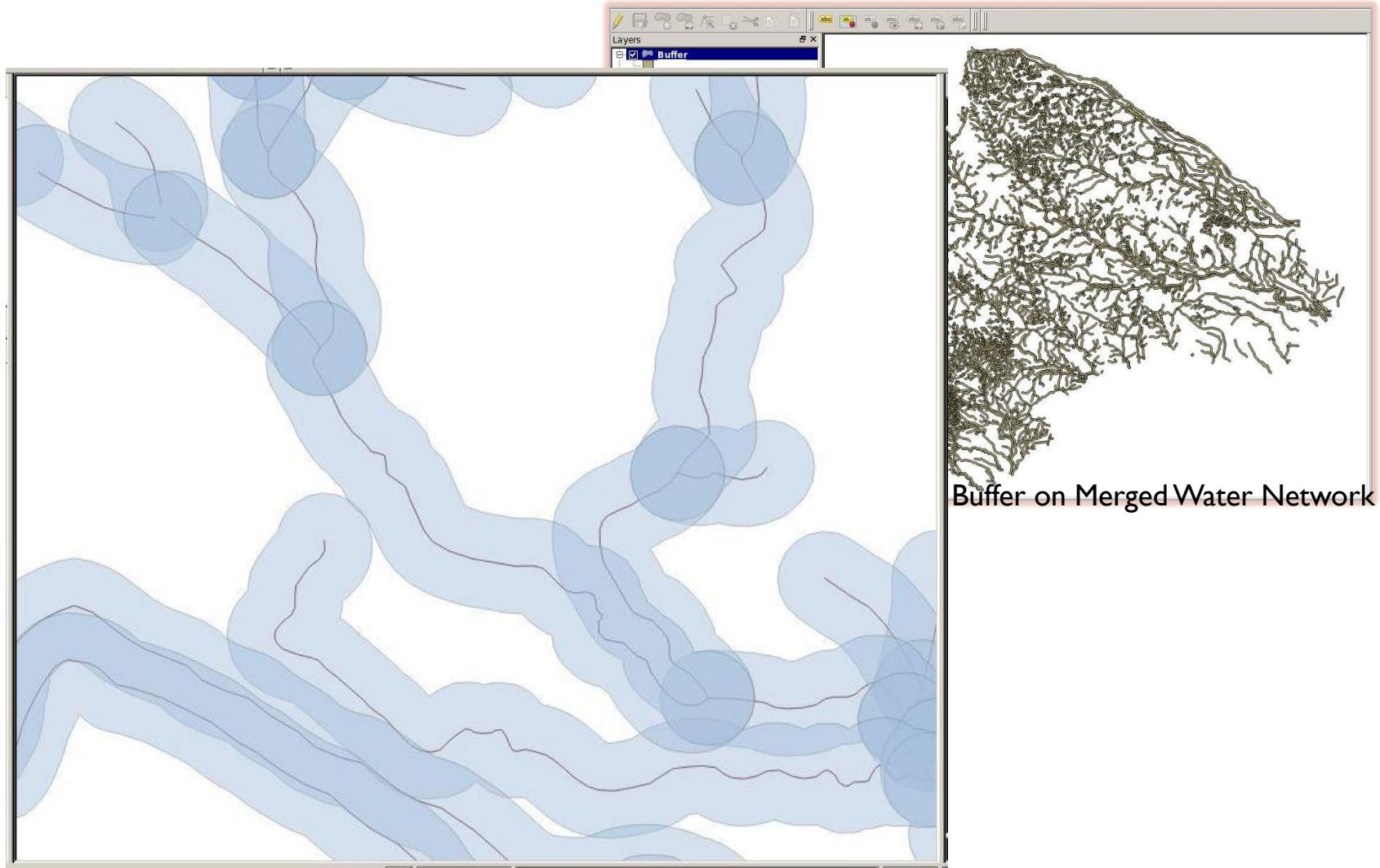
# Service Integration for Query in Cloud

Merged Water Network



▶ Buffer on Merged Water Network

# Service Integration for Query in Cloud



▶ Buffer on Merged Water Network (Zoomed)

---

# Challenges in Geospatial Cloud



# 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

# 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 !





IIT KHARAGPUR



NPTEL ONLINE  
CERTIFICATION COURSES

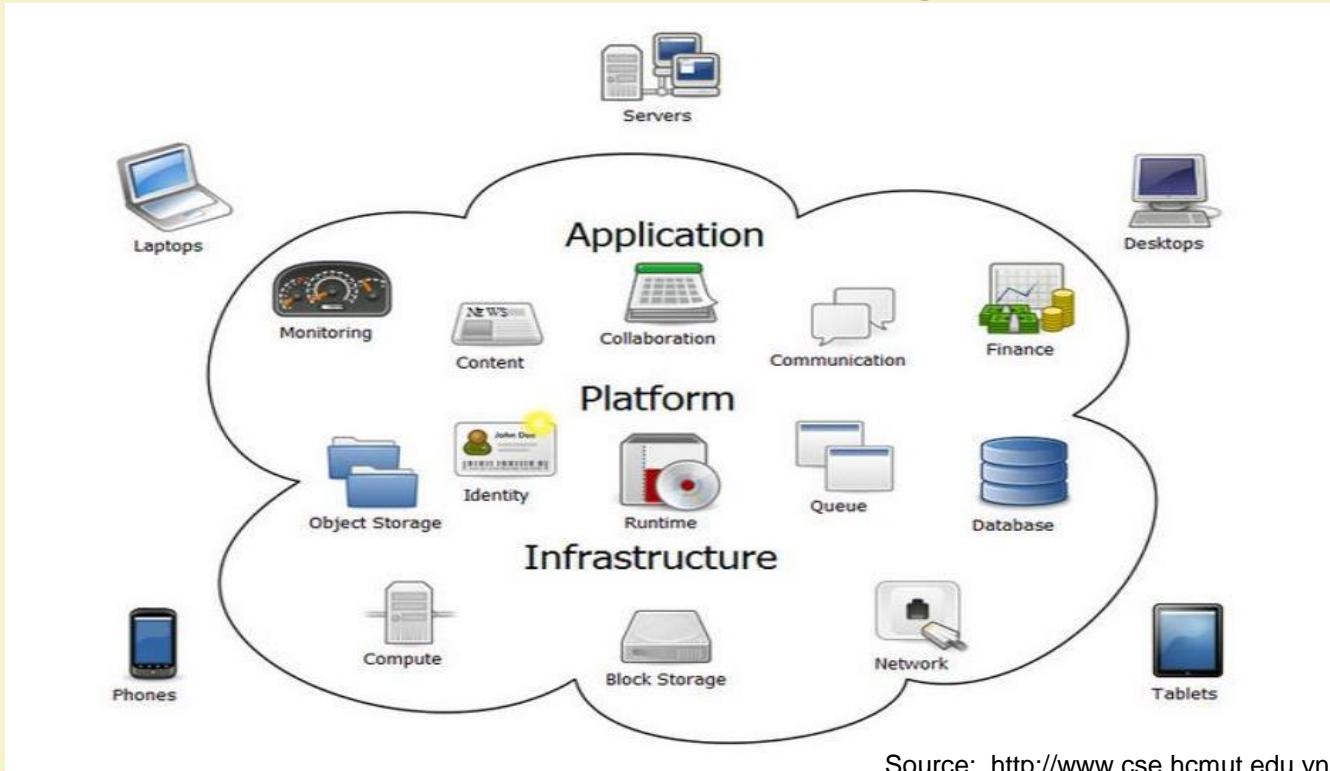
# CLOUD COMPUTING

## Resource Management - I

PROF. SOUMYA K. GHOSH

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING  
IIT KHARAGPUR

# Different Resources in Computing



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  
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 CO<sub>2</sub> produced by burning fossil fuels.
- Need to reduce the energy used with minimal performance impact.

*Ref: Efficient Resource Management for Cloud Computing Environments, by Andrew J. Younge, Gregor von Laszewski, Lizhe Wang, Sonia Lopez-Alarcon, Warren Carithers,*



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 CO<sub>2</sub> 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.

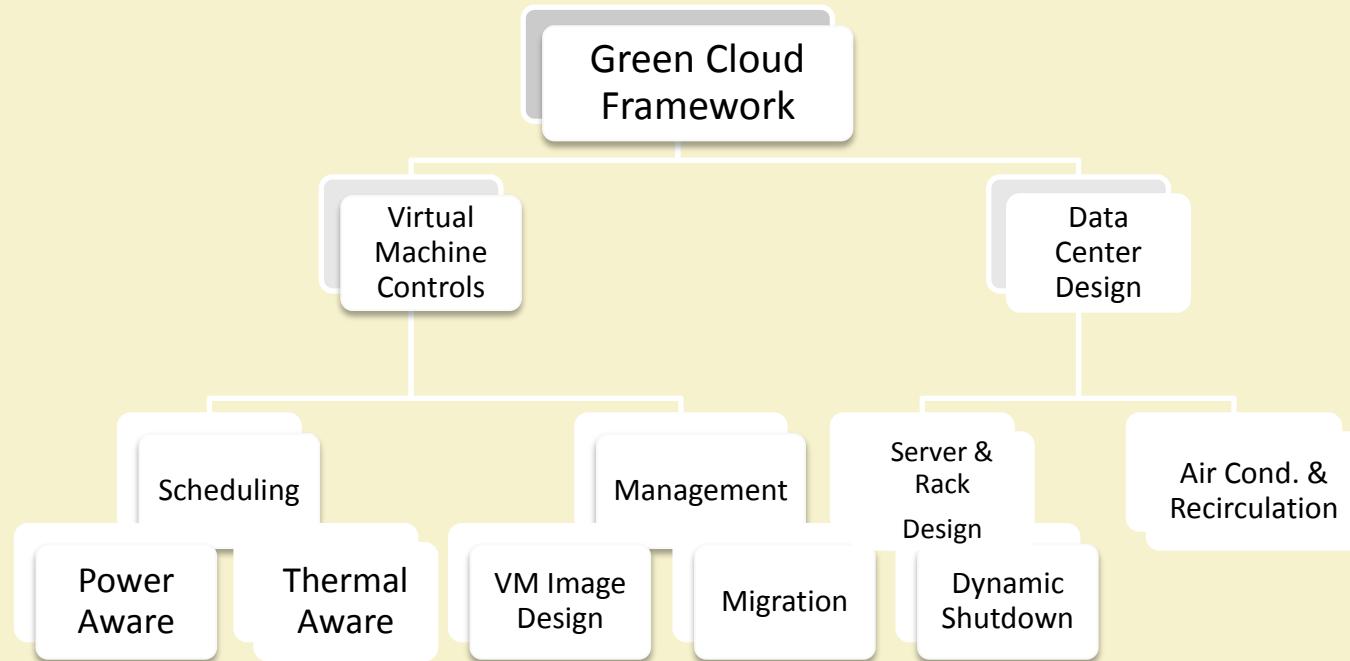


IIT KHARAGPUR



NPTEL ONLINE  
CERTIFICATION COURSES

# Steps towards Energy Efficiency



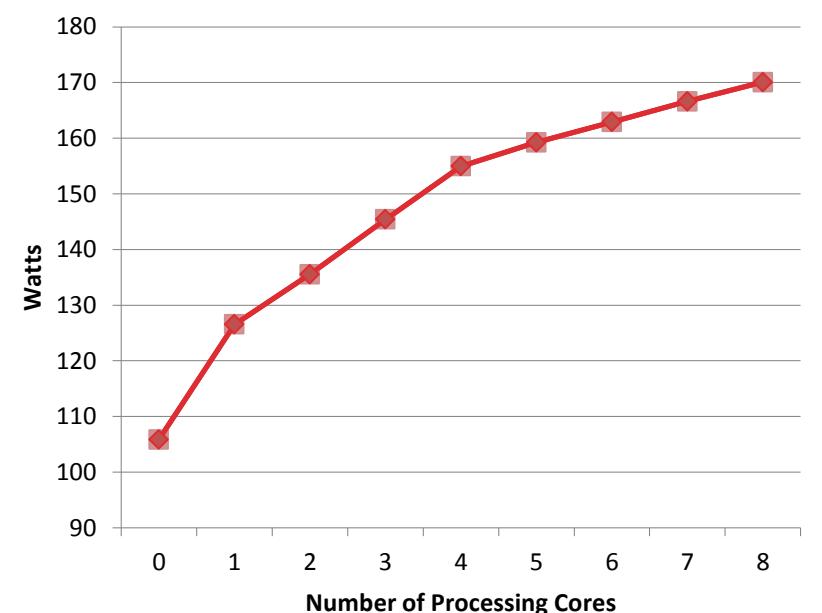
IIT KHARAGPUR



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
```

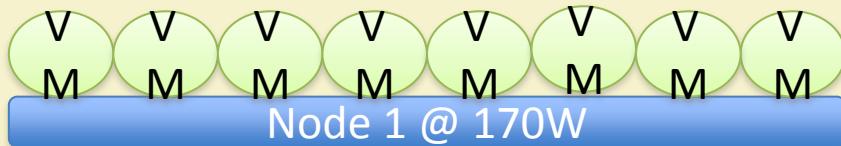


IIT KHARAGPUR



NPTEL  
NPTEL ONLINE  
CERTIFICATION COURSES

# 485 Watts vs. 552 Watts !



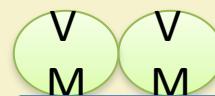
Node 2 @ 105W

Node 3 @ 105W

Node 4 @ 105W

---

VS.



Node 1 @ 138W

Node 2 @ 138W



Node 3 @ 138W

Node 4 @ 138W

# 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.

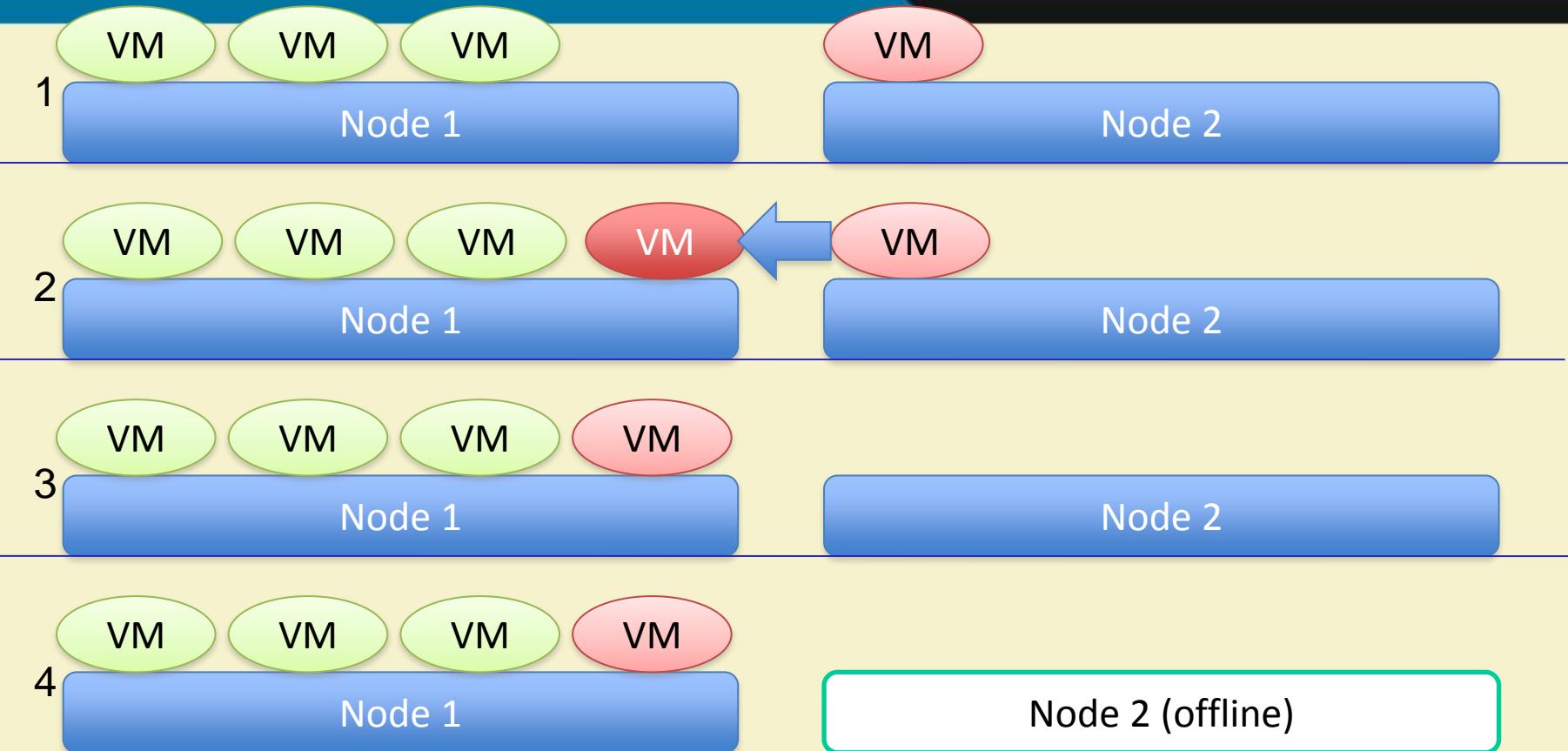
Management



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

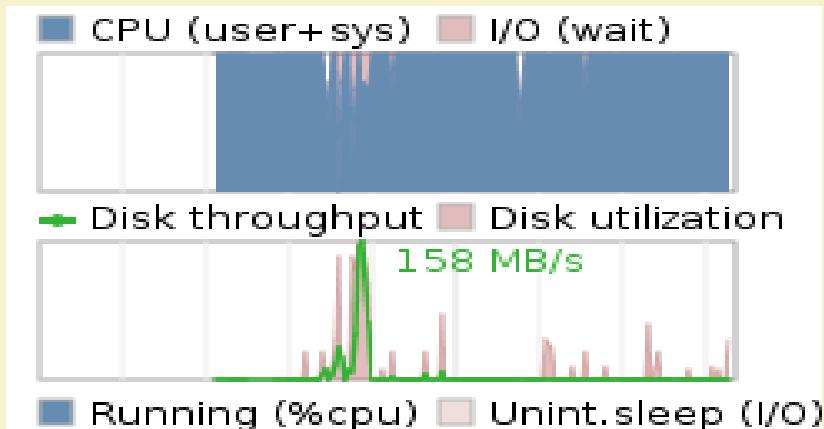


NPTEL  
NPTEL ONLINE  
CERTIFICATION COURSES

# 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 ONLINE  
CERTIFICATION COURSES

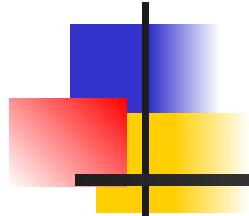
# Thank you!



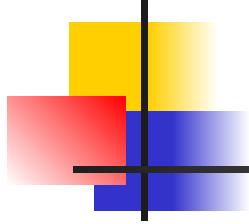
IIT KHARAGPUR



NPTEL ONLINE  
CERTIFICATION COURSES

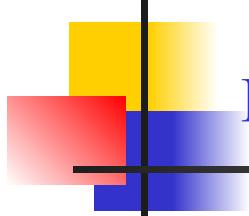


# **REST - Representational State Transfer**



# What is REST ?

REST is a term coined by Roy Fielding to describe an **architecture style** of networked systems. REST is an acronym standing for Representational State Transfer.



## Rest – An architectural Style

### Elements

- Components – Proxy , gateway etc
- Connectors – client , server etc
- Data – resource , representation etc

### REST

- Ignores component implementation details.
- Focus on roles of components,their interactions and their interpretation of data elements.

- **Resource**
- **URI-Uniform Resource Identifier (or URL)**
- **Web Page (HTML Page)**

URI

`http://weather.example.com/oaxaca`

Representation

Metadata:  
Content-type:  
`application/xhtml+xml`

Data:

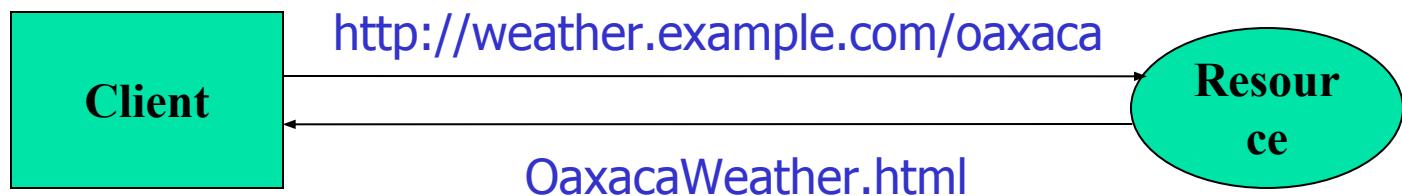
```
<!DOCTYPE html PUBLIC "...
  "http://www.w3.org/...
<html xmlns="http://www...
<head>
<title>5 Day Forecast for
Oaxaca</title>
...
</html>
```

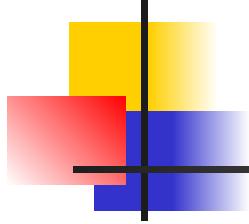


Resource  
*Oaxaca Weather Report*



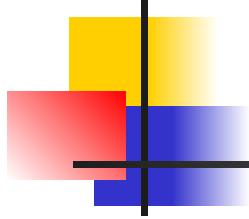
## Why is it called Representational State Transfer ?





**"Representational State Transfer is intended to evoke an image of how a well-designed Web application behaves: a network of web pages (a virtual state-machine), where the user progresses through an application by selecting links (state transitions), resulting in the next page (representing the next state of the application) being transferred to the user and rendered for their use."**

**Roy Fielding.**

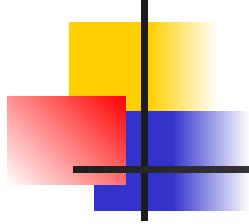


## **REST - An Architectural Style of Networked System**

- Underlying Architectural model of the world wide web.
- Guiding framework for Web protocol standards.

### **REST based web services**

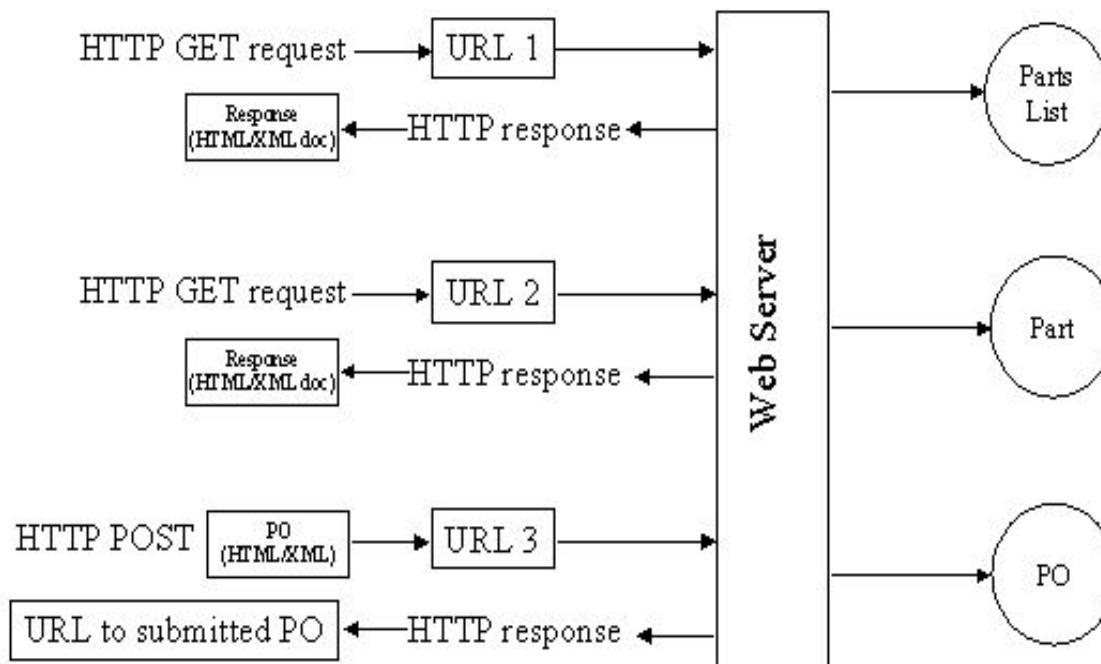
- Online shopping
- Search services
- Dictionary services

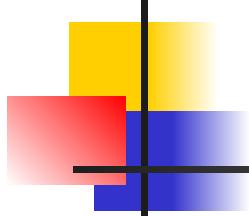


## Parts Depot Web Services

- Parts Depot, Inc has deployed some web services to enable its customers to:
  - get a list of parts
  - get detailed information about a particular part
  - submit a Purchase Order (PO)

## REST way of Implementing the web services





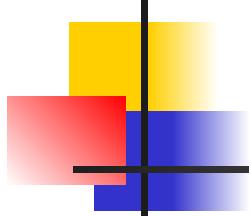
## Service – Get parts list

The web service makes available a URL to a parts list resource

Client uses : <http://www.parts-depot.com/parts>

**Document Client receives :**

```
<?xml version="1.0"?>
<p:Parts xmlns:p="http://www.parts-depot.com" xmlns:xlink="http://www.w3.org/1999/xlink">
    <Part id="00345" xlink:href="http://www.parts-depot.com/parts/00345"/>
    <Part id="00346" xlink:href="http://www.parts-depot.com/parts/00346"/>
    <Part id="00347" xlink:href="http://www.parts-depot.com/parts/00347"/>
    <Part id="00348" xlink:href="http://www.parts-depot.com/parts/00348"/>
</p:Parts>
```



## Service – Get detailed part data

The web service makes available a URL to each part resource.

**Client uses :** <http://www.parts-depot.com/parts/00345>

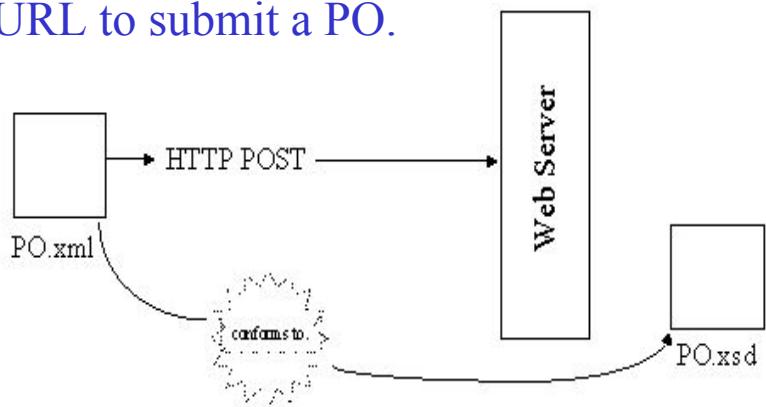
**Document Client receives :**

```
<?xml version="1.0"?>

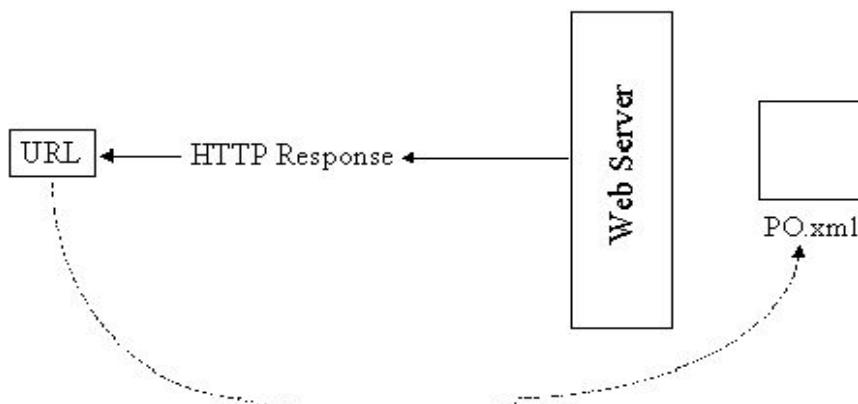
```

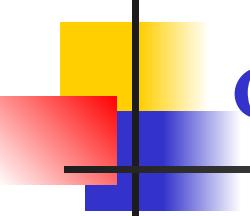
# Service – Submit purchase order (PO)

The web service makes available a URL to submit a PO.



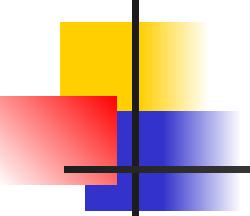
- 1)The client creates a PO instance document (PO.xml)
- 2)Submits the PO.xml(HTTP POST)
- 3)PO service reponds with a URL to the submitted PO.





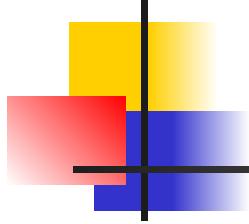
# Characteristics of a REST based network

- Client-Server: a pull-based interaction style(Client request data from servers as and when needed).
- Stateless: each request from client to server must contain all the information necessary to understand the request, and cannot take advantage of any stored context on the server.
- Cache: to improve network efficiency, responses must be capable of being labeled as cacheable or non-cacheable.
- Uniform interface: all resources are accessed with a generic interface (e.g., HTTP GET, POST, PUT, DELETE).
- Named resources - the system is comprised of resources which are named using a URL.
- Interconnected resource representations - the representations of the resources are interconnected using URLs, thereby enabling a client to progress from one state to another.



# Principles of REST web service design

- 1. Identify all the conceptual entities that we wish to expose as services. (Examples we saw include resources such as : parts list, detailed part data, purchase order)
- 2. Create a URL to each resource.
- 3. Categorize our resources according to whether clients can just receive a representation of the resource (using an HTTP GET), or whether clients can modify (add to) the resource using HTTP POST, PUT, and/or DELETE).
- 4. All resources accessible via HTTP GET should be side-effect free. That is, the resource should just return a representation of the resource. Invoking the resource should not result in modifying the resource.
- 5. Put hyperlinks within resource representations to enable clients to drill down for more information, and/or to obtain related information.
- 6. Design to reveal data gradually. Don't reveal everything in a single response document. Provide hyperlinks to obtain more details.
- 7. Specify the format of response data using a schema (DTD, W3C Schema, RelaxNG, or Schematron). For those services that require a POST or PUT to it, also provide a schema to specify the format of the response.
- 8. Describe how our services are to be invoked using either a WSDL document, or simply an HTML document.



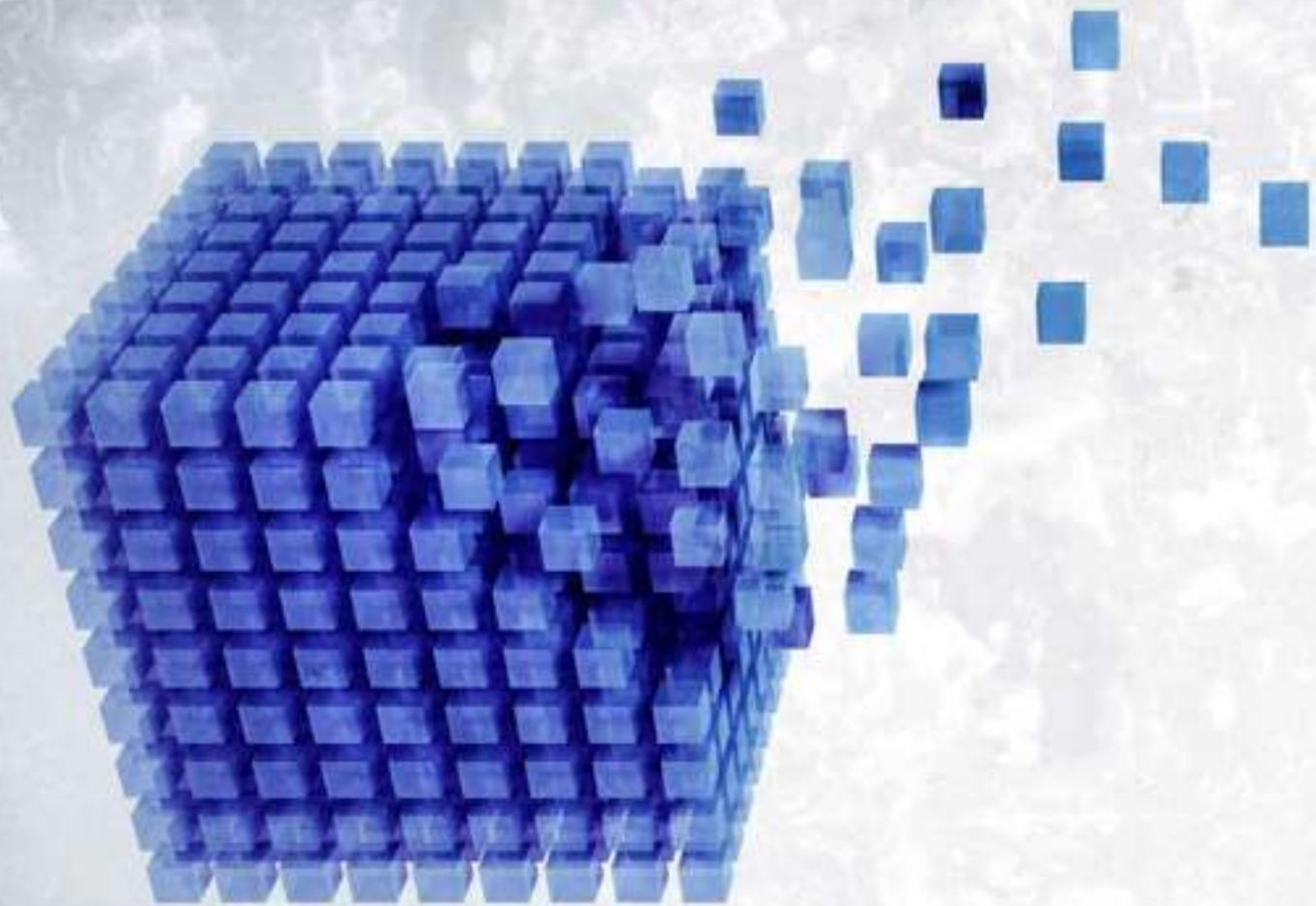
# Summary

---

- REST – Is an architectural style.
  - It is the architectural style of the WEB
- 
- Resource
- [http://www.ics.uci.edu/~fielding/pubs/dissertation/rest\\_arch\\_style.htm](http://www.ics.uci.edu/~fielding/pubs/dissertation/rest_arch_style.htm)



# VIRTUALIZATION 2.0



**tutorialspoint**  
SIMPLY EASY LEARNING

[www.tutorialspoint.com](http://www.tutorialspoint.com)



<https://www.facebook.com/tutorialspointindia>



<https://twitter.com/tutorialspoint>

## About the Tutorial

---

Virtualization 2.0 is a technology that helps users to install different Operating Systems on a hardware. They are completely separated and independent from each other. Virtualization hides the physical characteristics of computing resources from their users, their applications, or end users.

This is an introductory tutorial, which covers the basics of Virtualization 2.0 and explains how to deal with its various components and sub-components.

## Audience

---

This tutorial is created for IT Managers and System Administrators, who want to learn how to install different OS on a hardware. It provides simple, easy to understand explanations with useful working examples. We will go through most of the important modules of Virtualization 2.0, so you can also use this as a reference for your future projects.

This tutorial is intended to make you comfortable in getting started with Virtualization 2.0 and its various functions.

## Prerequisites

---

Since Virtualization 2.0 is all about operating systems and hardware, you will need to have a basic knowledge about the various OS and their elements.

Additionally, it will be helpful if you are familiar with various components such as a server, an application and various storage devices, if you want to understand all the information provided.

## Copyright and Disclaimer

---

© Copyright 2017 by Tutorials Point (I) Pvt. Ltd.

All the content and graphics published in this e-book are the property of Tutorials Point (I) Pvt. Ltd. The user of this e-book is prohibited to reuse, retain, copy, distribute or republish any contents or a part of contents of this e-book in any manner without written consent of the publisher.

We strive to update the contents of our website and tutorials as timely and as precisely as possible, however, the contents may contain inaccuracies or errors. Tutorials Point (I) Pvt. Ltd. provides no guarantee regarding the accuracy, timeliness or completeness of our website or its contents including this tutorial. If you discover any errors on our website or in this tutorial, please notify us at [contact@tutorialspoint.com](mailto:contact@tutorialspoint.com)

## Table of Contents

---

About the Tutorial .....	i
Audience.....	i
Prerequisites.....	i
Copyright and Disclaimer .....	i
Table of Contents .....	ii
<b>1. Virtualization – Overview .....</b>	<b>1</b>
Types of Virtualization .....	1
Understanding Different Types of Hypervisors .....	4
Understanding Local Virtualization and Cloud .....	6
<b>2. Virtualization – Pros and Cons .....</b>	<b>7</b>
Advantages of Virtualization .....	7
Disadvantages of Virtualization.....	8
<b>3. Virtualization – Common Uses .....</b>	<b>10</b>
Virtualizing Desktop Computers.....	10
Running a Specific Program Solution.....	11
Setting up Test and Development Environments.....	12
<b>4. Virtualization – Hardware .....</b>	<b>13</b>
Understanding Virtual CPU.....	13
Understanding Virtual Memory.....	14
Understanding Virtual Storage.....	15
Understanding Virtual Networking.....	16
<b>5. Virtualization – Microsoft Hyper-V.....</b>	<b>17</b>
Installing Hyper-V in Windows Server 2012 .....	17
Installing Hyper-V in a windows 10 workstation .....	22
Creating a Virtual Machine with Hyper-V.....	24
Setting up Networking with Hyper-V.....	29
Allocating Processors & Memory to a VM using Hyper-V .....	32
Using Checkpoints in Hyper-V .....	37
<b>6. Virtualization – VMware Workstation Player .....</b>	<b>39</b>
Installing VMware Workstation Player .....	39
Creating a VM with VMware Workstation .....	44
Setting up Networking with VMware Workstation .....	47
Allocating Processors & Memory to a VM using VMware Workstation .....	50
Duplicating a VM Using VMware Workstation .....	53
<b>7. Virtualization – VirtualBox .....</b>	<b>58</b>
Installing VirtualBox.....	58
Creating a VM with VirtualBox .....	62
Setting up Networking with VirtualBox .....	69
Allocating Processors & Memory to a VM.....	74
Duplicating a VM Using VirtualBox.....	76
Deleting a VM on VirtualBox .....	78

<b>8. Virtualization – Openstack .....</b>	<b>79</b>
Understanding Openstack .....	79
Installing Openstack .....	81
Installing Openstack on Ubuntu 14.04 .....	82
<b>9. Virtualization – Preparing the Infrastructure .....</b>	<b>88</b>
Understanding Different File Systems .....	88
Choosing Between Different Types of Storage.....	89
<b>10. Virtualization – Troubleshooting .....</b>	<b>92</b>
Troubleshooting Network Communication .....	92
Troubleshooting Slow Performance .....	93
<b>11. Virtualization – Backing Up, Restoring &amp; Migrating VM .....</b>	<b>94</b>
Duplicating a VM .....	94
Backing Up and Recovering a VM.....	94
Converting a Physical Server into a Virtual Server .....	96
Converting a Virtual Server into a Physical Server .....	98

# 1. Virtualization – Overview

Virtualization is a technology that helps us to install different Operating Systems on a hardware. They are completely separated and independent from each other. In Wikipedia, you can find the definition as – “In computing, virtualization is a broad term that refers to the abstraction of computer resources.

Virtualization hides the physical characteristics of computing resources from their users, their applications or end users. This includes making a single physical resource (such as a server, an operating system, an application or a storage device) appear to function as multiple virtual resources. It can also include making multiple physical resources (such as storage devices or servers) appear as a single virtual resource...”

Virtualization is often:

- The creation of many virtual resources from one physical resource.
- The creation of one virtual resource from one or more physical resource

## Types of Virtualization

---

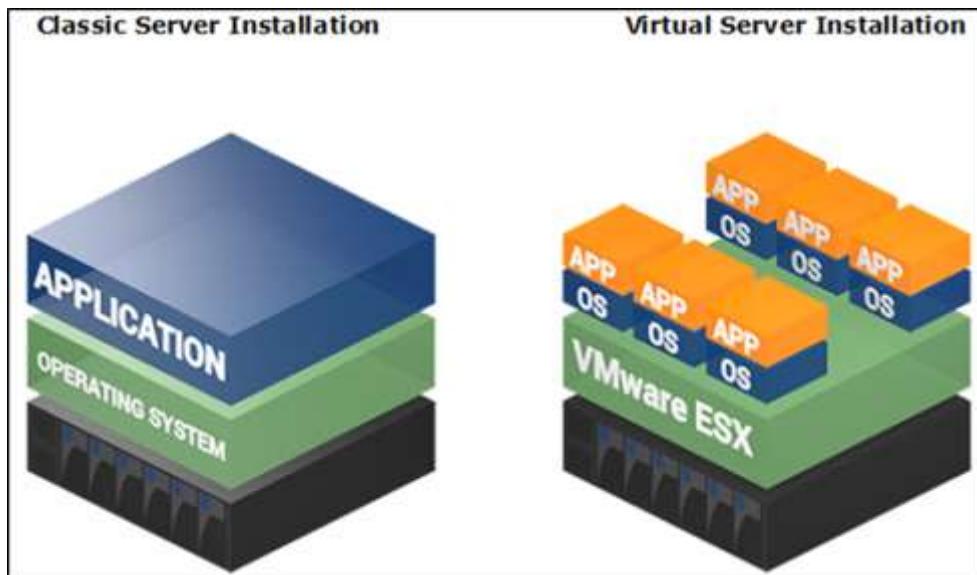
Today the term virtualization is widely applied to a number of concepts, some of which are described below:

- Server Virtualization
- Client & Desktop Virtualization
- Services and Applications Virtualization
- Network Virtualization
- Storage Virtualization

Let us now discuss each of these in detail.

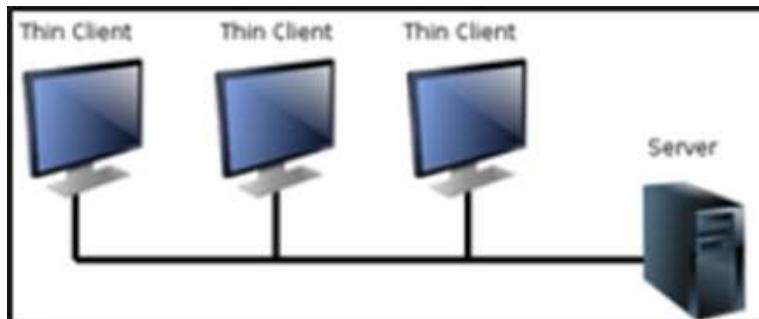
### Server Virtualization

It is virtualizing your server infrastructure where you do not have to use any more physical servers for different purposes.



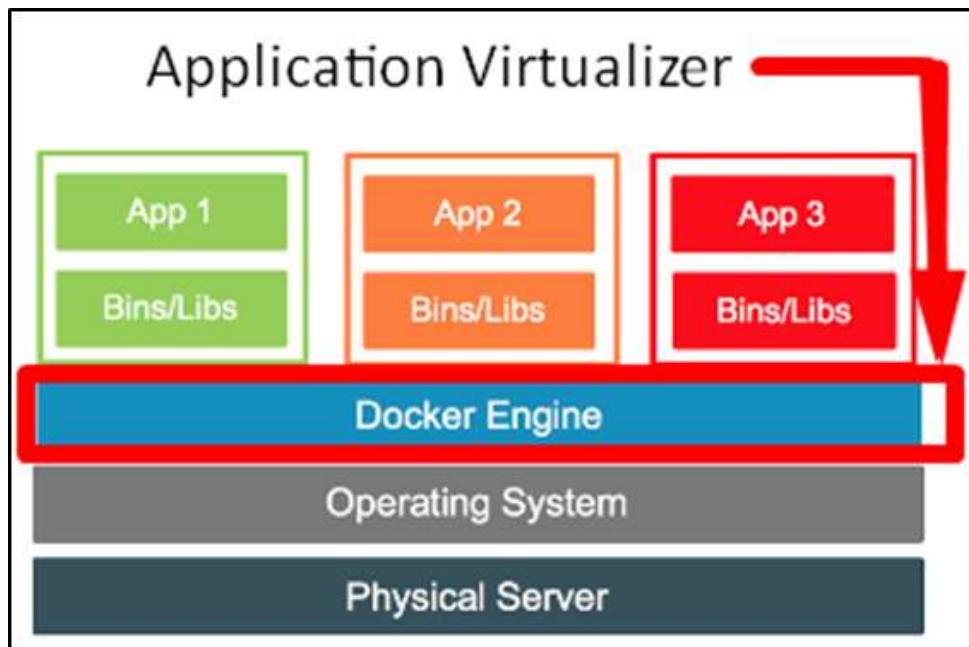
## Client & Desktop Virtualization

This is similar to server virtualization, but this time is on the user's site where you virtualize their desktops. We change their desktops with thin clients and by utilizing the datacenter resources.



## Services and Applications Virtualization

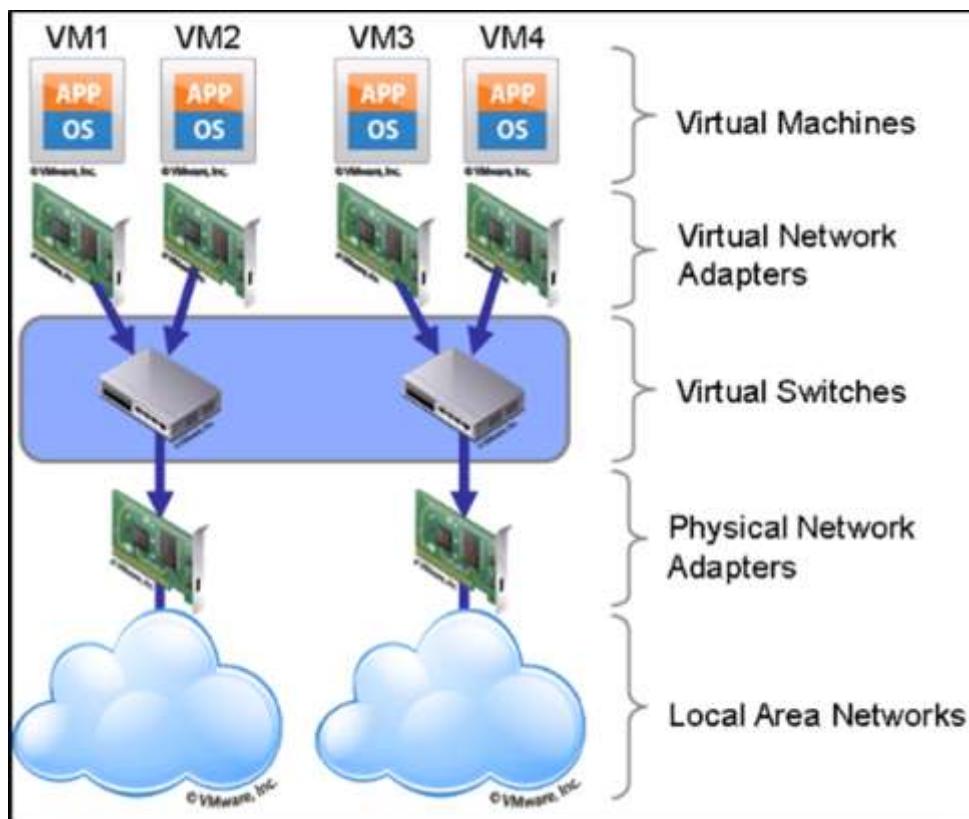
The virtualization technology isolates applications from the underlying operating system and from other applications, in order to increase compatibility and manageability. For example – Docker can be used for that purpose.



## Network Virtualization

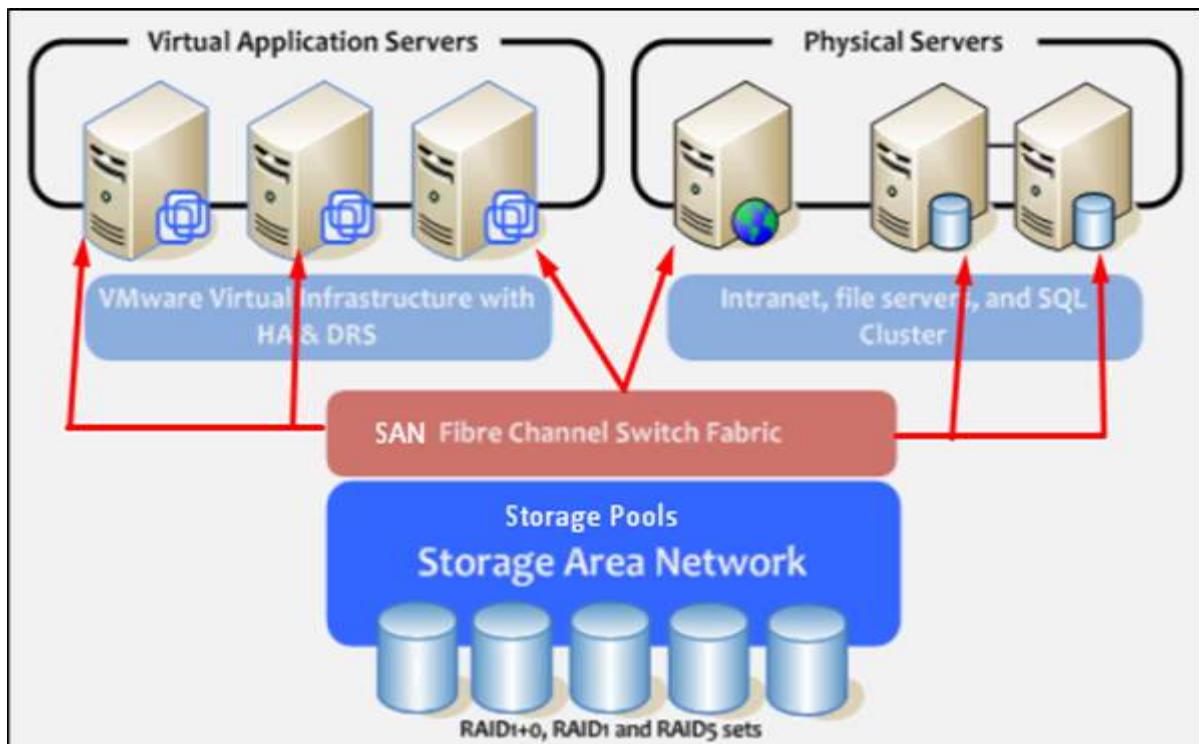
It is a part of virtualization infrastructure, which is used especially if you are going to visualize your servers. It helps you in creating multiple switching, Vlans, NAT-ing, etc.

The following illustration shows the VMware schema:



## Storage Virtualization

This is widely used in datacenters where you have a big storage and it helps you to create, delete, allocated storage to different hardware. This allocation is done through network connection. The leader on storage is SAN. A schematic illustration is given below:



## Understanding Different Types of Hypervisors

A hypervisor is a thin software layer that intercepts operating system calls to the hardware. It is also called as the **Virtual Machine Monitor** (VMM). It creates a virtual platform on the host computer, on top of which multiple guest operating systems are executed and monitored.

Hypervisors are two types:

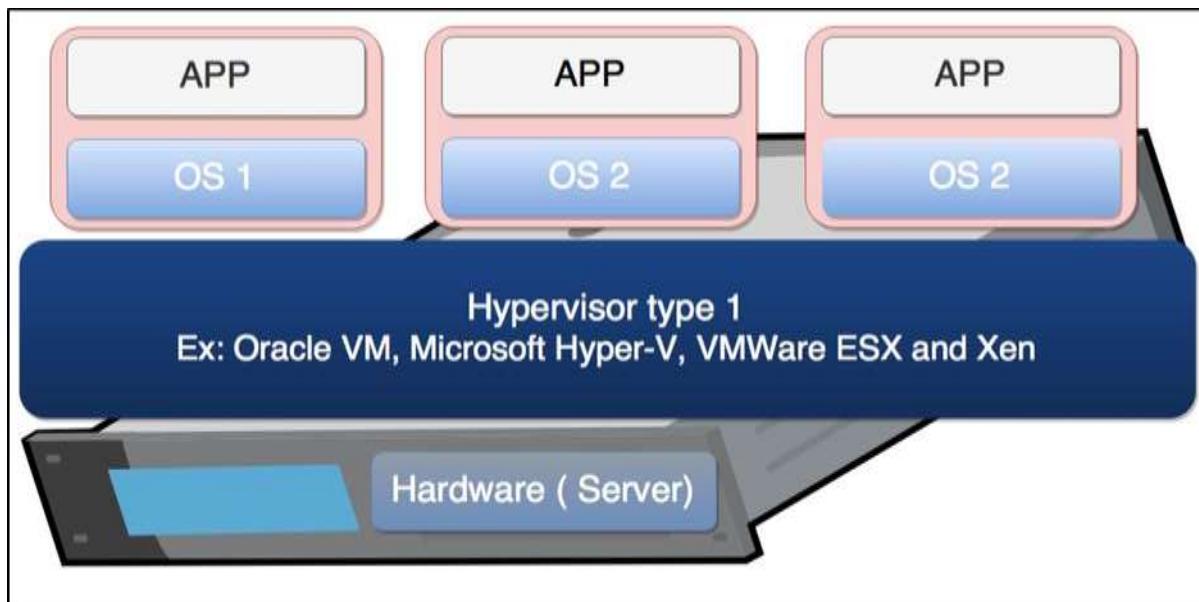
- Native or Bare Metal Hypervisor and
- Hosted Hypervisor

Let us now discuss both of these in detail.

### Native or Bare Metal Hypervisor

Native hypervisors are software systems that run directly on the host's hardware to control the hardware and to monitor the **Guest Operating Systems**. The guest operating system runs on a separate level above the hypervisor. All of them have a Virtual Machine Manager.

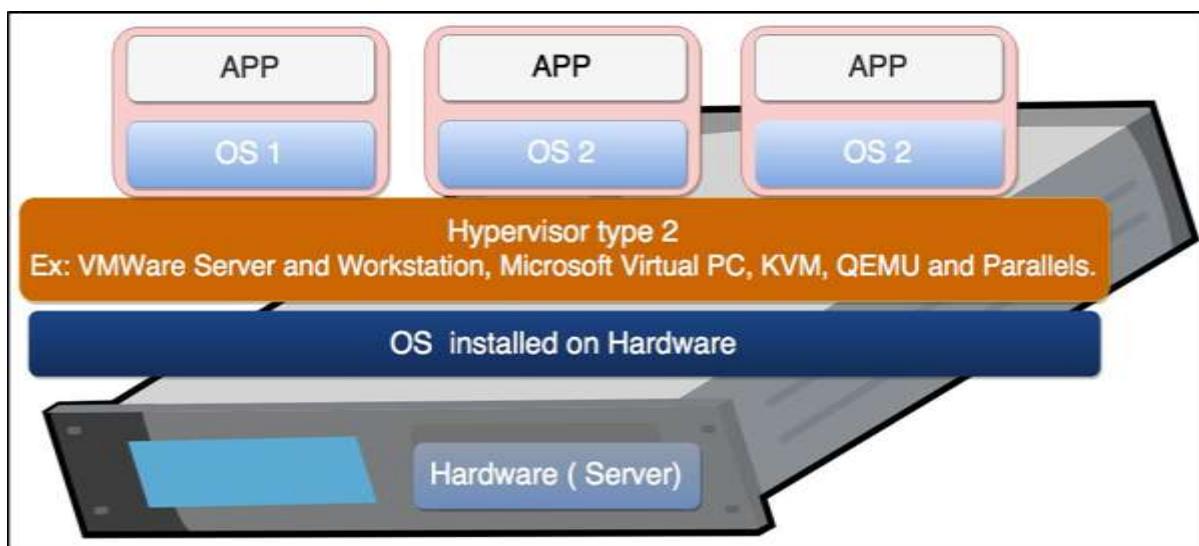
Examples of this virtual machine architecture are **Oracle VM**, **Microsoft Hyper-V**, **VMWare ESX** and **Xen**.



## Hosted Hypervisor

Hosted hypervisors are designed to run within a traditional operating system. In other words, a hosted hypervisor adds a distinct software layer on top of the host operating system. While, the guest operating system becomes a third software level above the hardware.

A well-known example of a hosted hypervisor is **Oracle VM VirtualBox**. Others include **VMWare Server and Workstation**, **Microsoft Virtual PC**, **KVM**, **QEMU** and **Parallels**.



## Understanding Local Virtualization and Cloud

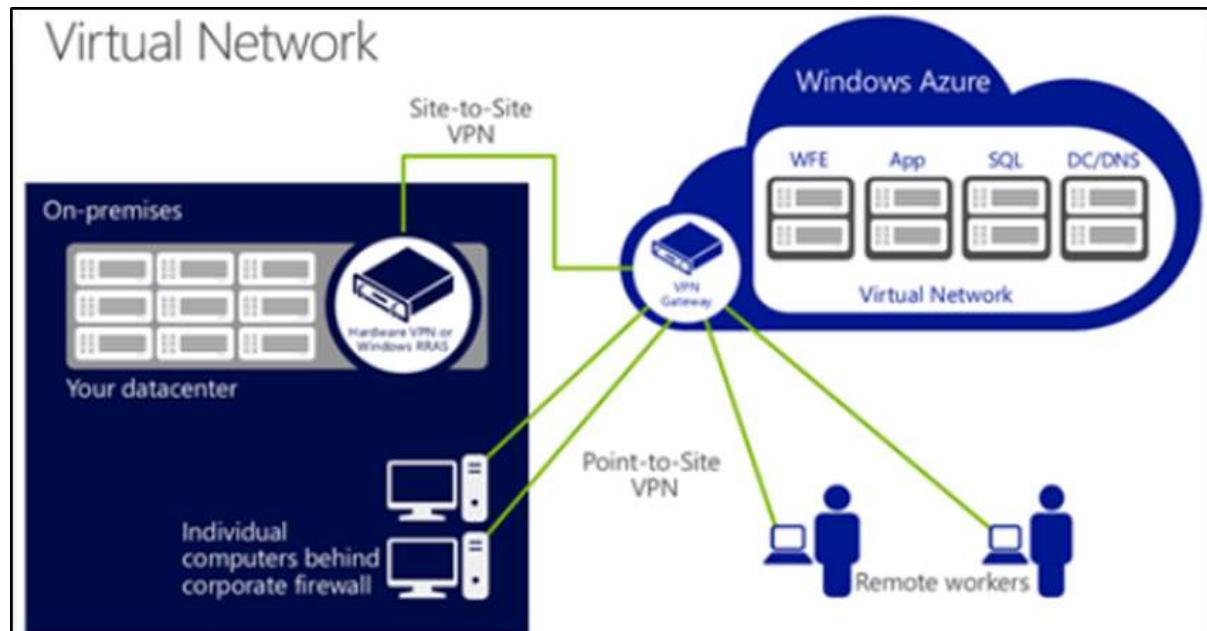
Virtualization is one of the fundamental technologies that makes cloud-computing work. However, virtualization is not cloud computing. Cloud computing is a service that different providers offer to you based on some costs.

In enterprise networks, virtualization and cloud computing are often used together to build a public or private cloud infrastructure. In small businesses, each technology will be deployed separately to gain measurable benefits. In different ways, virtualization and cloud computing can help you keep your equipment spending to a minimum and get the best possible use from the equipment you already have.

As mentioned before, virtualization software allows one physical server to run several individual computing environments. In practice, it is like getting multiple servers for each physical server you buy. This technology is fundamental to cloud computing. Cloud providers have large data centers full of servers to power their cloud offerings, but they are not able to devote a single server to each customer. Thus, they virtually partition the data on the server, enabling each client to work with a separate "virtual" instance (which can be a private network, servers farm, etc.) of the same software.

Small businesses are most likely to adopt cloud computing by subscribing to a cloud-based service. The largest providers of cloud computing are **Microsoft with Azure** and **Amazon**.

The following illustration is provided by Microsoft where you can understand how utilizing extra infrastructure for your business without the need to spend extra money helps. You can have the on-premises base infrastructure, while on cloud you can have all your services, which are based on Virtualized technology.



## 2. Virtualization – Pros and Cons

In this chapter, we will discuss some of the most common advantages and disadvantages of Virtualization.

### **Advantages of Virtualization**

---

Following are some of the most recognized advantages of Virtualization, which are explained in detail.

#### **Using Virtualization for Efficient Hardware Utilization**

Virtualization decreases costs by reducing the need for physical hardware systems. Virtual machines use efficient hardware, which lowers the quantities of hardware, associated maintenance costs and reduces the power along with cooling the demand. You can allocate memory, space and CPU in just a second, making you more self-independent from hardware vendors.

#### **Using Virtualization to Increase Availability**

Virtualization platforms offer a number of advanced features that are not found on physical servers, which increase uptime and availability. Although the vendor feature names may be different, they usually offer capabilities such as live migration, storage migration, fault tolerance, high availability and distributed resource scheduling. These technologies keep virtual machines chugging along or give them the ability to recover from unplanned outages.

The ability to move a virtual machine from one server to another is perhaps one of the greatest single benefits of virtualization with far reaching uses. As the technology continues to mature to the point where it can do long-distance migrations, such as being able to move a virtual machine from one data center to another no matter the network latency involved.

#### **Disaster Recovery**

Disaster recovery is very easy when your servers are virtualized. With up-to-date snapshots of your virtual machines, you can quickly get back up and running. An organization can more easily create an affordable replication site. If a disaster strikes in the data center or server room itself, you can always move those virtual machines elsewhere into a cloud provider. Having that level of flexibility means your disaster recovery plan will be easier to enact and will have a 99% success rate.

#### **Save Energy**

Moving physical servers to virtual machines and consolidating them onto far fewer physical servers' means lowering monthly power and cooling costs in the data center. It reduces carbon footprint and helps to clean up the air we breathe. Consumers want to see companies reducing their output of pollution and taking responsibility.

## Deploying Servers too fast

You can quickly clone an image, master template or existing virtual machine to get a server up and running within minutes. You do not have to fill out purchase orders, wait for shipping and receiving and then rack, stack, and cable a physical machine only to spend additional hours waiting for the operating system and applications to complete their installations. With virtual backup tools like [Veeam](#), redeploying images will be so fast that your end users will hardly notice there was an issue.

## Save Space in your Server Room or Datacenter

Imagine a simple example: you have two racks with 30 physical servers and 4 switches. By virtualizing your servers, it will help you to reduce half the space used by the physical servers. The result can be two physical servers in a rack with one switch, where each physical server holds 15 virtualized servers.

## Testing and setting up Lab Environment

While you are testing or installing something on your servers and it crashes, do not panic, as there is no data loss. Just revert to a previous snapshot and you can move forward as if the mistake did not even happen. You can also isolate these testing environments from end users while still keeping them online. When you have completely done your work, deploy it in live.

## Shifting all your Local Infrastructure to Cloud in a day

If you decide to shift your entire virtualized infrastructure into a cloud provider, you can do it in a day. All the hypervisors offer you tools to export your virtual servers.

## Possibility to Divide Services

If you have a single server, holding different applications this can increase the possibility of the services to crash with each other and increasing the fail rate of the server. If you virtualize this server, you can put applications in separated environments from each other as we have discussed previously.

## Disadvantages of Virtualization

---

Although you cannot find many disadvantages for virtualization, we will discuss a few prominent ones as follows:

### Extra Costs

Maybe you have to invest in the virtualization software and possibly additional hardware might be required to make the virtualization possible. This depends on your existing network. Many businesses have sufficient capacity to accommodate the virtualization without requiring much cash. If you have an infrastructure that is more than five years old, you have to consider an initial renewal budget.

## Software Licensing

This is becoming less of a problem as more software vendors adapt to the increased adoption of virtualization. However, it is important to check with your vendors to understand how they view software use in a virtualized environment.

## Learn the new Infrastructure

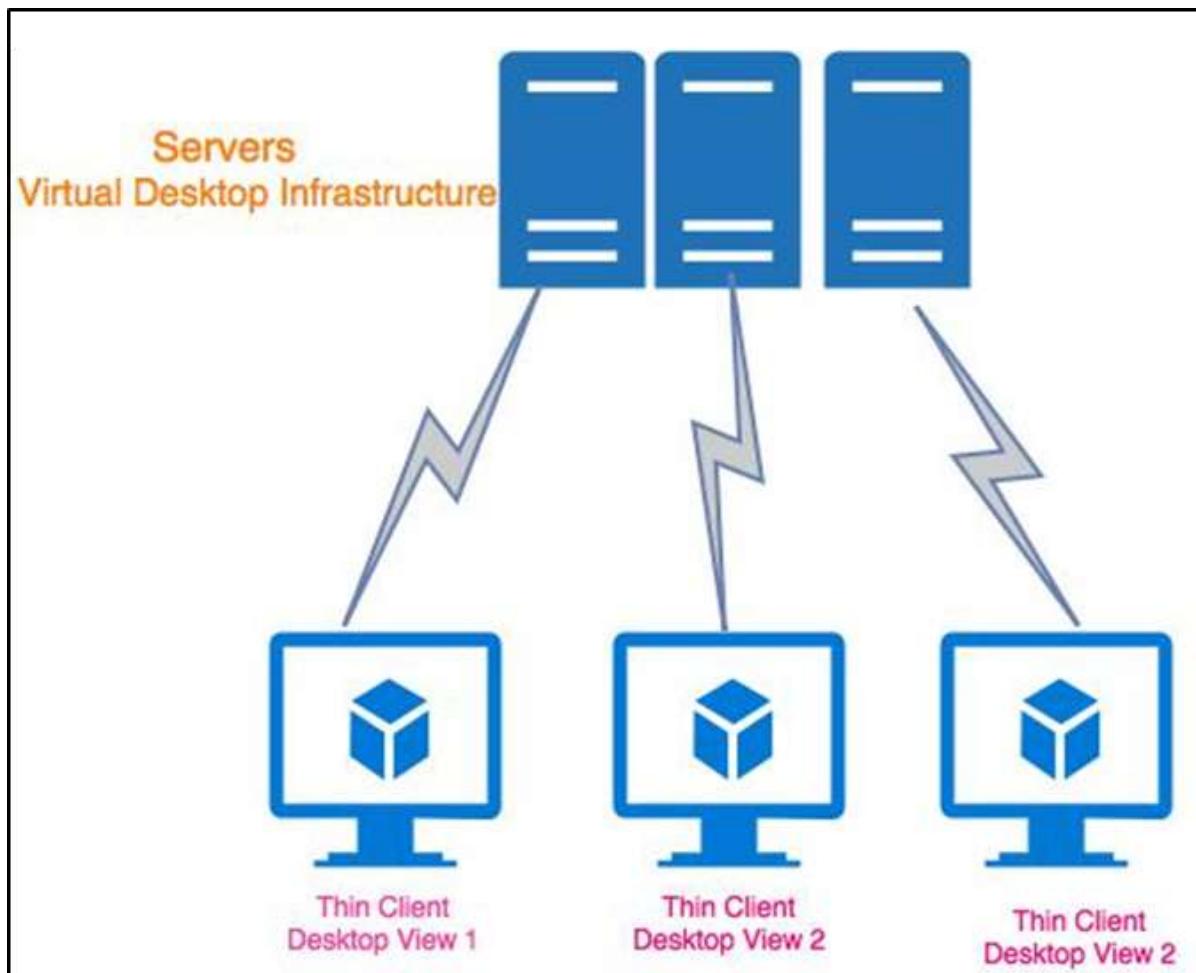
Implementing and managing a virtualized environment will require IT staff with expertise in virtualization. On the user side, a typical virtual environment will operate similarly to the non-virtual environment. There are some applications that do not adapt well to the virtualized environment.

### 3. Virtualization – Common Uses

There are several types of uses in virtualization, but the most commonly used are **Server Virtualization** and **Client Desktops Virtualization**. We have already discussed their advantages in the previous chapter and why are they used widely.

#### Virtualizing Desktop Computers

Client virtualization also called as VDI runs the entire desktop environment within a centralized server. All processing is done within the server. Client devices are typically thin clients that serve as an end node to connect I/O peripherals such as keyboard, mouse, a display, audio connectors and even USB ports over the LAN.



As discussed earlier, a thin client hardware is a computer terminal, which provides I/O for a keyboard, mouse, monitor, jacks for sound peripherals, and open ports for USB devices.

For example – Printer, Flash Drive, Web Cam, Card Reader, Smartphone, etc. Some thin clients include legacy serial and/or parallel ports to support older devices such as Receipt Printers, Scales, Time Clocks, etc. Thin client software typically consists of a GUI (graphical

user interface), Cloud Access Agents (for e.g. RDP, ICA, PCoIP), a local web browser, terminal emulations (in some cases) and a basic set of local utilities.

The largest producers of thin clients are HP, Dell and IBM.



## Running a Specific Program Solution

One of the best software known for Desktop Virtualization is [XenApp](#) & [XenDesktop](#). Deliver Windows, Linux, Web and SaaS applications or full virtual desktops to workers on any device and anywhere.

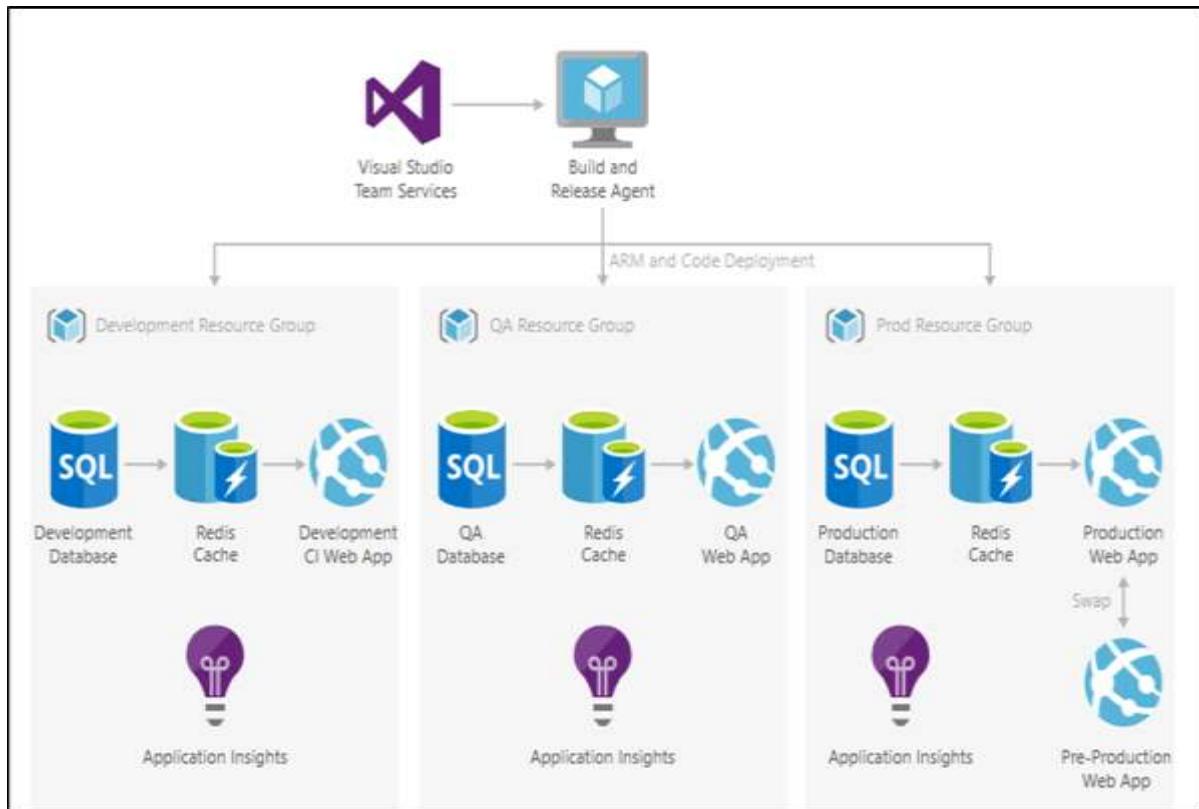
More information can be found on <https://www.citrix.com/products/xenapp-xendesktop/>.

Another major provider is VMware with its platform called **VMware Horizon 7**. To purchase and understand all their features and specifications, click on the following link – <http://www.vmware.com/products/horizon.html>.

Microsoft is another provider with a combination of Remote desktop services along with **Hyper-V**. For any further Information on this, click on the following link – <https://technet.microsoft.com/en-us/windowsserver/ee236407.aspx>.

## Setting up Test and Development Environments

One of the most powerful features of a virtualized environment is the possibility to create labs for different approaches in a minute, especially in software development and then to import the same infrastructure in production.



Regarding the test environment, it brings cross-platform functionality to your dev-test environment and uses your preferred coding language to build natively. It tests your applications on the devices and platforms you use today: from Linux, Windows to iOS and Android.

All the features mentioned in the above can be done through private cloud or public cloud. It depends on what we would like to use as per the requirement. You should take into consideration the human resources with which you have to manage this cloud and the budget that you want to spend.

## 4. Virtualization – Hardware

In this chapter, we will discuss various components of hardware such as CPU, Memory, Storage and Networking.

### **Understanding Virtual CPU**

---

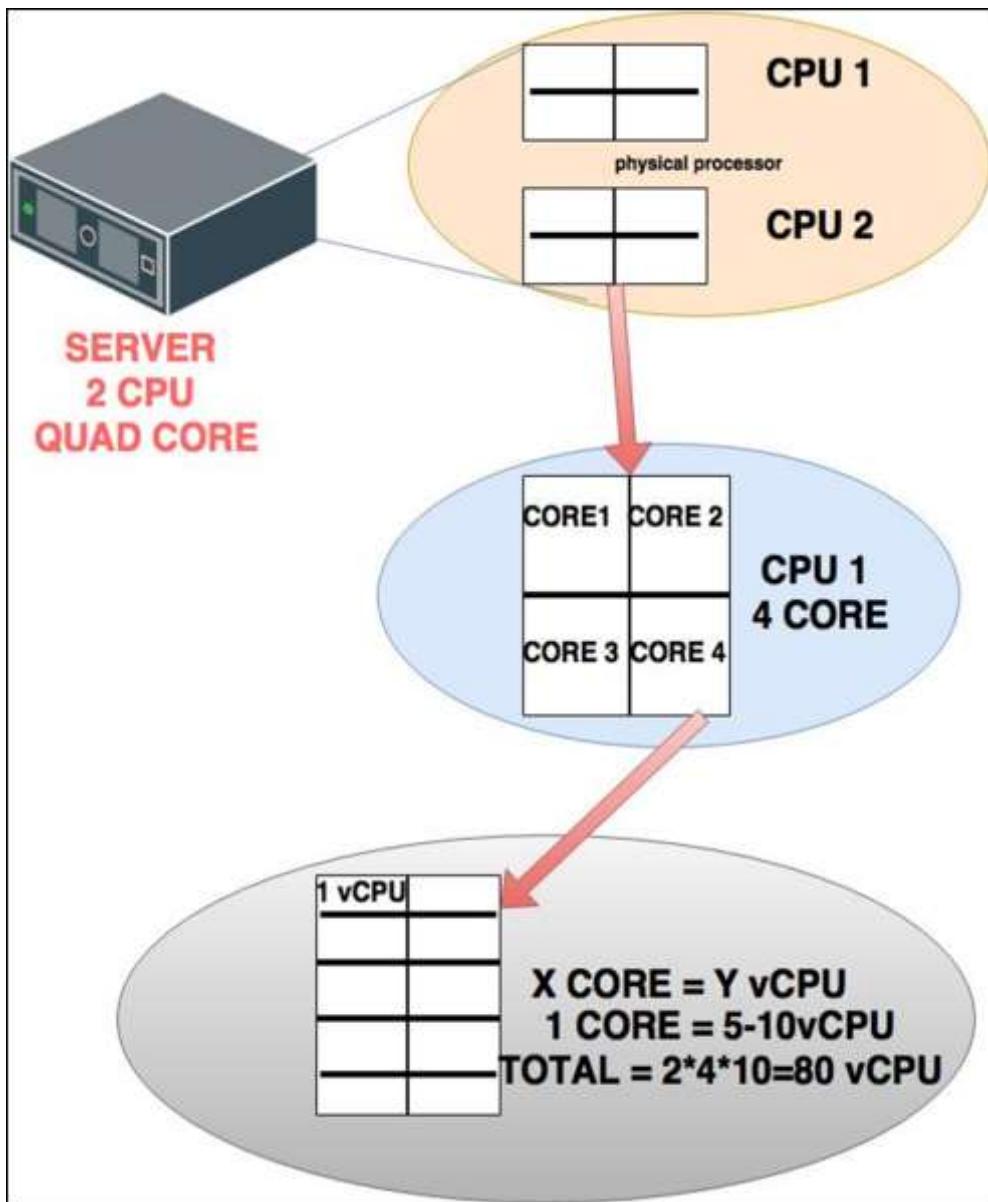
When we install a hypervisor, each physical CPU is abstracted into virtual CPUs. This divides the available CPU cycles for each core and allows multiple VMs to "time share" a given physical processor core. Generally talking, the hypervisor typically assigns one workload per vCPU (per core). If the workloads on a server need more CPU cycles, it is better to deploy fewer VMs on a particular virtual CPU.

Let us consider the following example to understand the logic of virtual CPU.

I have a physical server with two processors (CPU 1 and CPU 2) and each of them has four physical cores. In total, we have  $2 \times 4 = 8$  physical cores.

Based on some calculations our hypervisor provided for each physical core, we can get 5-10 vCPUs.

In total, we will have [8 physical cores \* (5 to 10 vCPUs)] 40-80 vCPUs, which means that we can assign a maximum of 80 vCPUs to virtual machines.



## Understanding Virtual Memory

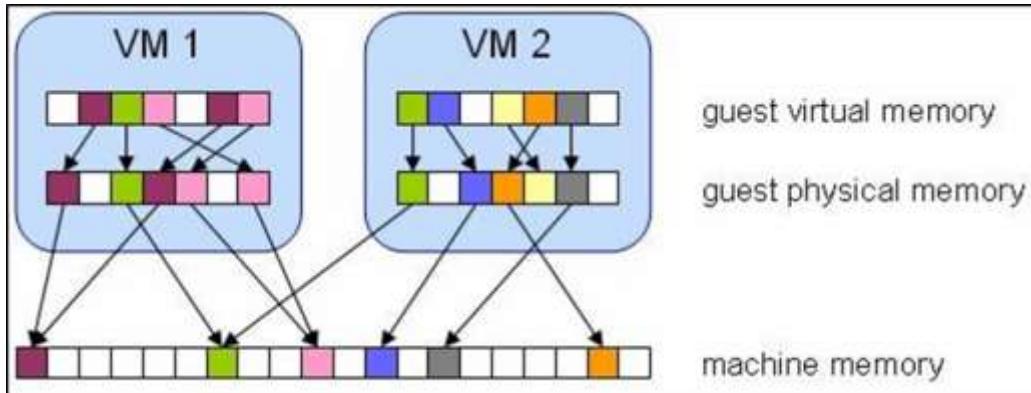
Virtual Memory in simple words is the RAM of the machine. The memory resource settings for a virtual machine determines how much of the host's memory is allocated to the virtual machine. The virtual hardware memory size determines how much memory is available to applications that run in the virtual machine.

A virtual machine cannot benefit from more memory resources than its configured virtual hardware memory size. The **ESXi hosts** limit the memory resource use to the maximum amount useful for the virtual machine, so that you can accept the default of unlimited memory resources.

You can add, change, and configure virtual machine memory resources or options to enhance virtual machine performance. You can set most of the memory parameters while creating the virtual machine or it can also be done after the **Guest Operating System** is

installed. Most of the hypervisors require to power off the virtual machine before changing the settings.

In the following schematic illustration, you can see that the total physical memory is divided between two virtual machines.

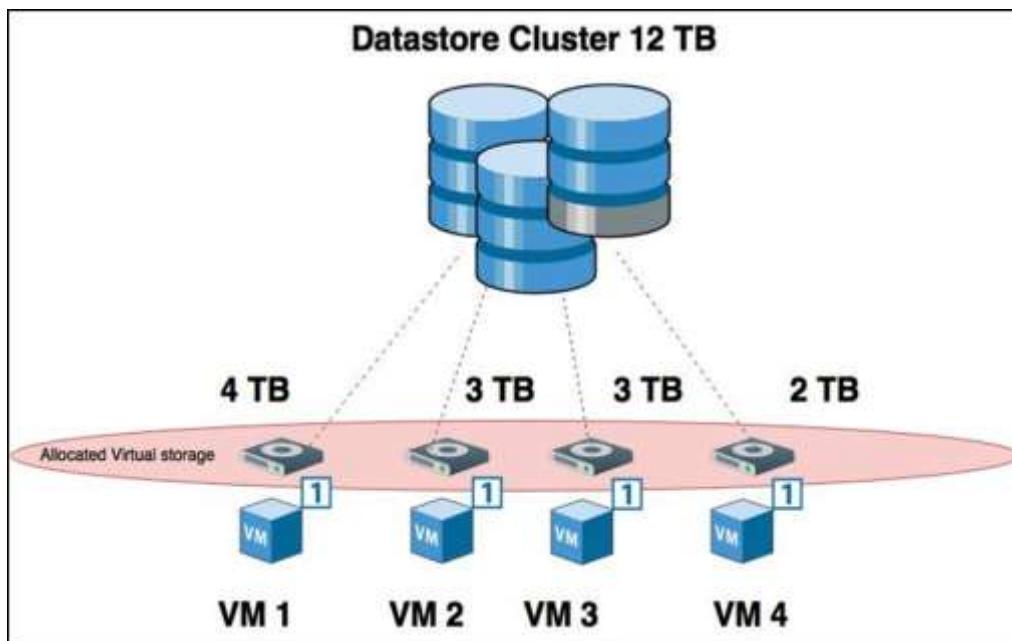


## Understanding Virtual Storage

Storage virtualization is the pooling of physical storage (Data cluster) from multiple network storage devices into what appears to be a single storage device that is managed from a central console. We cannot assign more storage to virtual machines than data cluster offers physically.

You will see these extensions on the end of a file. Of all the files used as part of a virtual machine, different hypervisors like to use different file types. Some of the more common are **VDI**, **VHDX**, **VMDK** and **HDD**.

In the following example, we have a data cluster of 12 TB in total and four virtual machines to which we have allocated storage to each of them. In total, the maximum storage allocated to them is 12 TB.



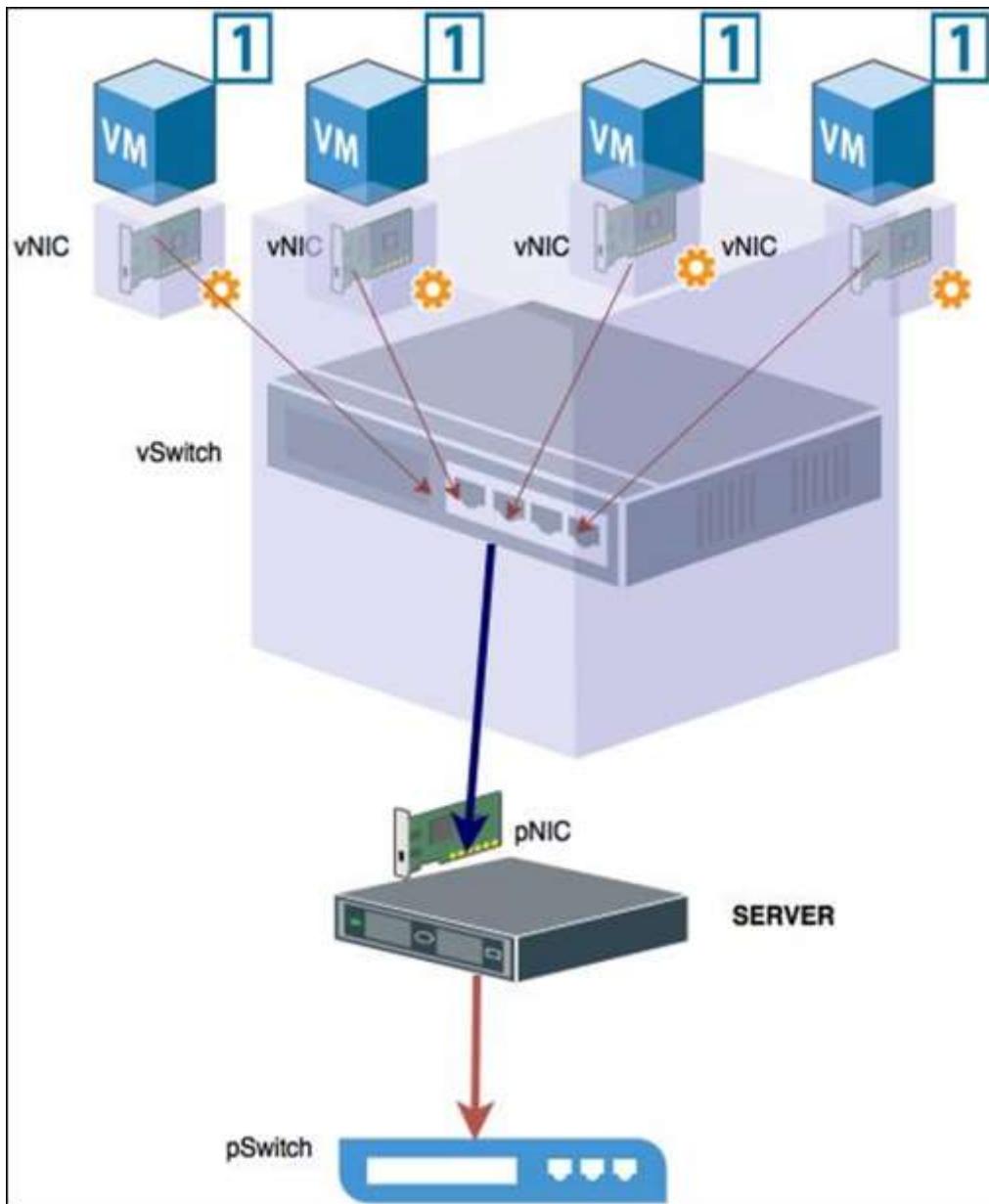
## Understanding Virtual Networking

We will discuss this with a simple example as to how virtual networking done.

We have Virtual Machine 1, 2, 3 and 4 running on the same host. They would like to send the network traffic back and forth. This is done by virtual networking cards as shown in the following illustration (vNIC), which connects virtually with a virtual switch (vSwitch) that is created by the hypervisor.

This virtual switch communicates with a physical card of the server (pNIC), which is connected with a physical switch (pSwitch) and then communicates with the rest of the network equipment.

Please see the following schematically done up scenario.



## 5. Virtualization – Microsoft Hyper-V

In this chapter, we will discuss Microsoft Hyper-V along with its various modules.

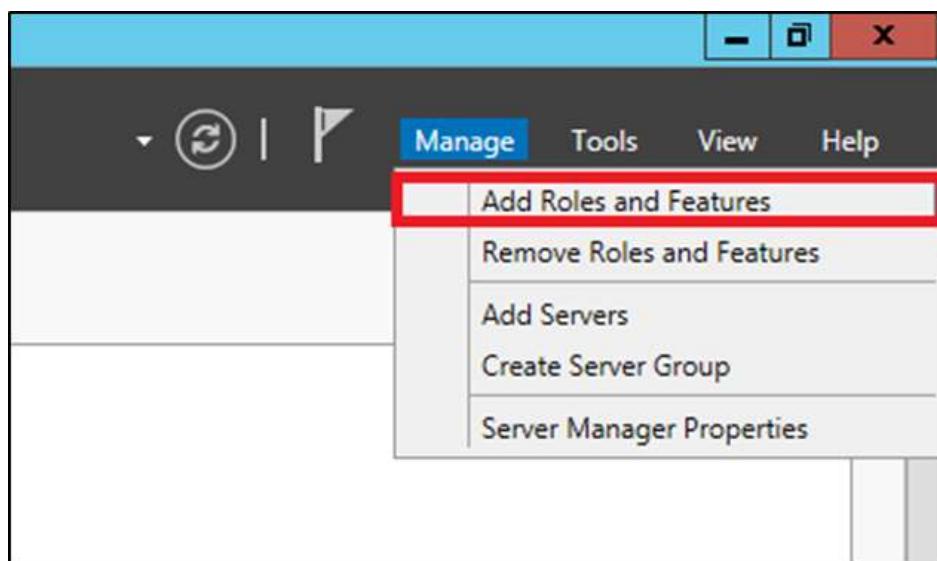
### Installing Hyper-V in Windows Server 2012

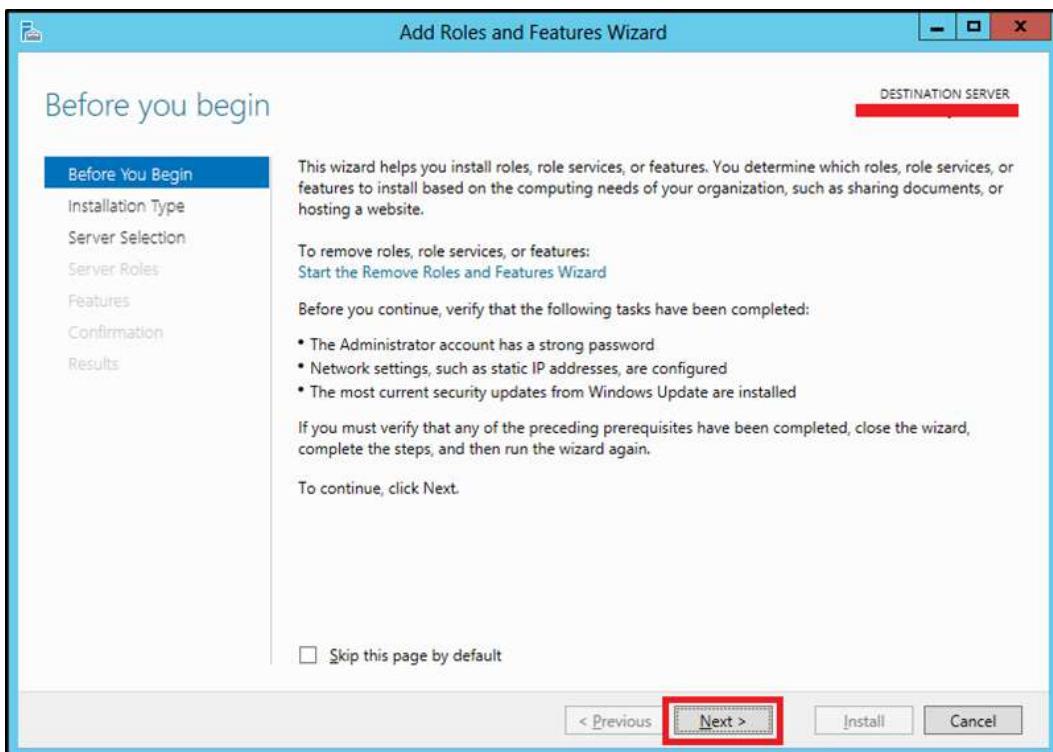
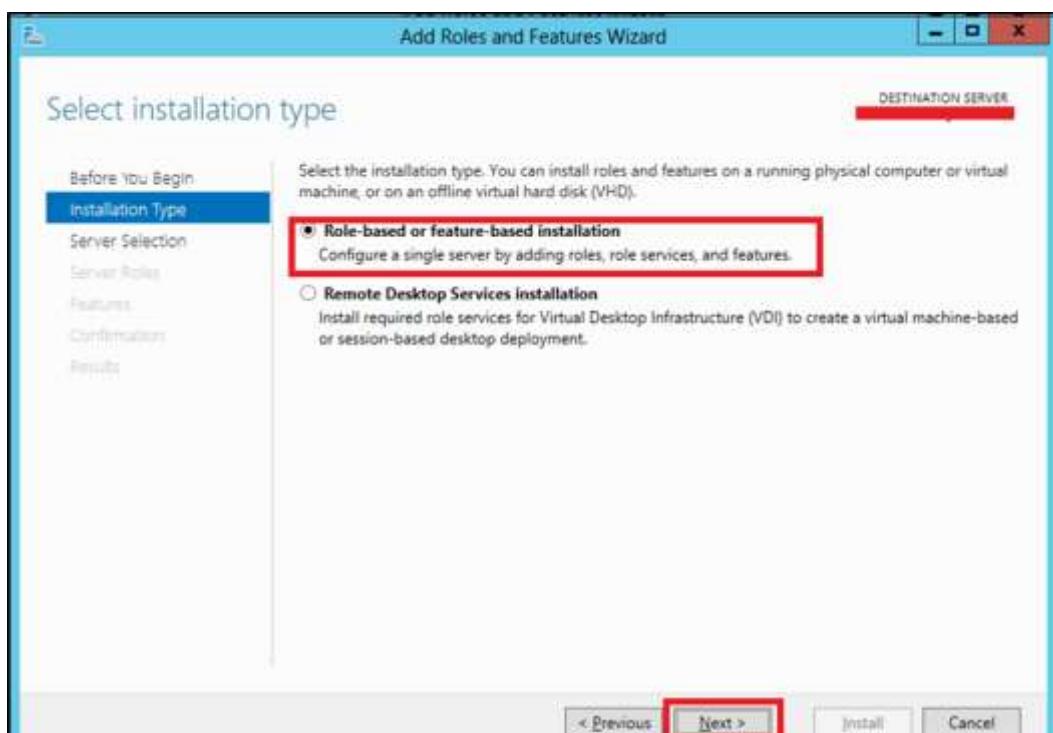
Microsoft Hyper-V, known as Windows Server Virtualization, is a native (bare) hypervisor. It can create virtual machines on x86-64 systems running Windows OS, starting with Windows 8. Hyper-V supersedes **Windows Virtual PC** as the hardware virtualization component of the client editions of Windows NT. A server computer running Hyper-V can be configured to expose individual virtual machines to one or more networks.

Hyper-V was first released alongside Windows Server 2008 and Windows 7 and has been available without charge for all the Windows Server versions and some client operating systems since that time.

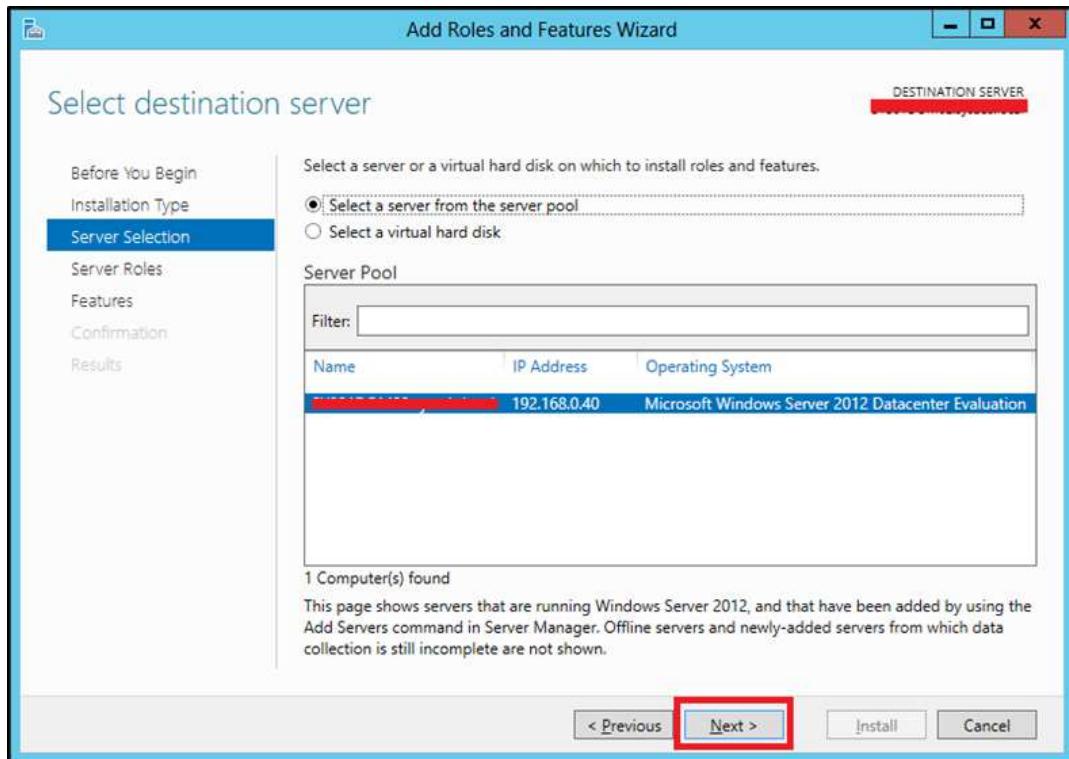
Let us see how to install a Hyper-V role in a Windows Server 2012 by following the steps given below.

**Step 1:** To Install Hyper-V role go to “Server Manager” → Manage → Add Roles and Features.

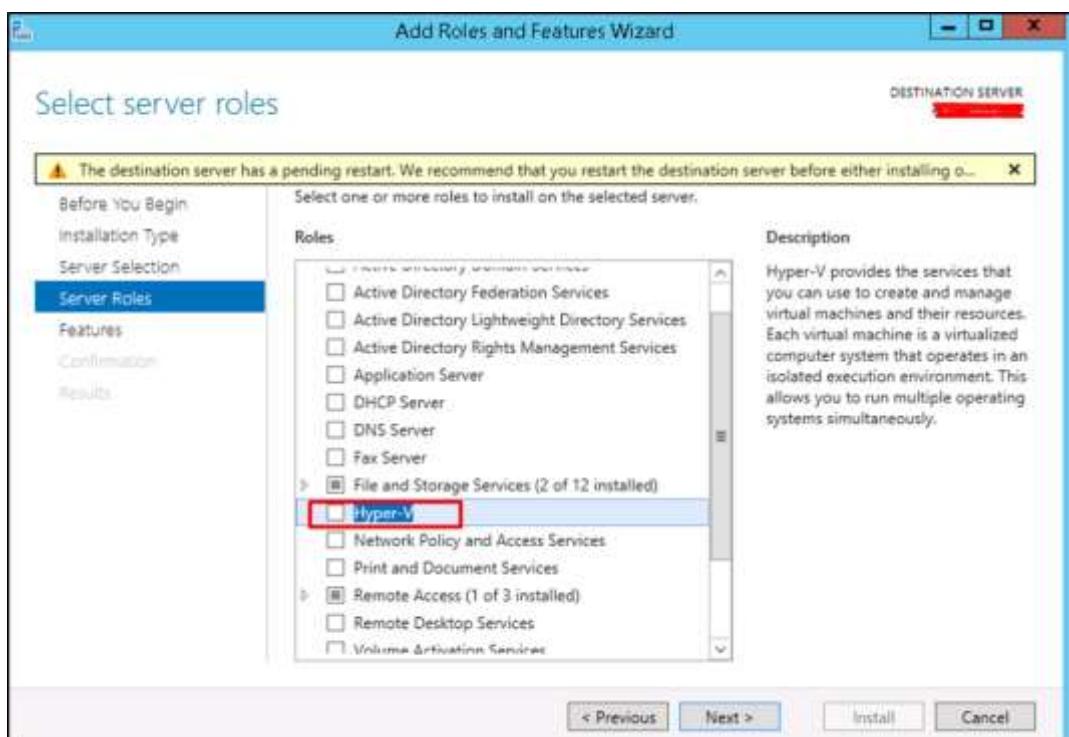


**Step 2:** Click on "Next".**Step 3:** Select "Role-based or feature-based installation" option → click on "Next".

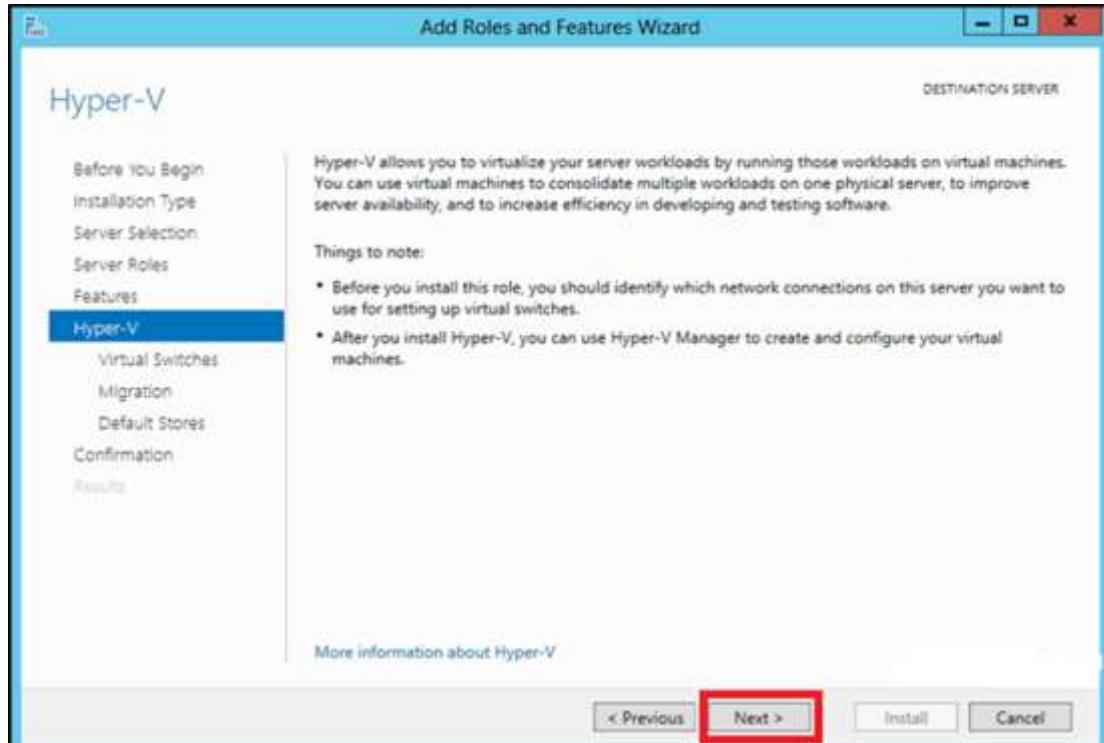
**Step 4:** We will locally install the Hyper-V role as such “Select a server from the server pool” → click “Next”.



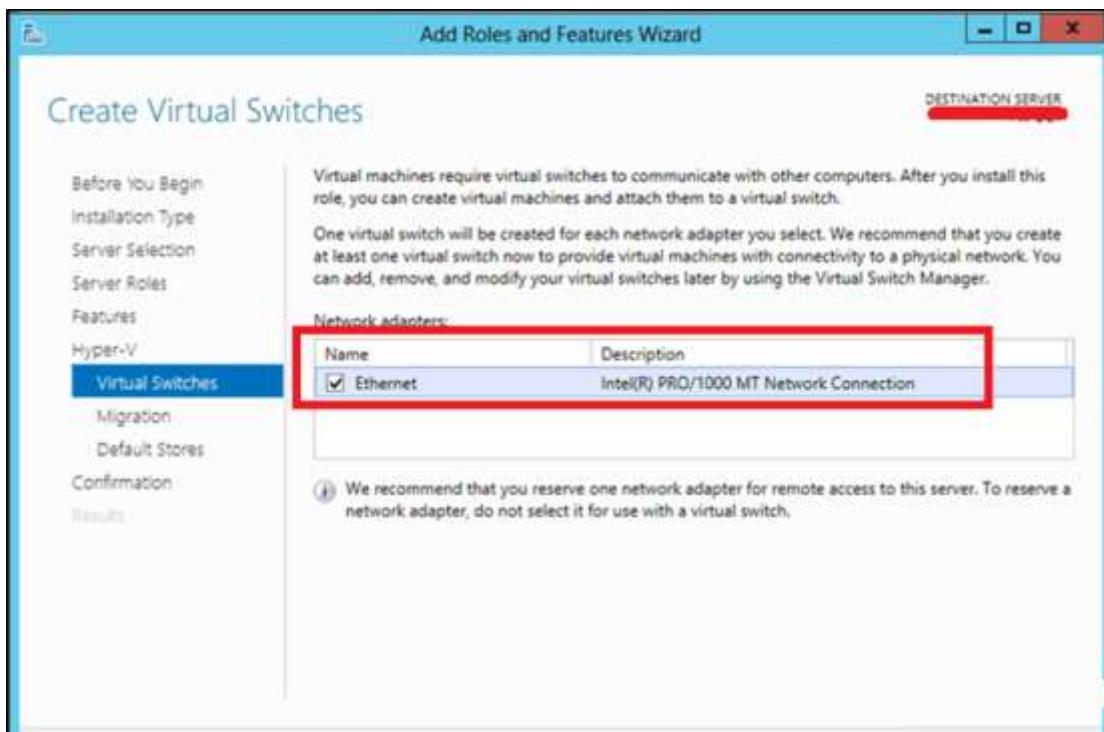
**Step 5:** From the Roles lists, check the “Hyper-V” Server role → click on Add Features on the popup window → click “Next”.



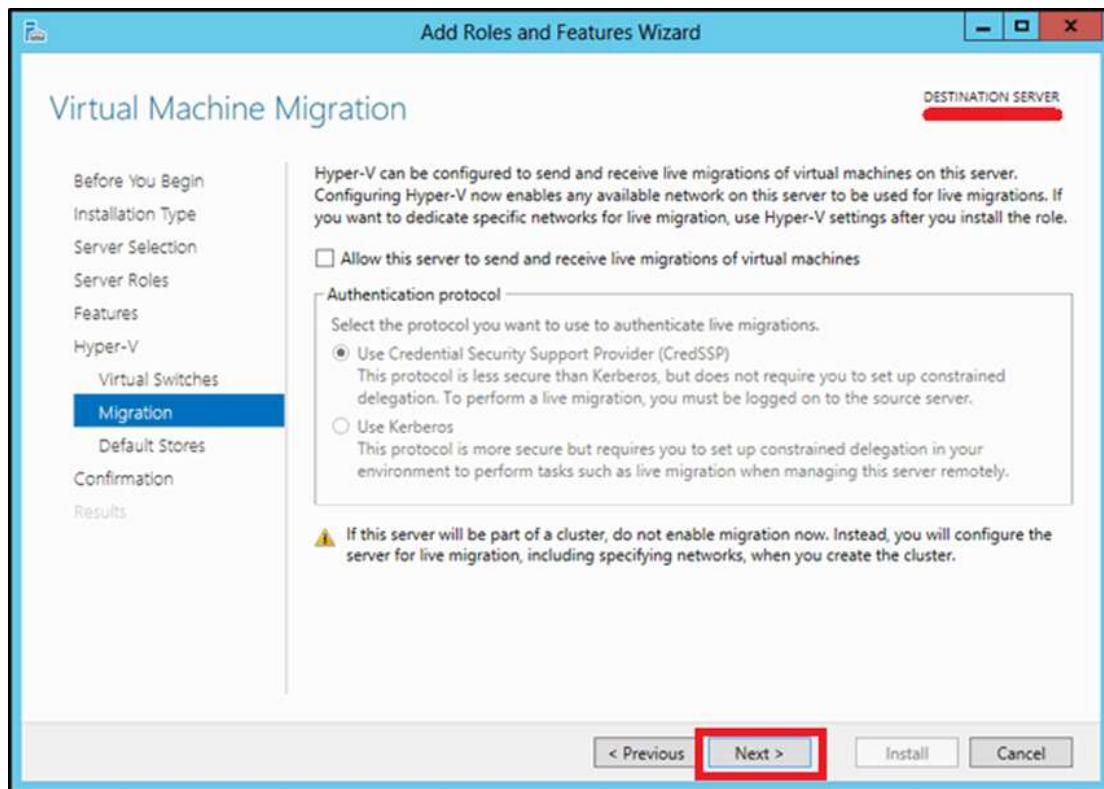
**Step 6:** Click “Next”.



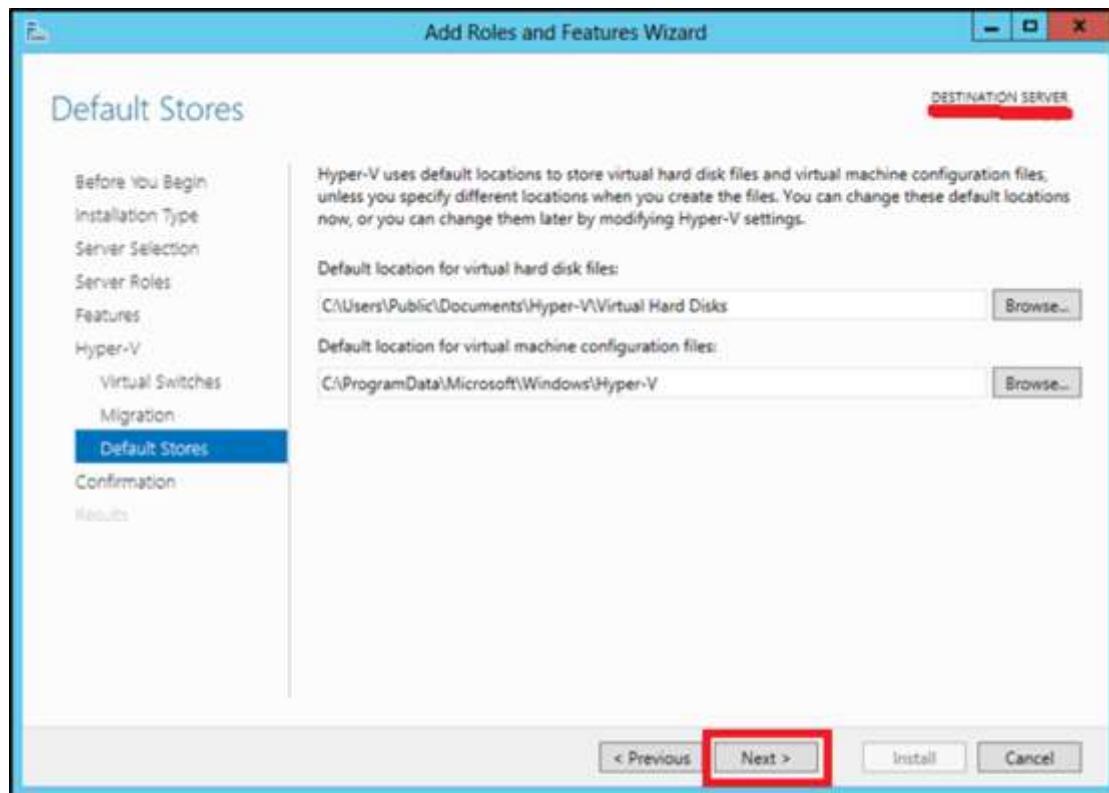
**Step 7:** Choose your server’s physical network adapters that will take part in the virtualization and responsible for network switching → click on “Next”



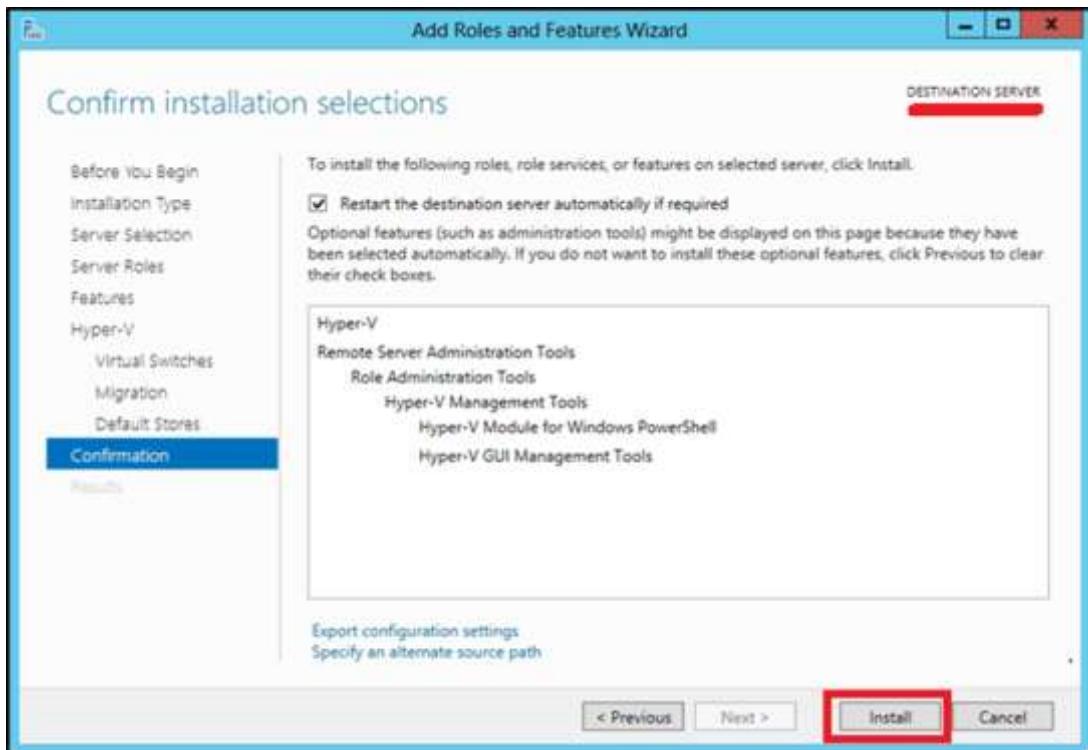
**Step 8:** Under Migration, leave the default settings → click on “Next”.



**Step 9:** Choose the path where you want to save the file → click on “Next”.



**Step 10:** Click “Install” and wait for the installation bar to finish.



## Installing Hyper-V in a windows 10 workstation

To install it in Windows 7, 8, 10 versions, you have to check if your computer supports virtualization. Following are the basic requirements:

- Windows 10 Pro or Enterprise 64-bit Operating System.
- A 64-bit processor with Second Level Address Translation (SLAT).
- 4GB system RAM at minimum.
- BIOS-level Hardware Virtualization support.

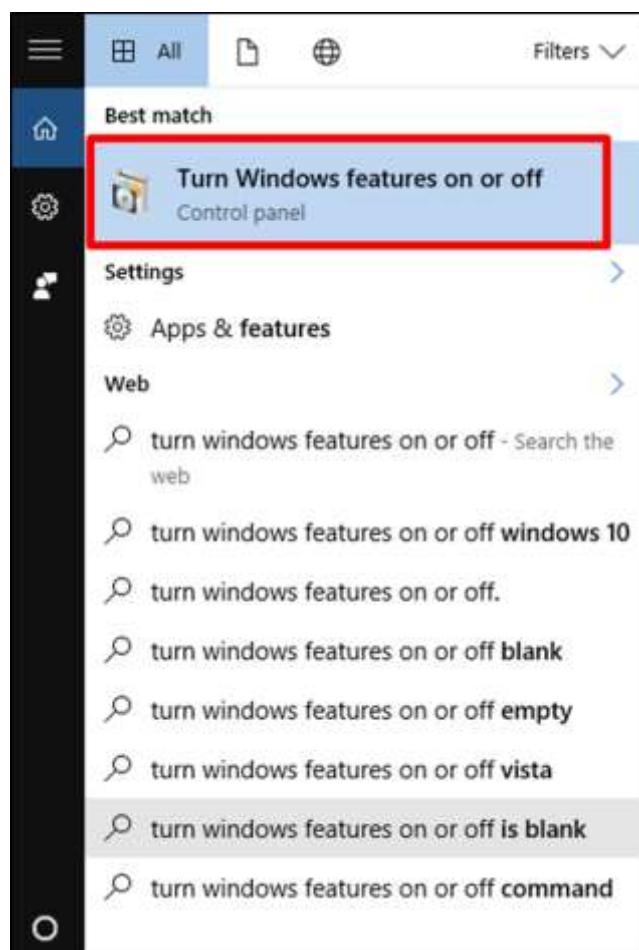
In my case, we have a laptop **HP Probook 450 G3**, which supports it.

Before continuing with the installation, follow the steps given below.

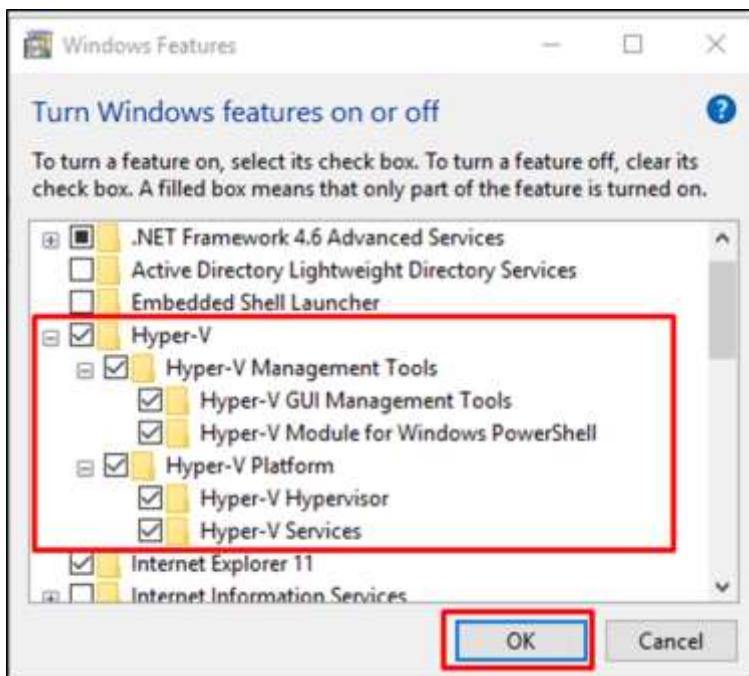
**Step 1:** Ensure that hardware virtualization support is turned on in the BIOS settings as shown below:



**Step 2:** Type in the search bar “turn windows features on or off” and click on that feature as shown below.



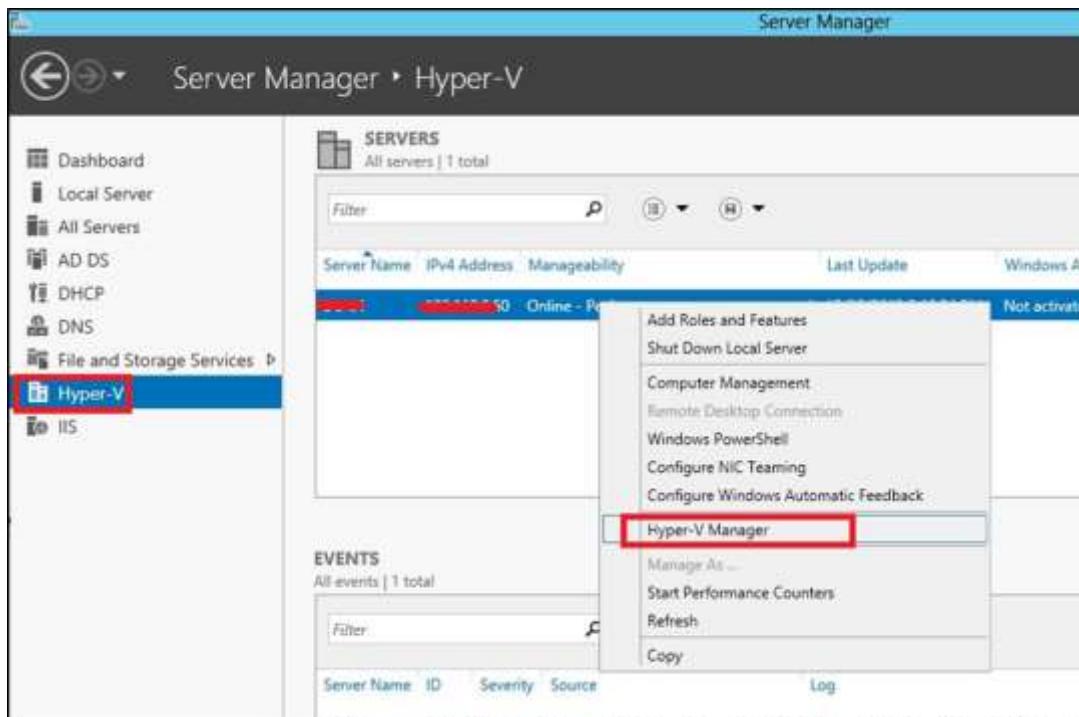
**Step 3:** Select and enable Hyper-V.



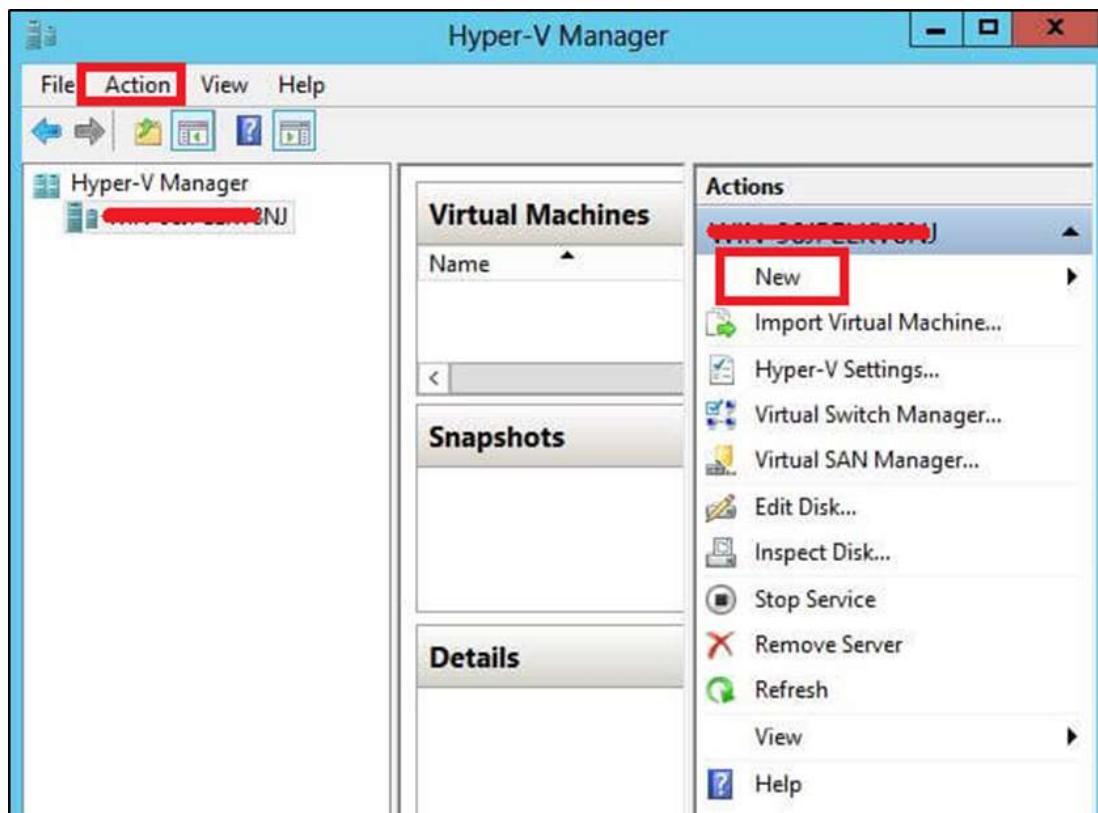
## Creating a Virtual Machine with Hyper-V

In this section, we will learn how to create a virtual machine. To begin with, we have to open the Hyper-V manager and then follow the steps given below.

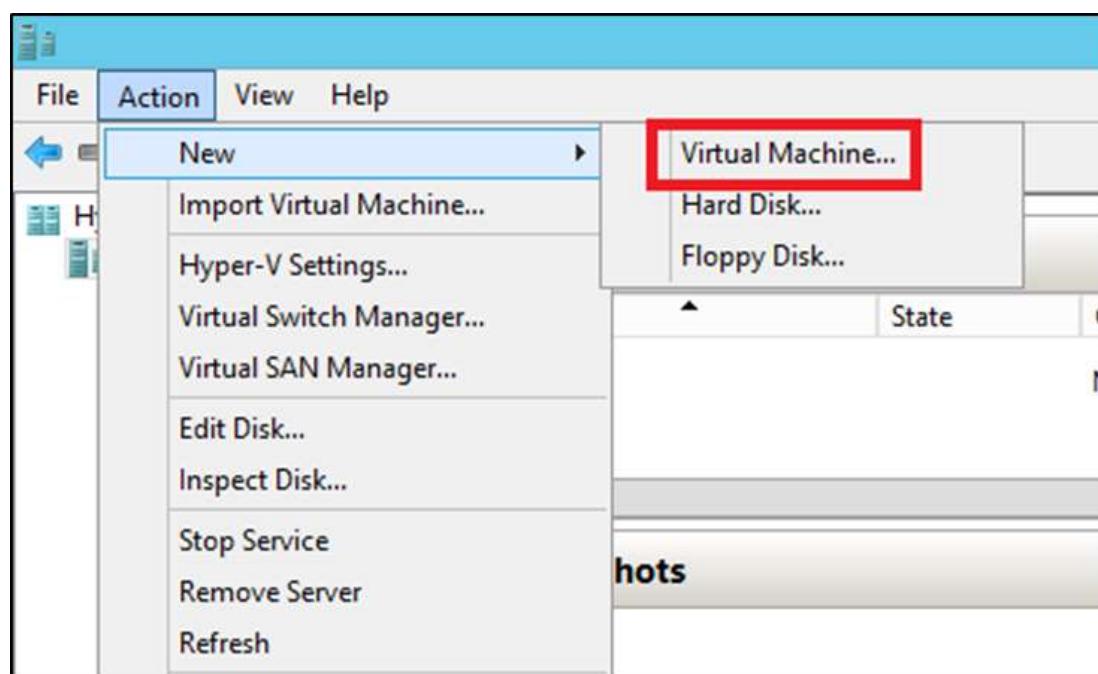
**Step 1:** Go to "Server Manager" → Click on "Hyper-V Manager".



**Step 2:** Click "New" on the left Panel or on the "Actions" button.



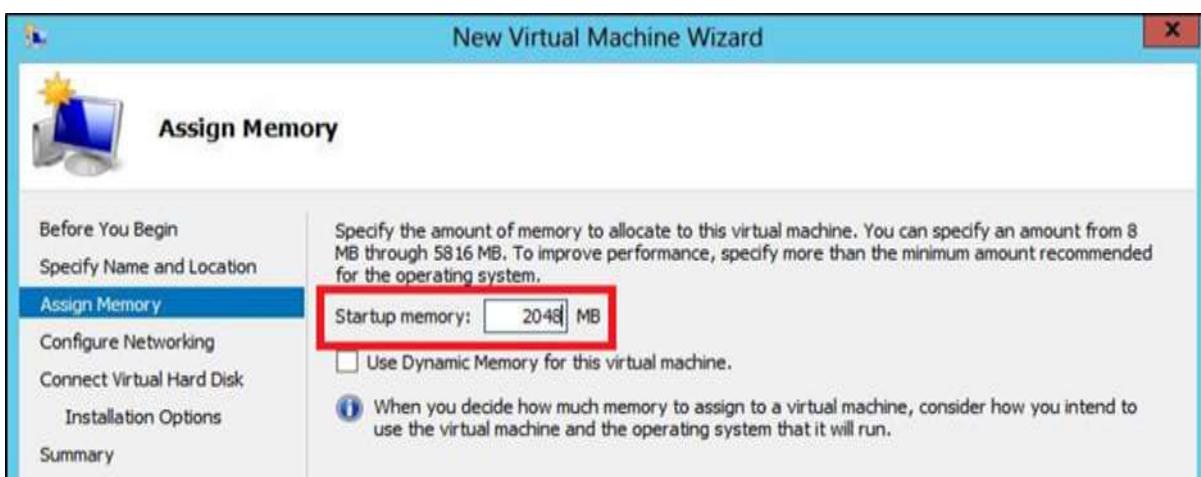
**Step 3:** Double-click on "Virtual Machine..."



**Step 4:** A new table will open → Type Name of your new machine → click “Next”.



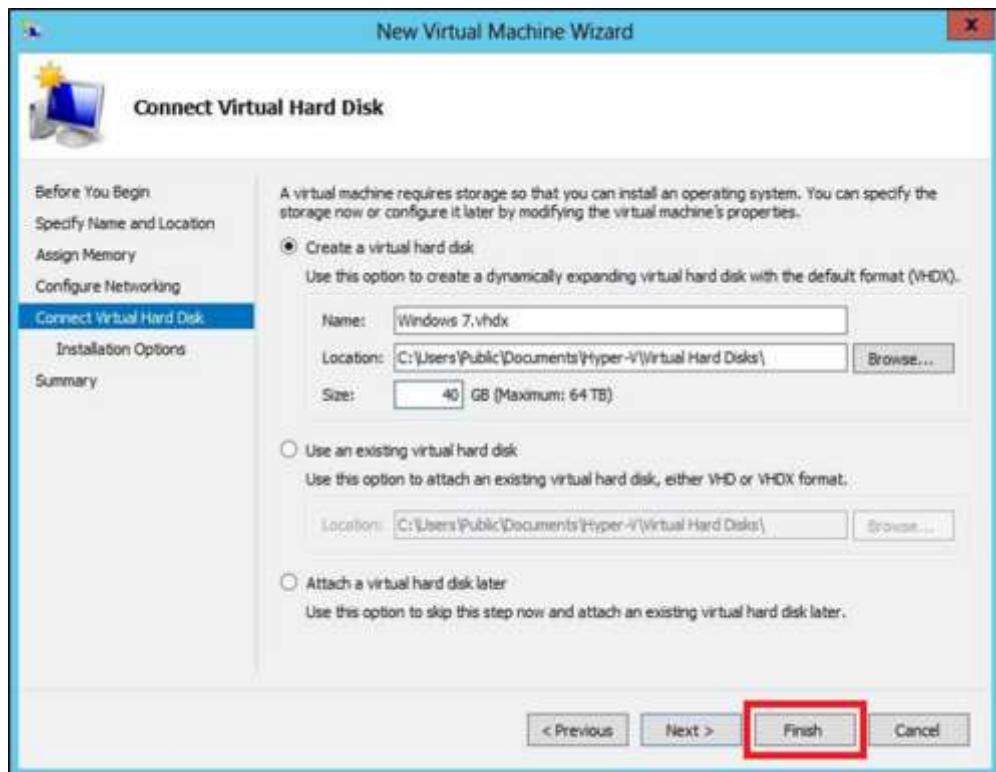
**Step 5:** A new table will be opened where you have to allocate the memory. Keep in mind you cannot choose more memory than you have physically.



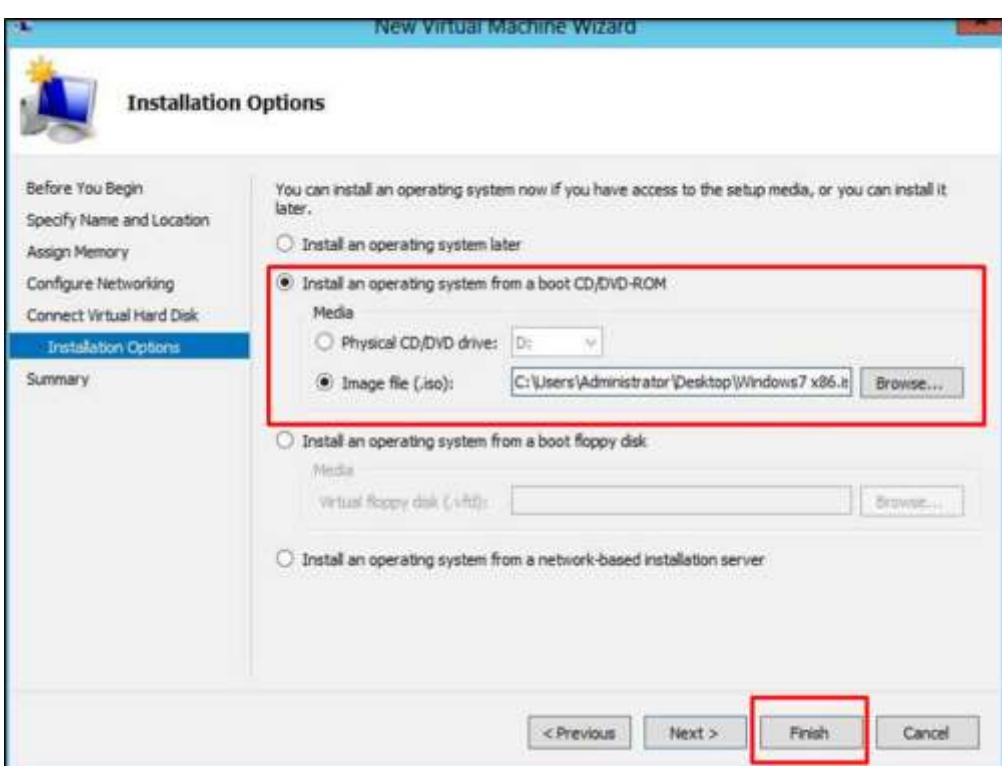
**Step 6:** In the “Connection” drop down box, choose your physical network adaptor → click on “Next”.



**Step 7:** Now it is time to create a Virtual Hard disk, if you already have one, choose the second option.



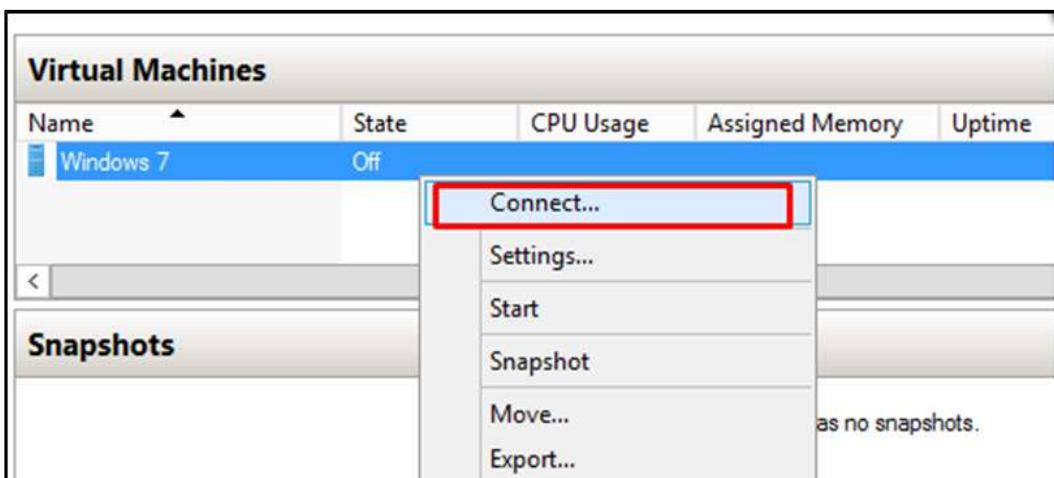
**Step 8:** Select the Image of ISO that has to be installed → click on "Finish".



**Step 9:** After clicking on finish, you would get the following message as shown in the screenshot below.



**Step 10:** To connect to the Virtual machine, Right Click on the created machine → click on "Connect..."



**Step 11:** After that, installation of your ISO will continue.

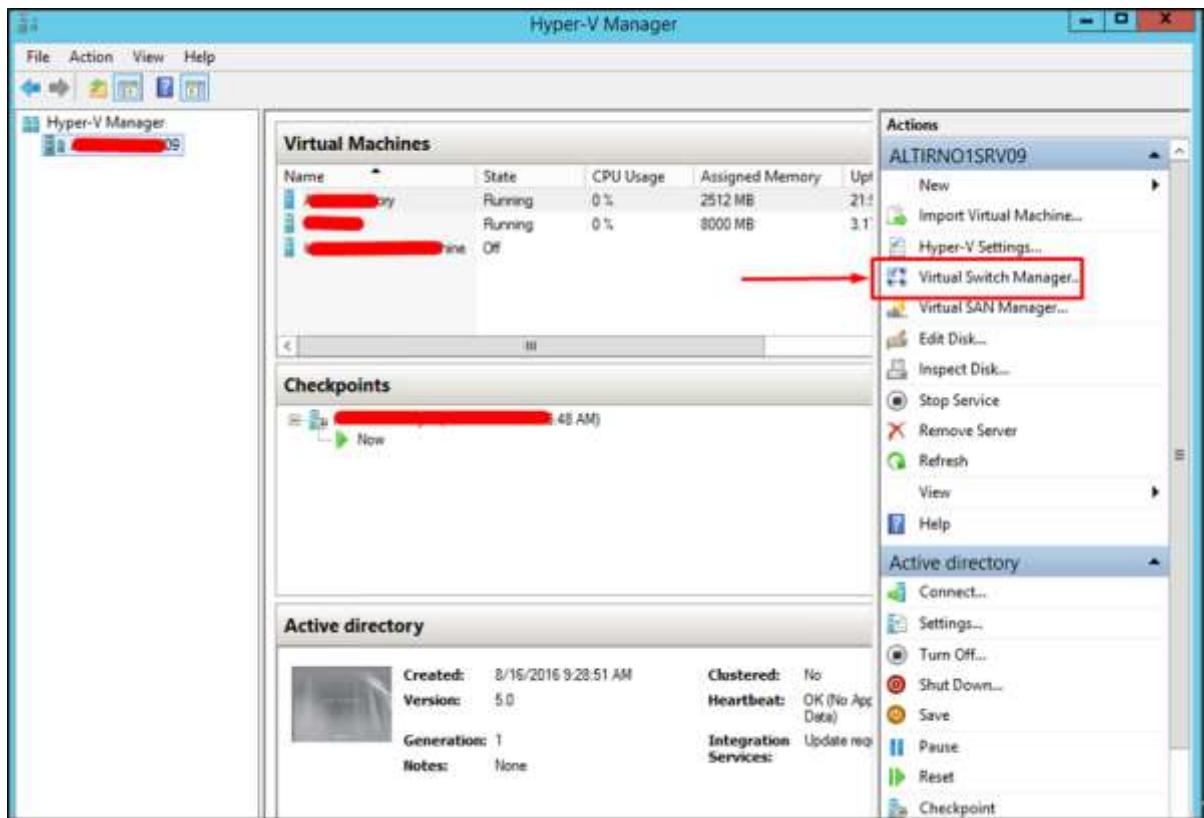


## Setting up Networking with Hyper-V

The Hyper-V vSwitch is a software, layer-2 Ethernet network-traffic switch. It allows administrators to connect VMs to either physical or virtual networks. It is available by default within the Hyper-V Manager installation and contains extended capabilities for security and resource tracking.

If you attempt to create a VM right after the set-up process, you will not be able to connect it to a network.

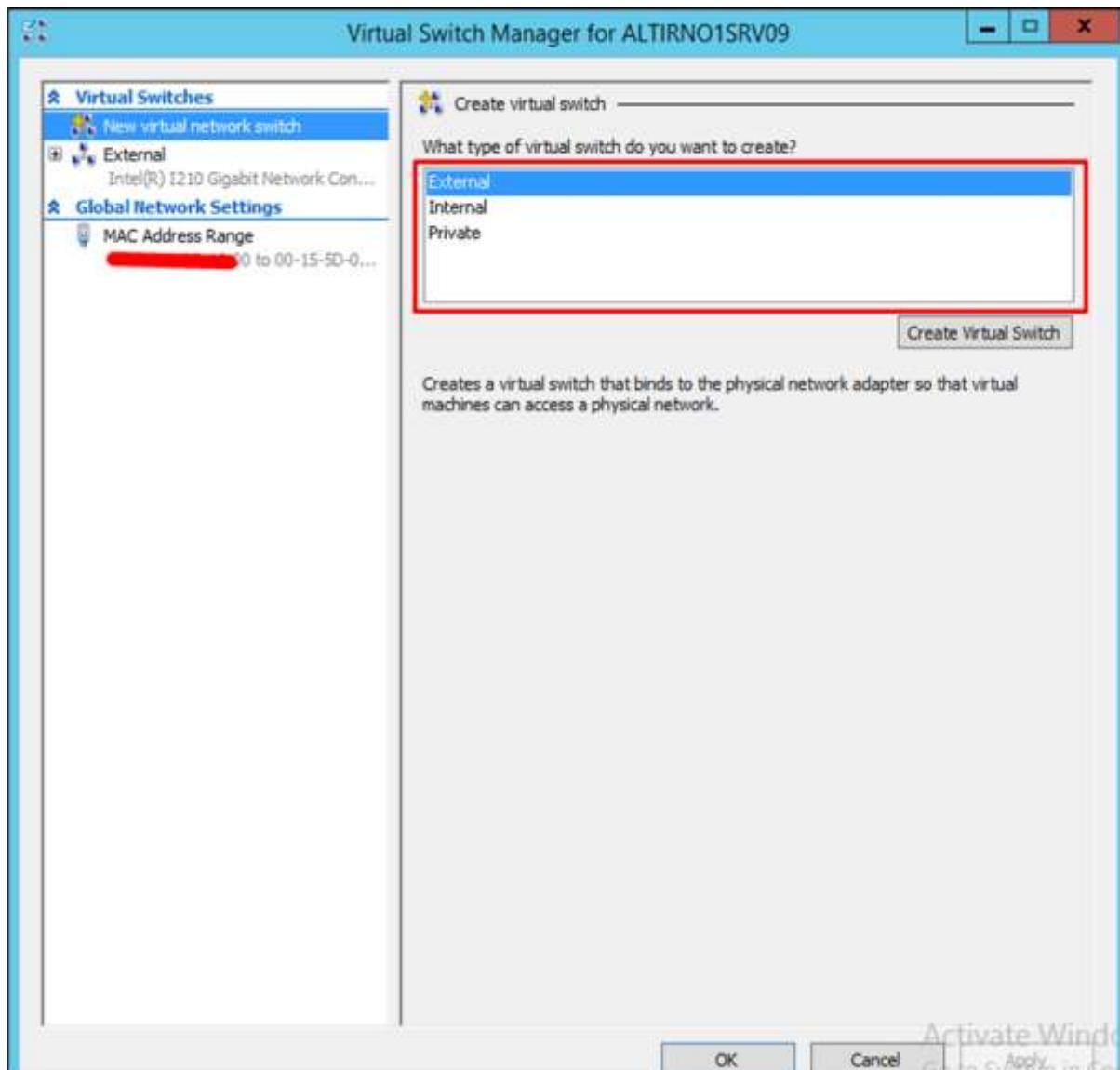
To set up a network environment, you will need to select the **Virtual Switch Manager** in the right hand side panel of Hyper-V Manager as shown in the screenshot below.



The Virtual Switch Manager helps configure the vSwitch and the Global Network Settings, which simply lets you change the default 'MAC Address Range', if you see any reason for that.

Creation of the virtual switch is easy and there are three vSwitch types available, which are described below:

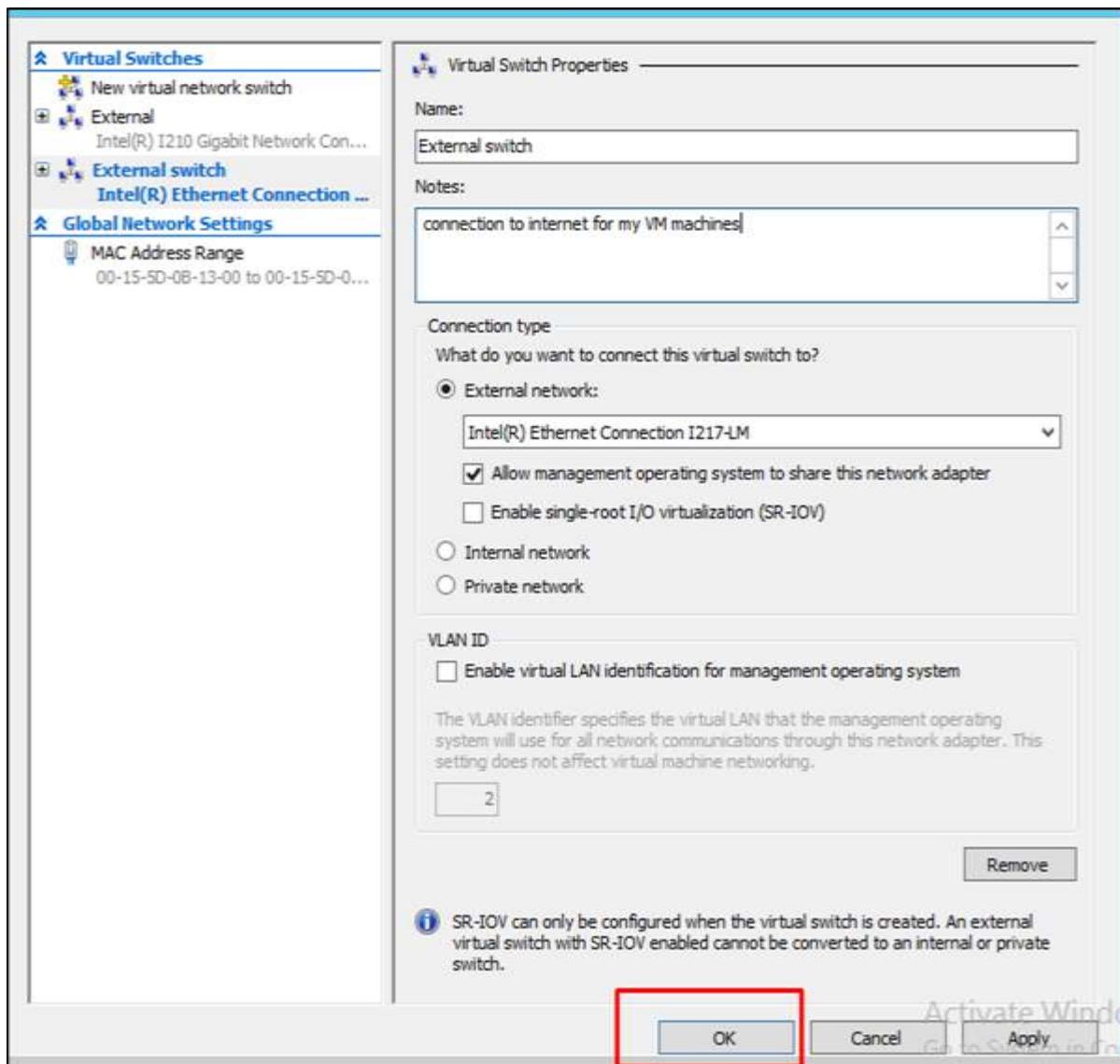
- **External vSwitch** will link a physical NIC of the Hyper-V host with a virtual one and then give your VMs access outside of the host. This means that your physical network and internet (if your physical network is connected to internet).
- **Internal vSwitch** should be used for building an independent virtual network, when you need to connect VMs to each other and to a hypervisor as well.
- **Private vSwitch** will create a virtual network where all connected VMs will see each other, but not the Hyper-V host. This will completely isolate the VMs in that sandbox.



Here, we have selected “External” and then “Create Virtual Switch”. The table with the setting of the vSwitch will be open where we will fill the fields as shown below:

- **Name:** is the name that we will put to identify the vSwitch.
- **Notes:** is the description for us, generally, we put friendly descriptions to be understood.
- **Connection Type:** is external as explained earlier and selects a physical network card on my server.

Once all this is entered, Click on “OK”.

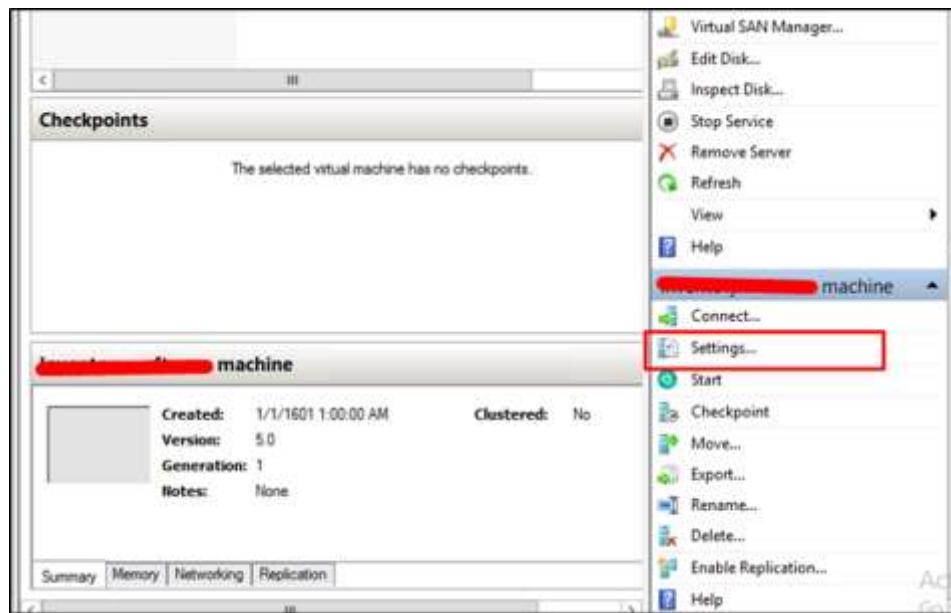


## Allocating Processors & Memory to a VM using Hyper-V

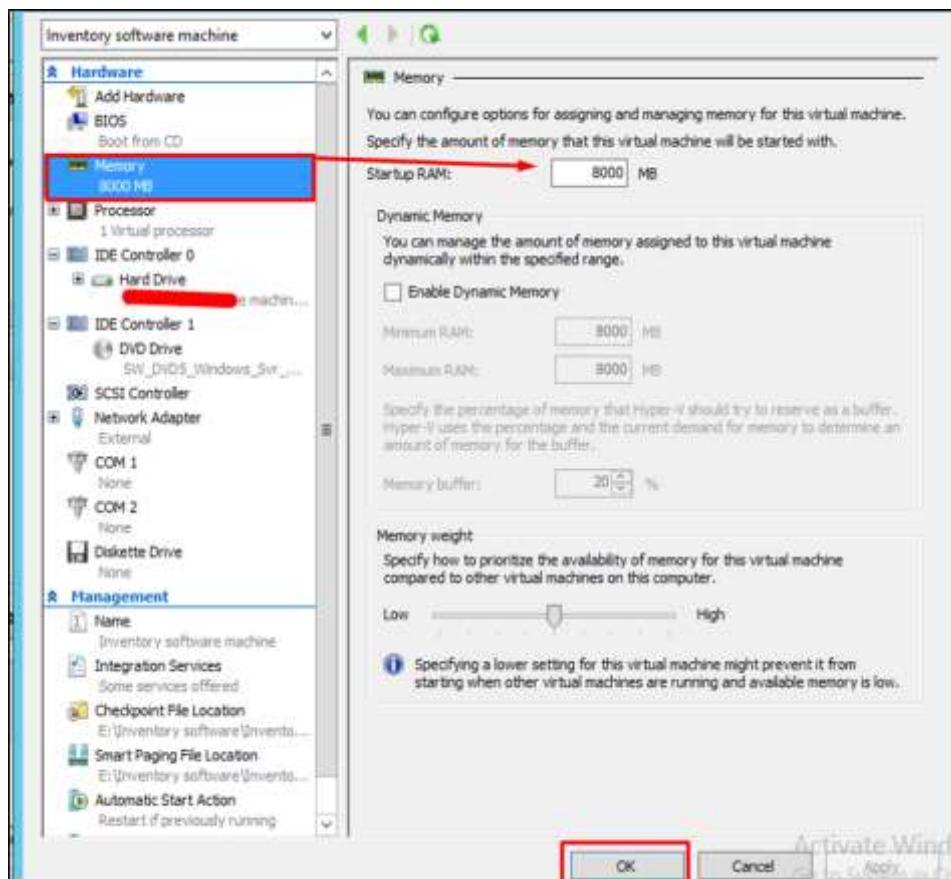
In this section, we will see the task of allocating CPU, Memory and Disk Resources to the virtual machines that are running on a server. The key to allocating CPU or any other type of resource in Hyper-V is to remember that everything is relative.

For example, Microsoft has released some guidelines for Virtualizing Exchange Server. One of the things that was listed was that the overall system requirements for Exchange Server are identical whether Exchange is being run on a virtual machine or on a dedicated server.

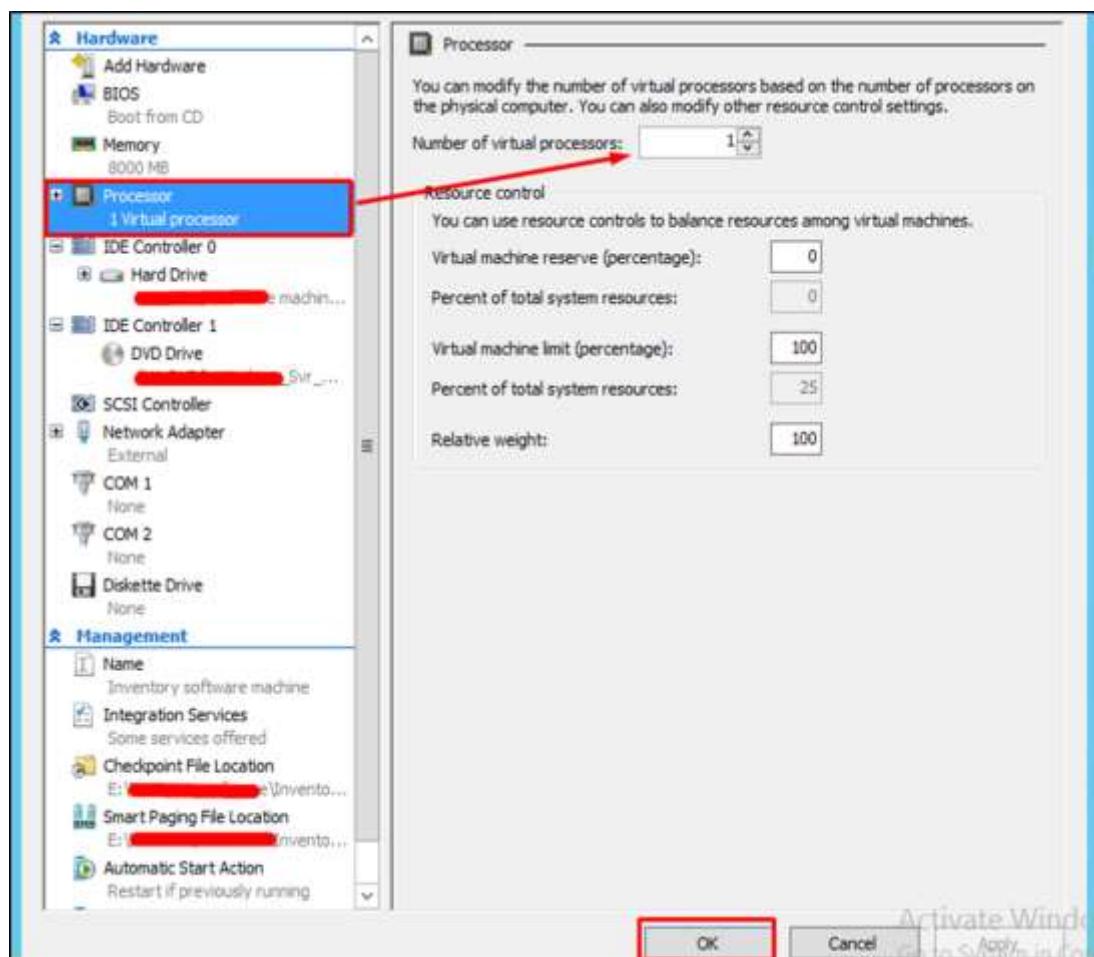
To allocate one of the features mentioned above, we need to click on the “Settings...” tab in the right hand side panel.



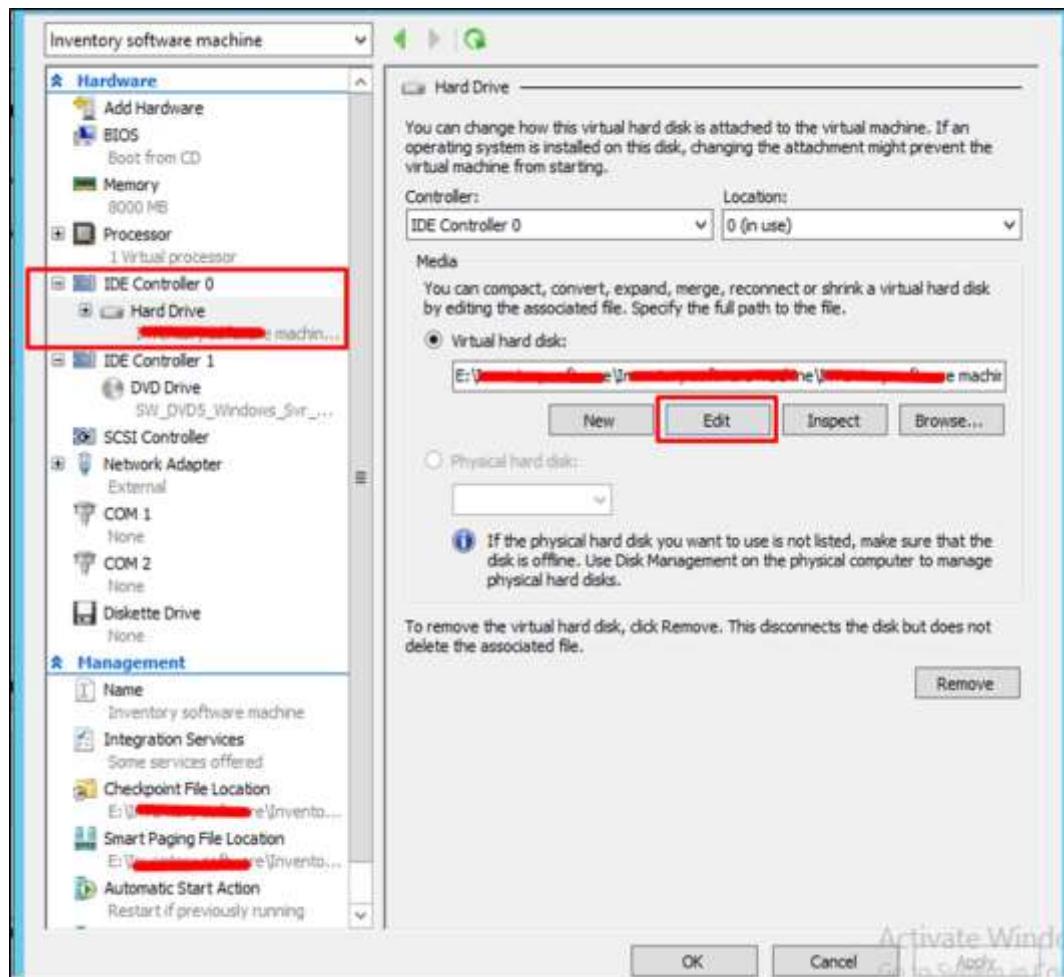
To allocate more memory to the selected virtual machine, click on the “Memory” tab on the left hand side of the screen. You will also have “Startup RAM”, where you can allocate as much ram as you have physically to a VM machine → Click on “Ok”.



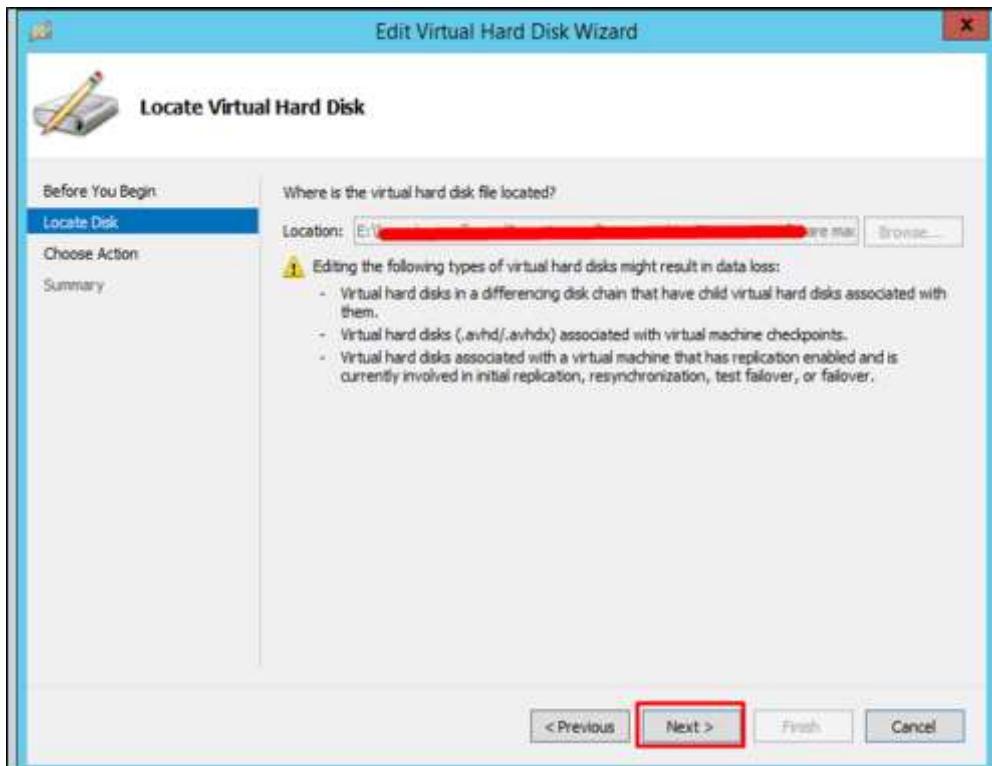
To allocate more processors, click on the "Processor" tab on the left hand side of the panel. Then you can enter the number of virtual processors for your machine.



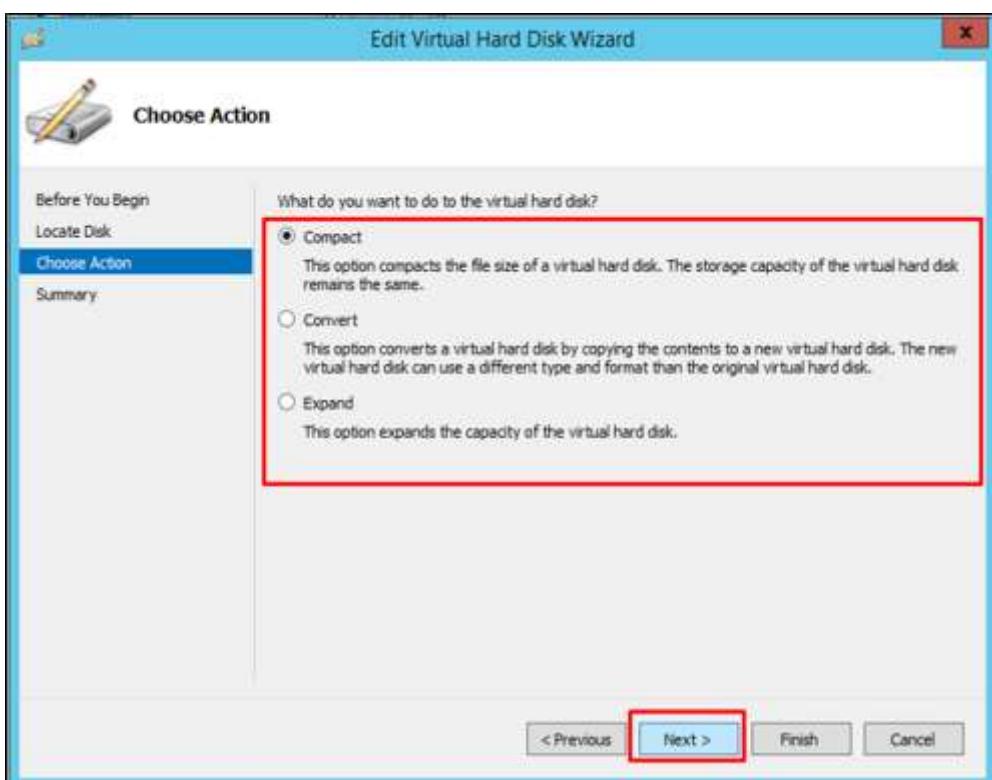
If you need to expand, compress the capacity of the virtual hard disk. Click on the "IDE controller 0" on the left hand side panel → click on "Edit".



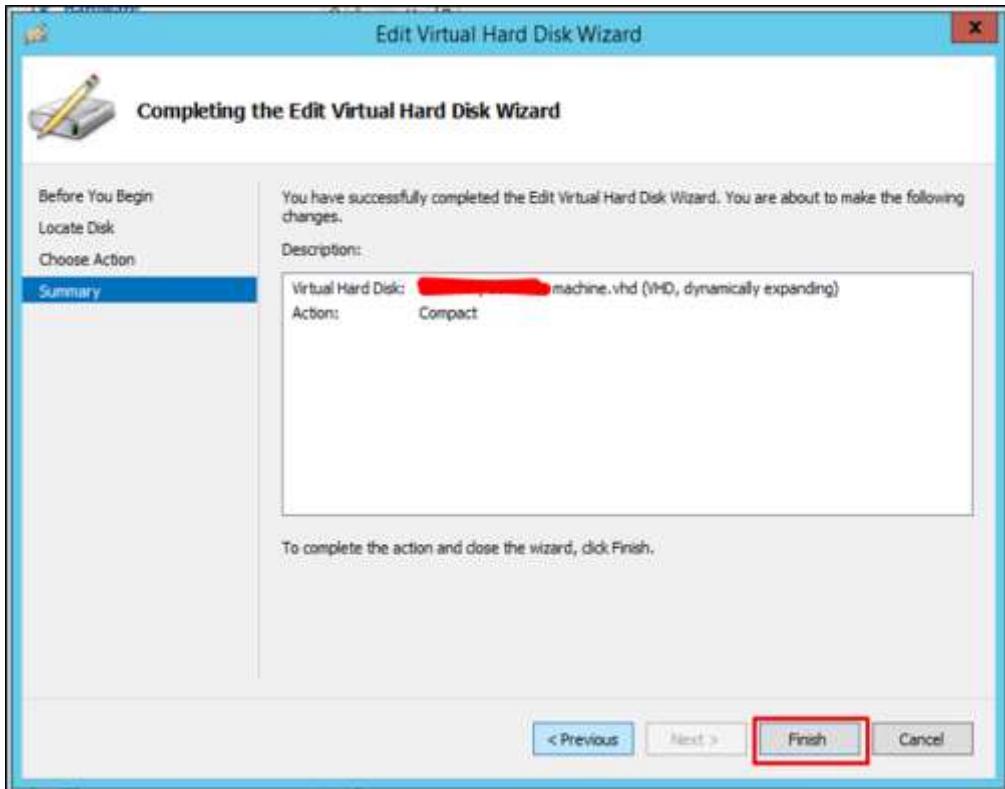
Once all the above changes are done, Click on "Next"



Select one of the options based on your need (all of them have their respective descriptions) and then click on "Next".



Click on “Finish” and wait for the process to finish.

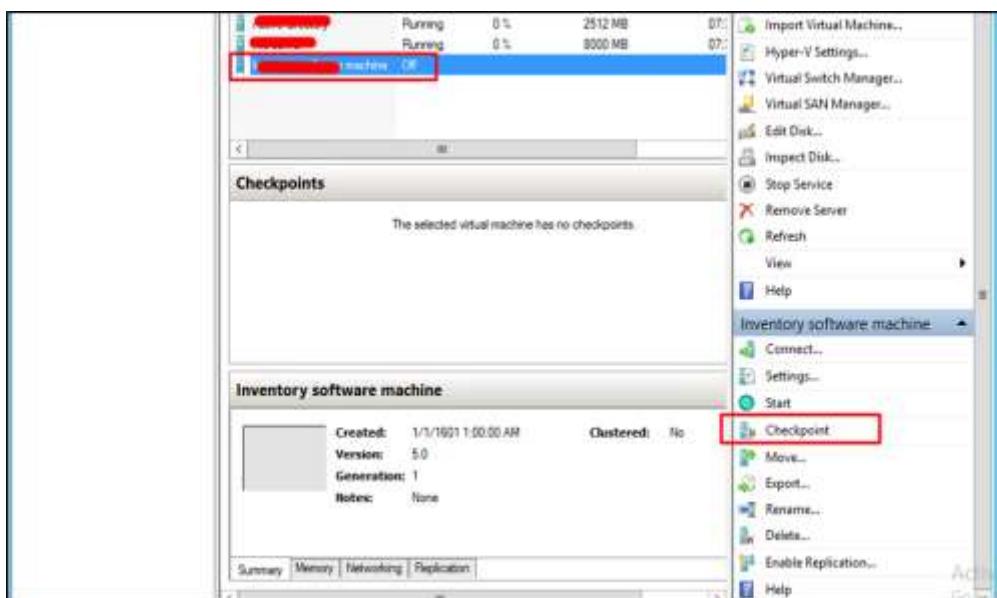


## Using Checkpoints in Hyper-V

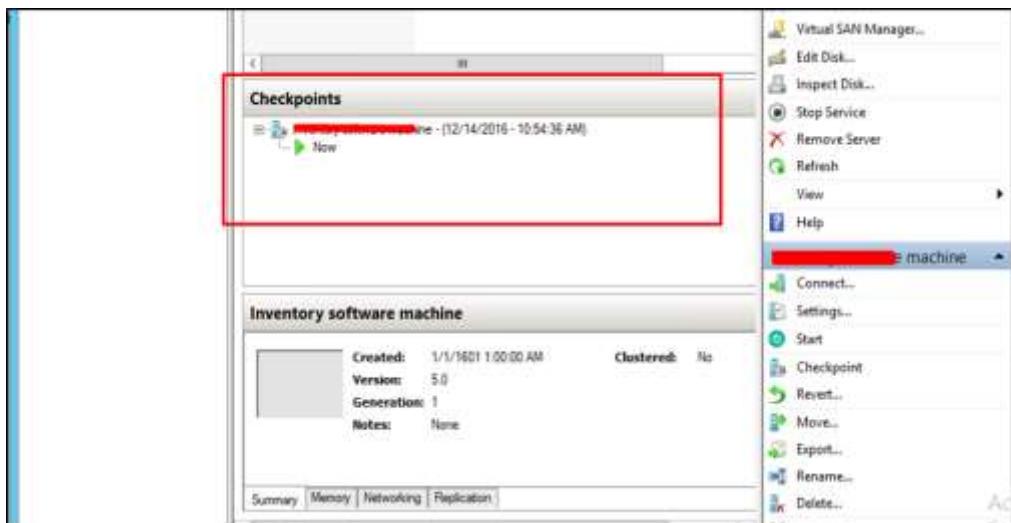
Checkpoints in Hyper-V are called snapshots and they help us to revert the machine in a desired state that we had in the past.

To create a checkpoint we should follow the steps given below.

**Step 1:** Select the VM by clicking on it. On the left hand side panel, click on “Checkpoint”.



**Step 2:** The following checkpoint will be created with the respective date and time in the main Hyper-V manager console.



# 6. Virtualization – VMware Workstation Player

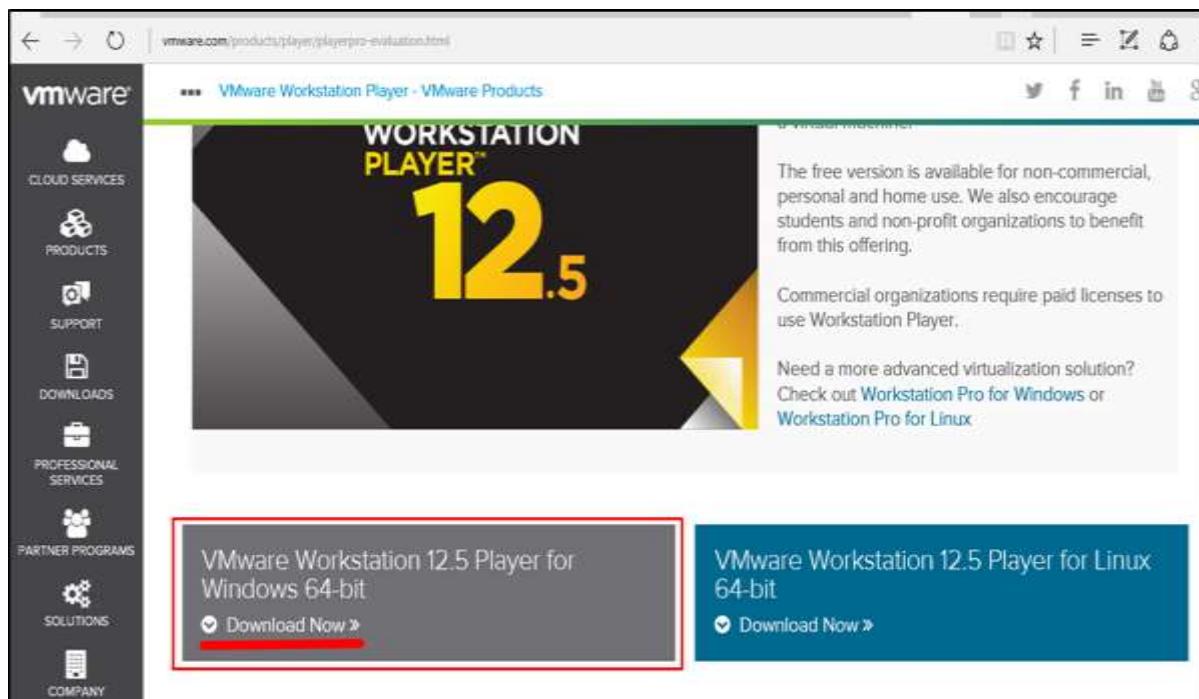
In this chapter, we will understand how to install the VMware Workstation Player and its usages.

## Installing VMware Workstation Player

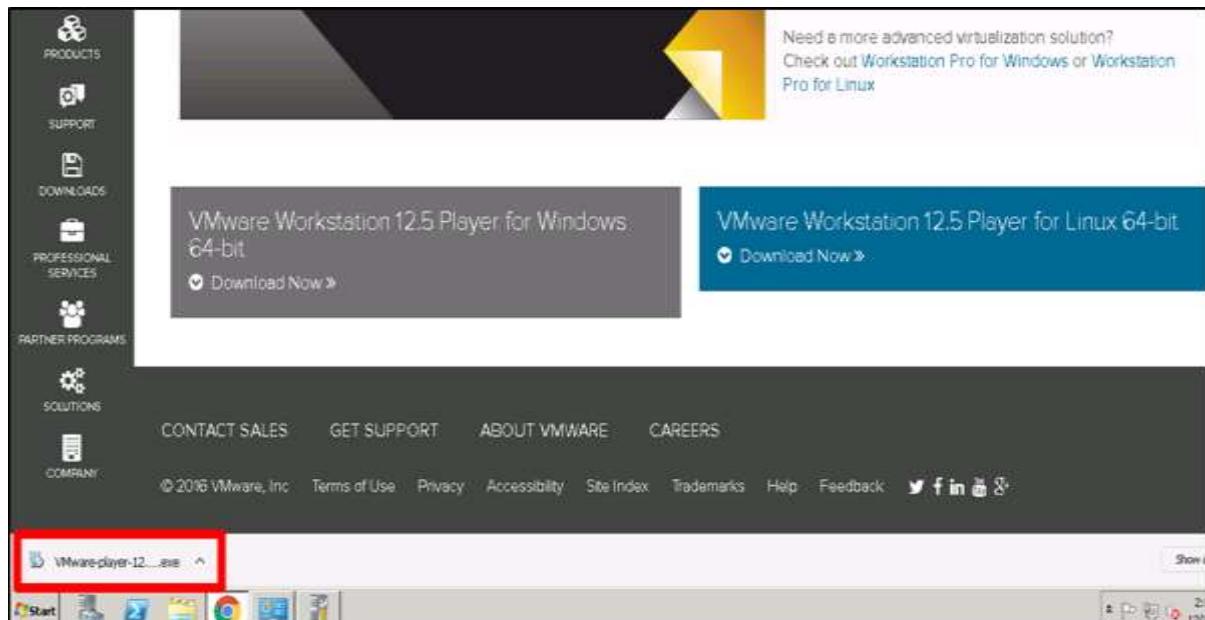
VMware workstation player is a “hosted hypervisor”, so you have to have a pre-installed OS before continuing to install it.

VMware workstation player is free version and available for non-commercial, personal and home use. They also encourage students and non-profit organizations to benefit from this offering. To download the VMware workstation player, you can click on the following link – <http://www.vmware.com/products/player/playerpro-evaluation.html>. To install the VMware workstation player, follow the steps given below.

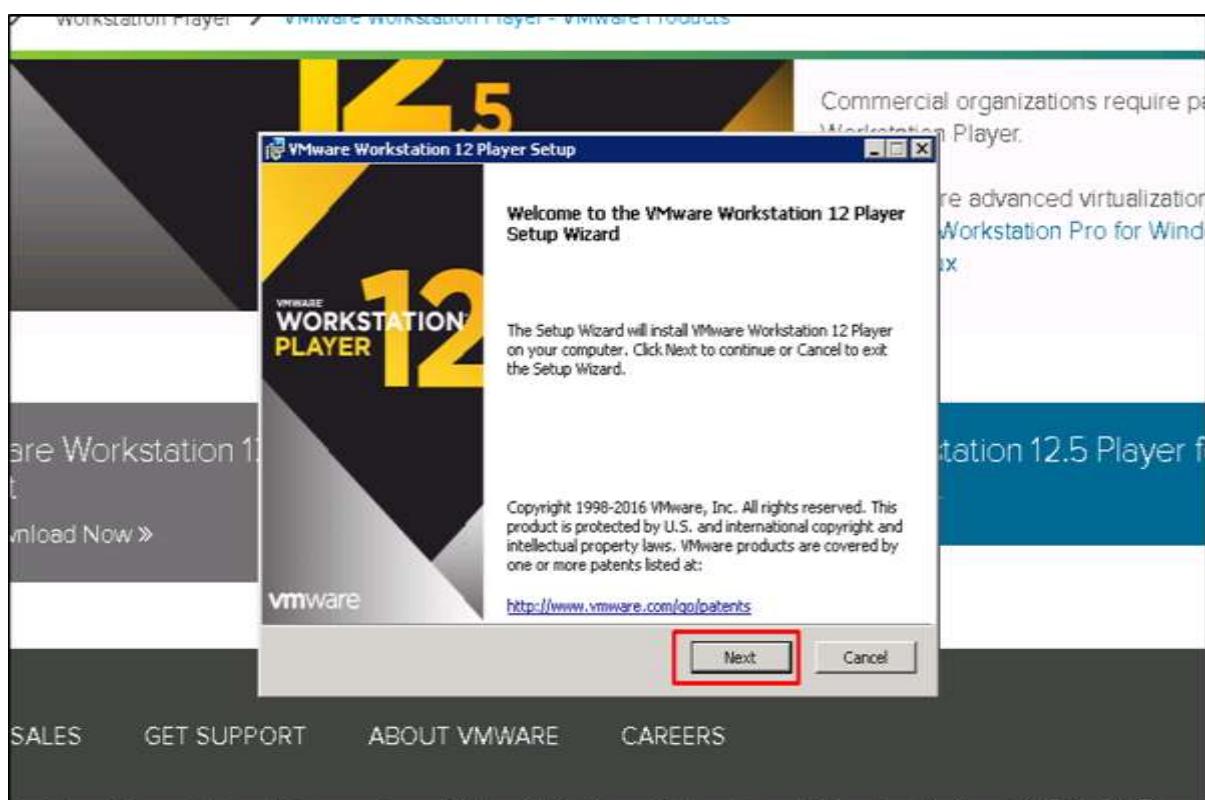
**Step 1:** Click on “Download Now” as shown in the screenshot below.



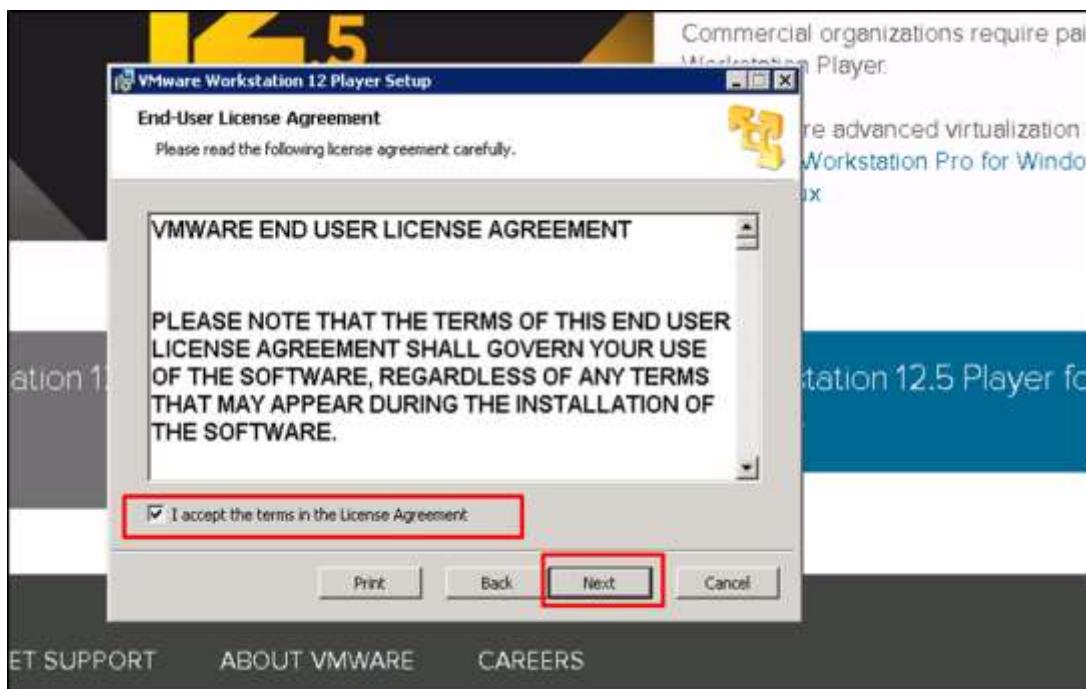
**Step 2:** You will see that a file has been downloaded → double click on it.



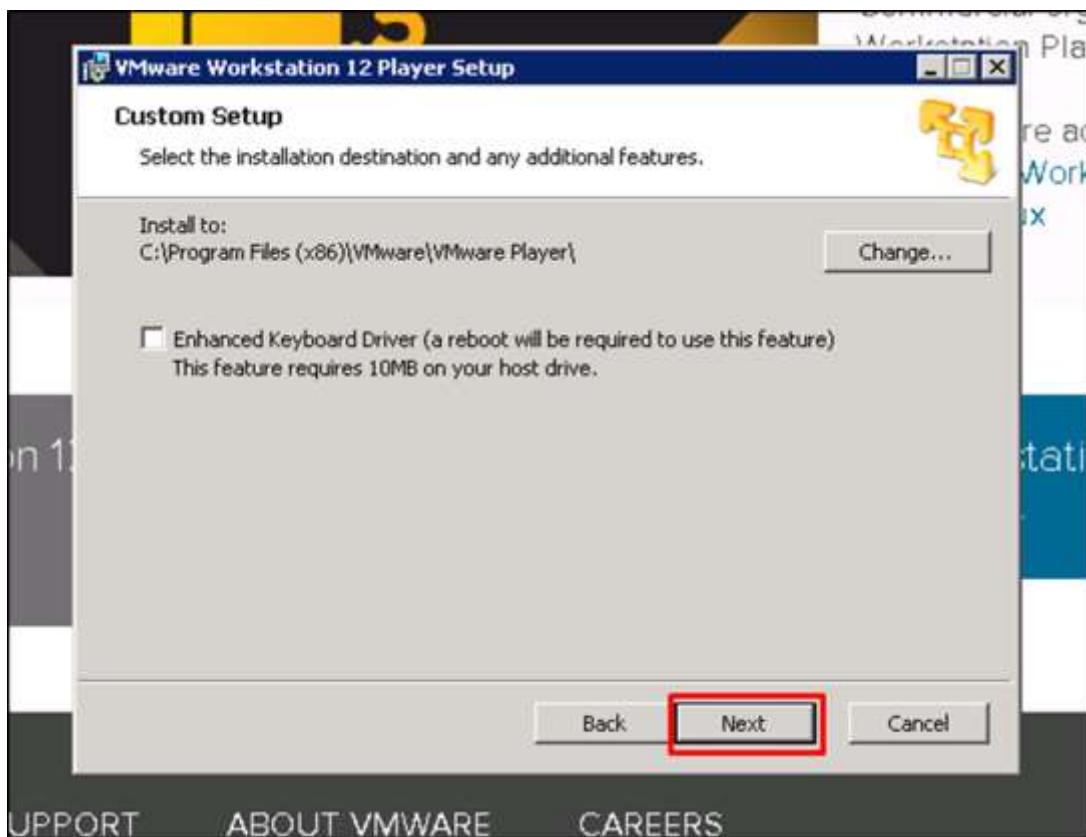
**Step 3:** A Table will pop-up initializing the installation of VMware -> Click "Next"



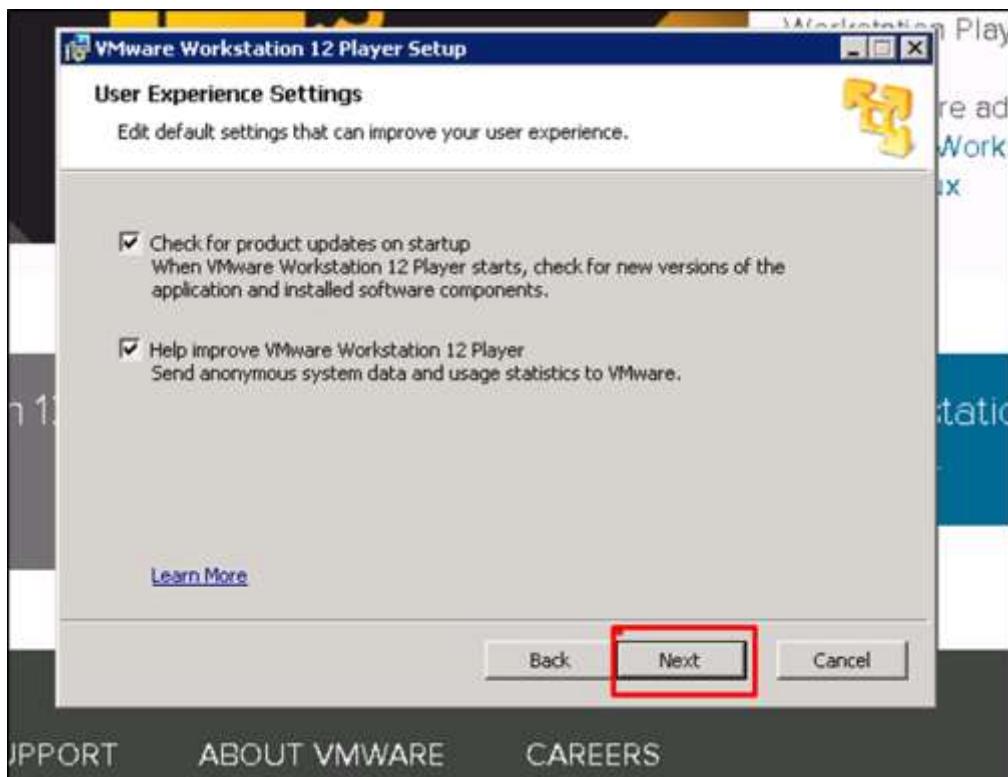
**Step 4:** Check the box "I accept the terms in the license agreement" → Click on "Next".



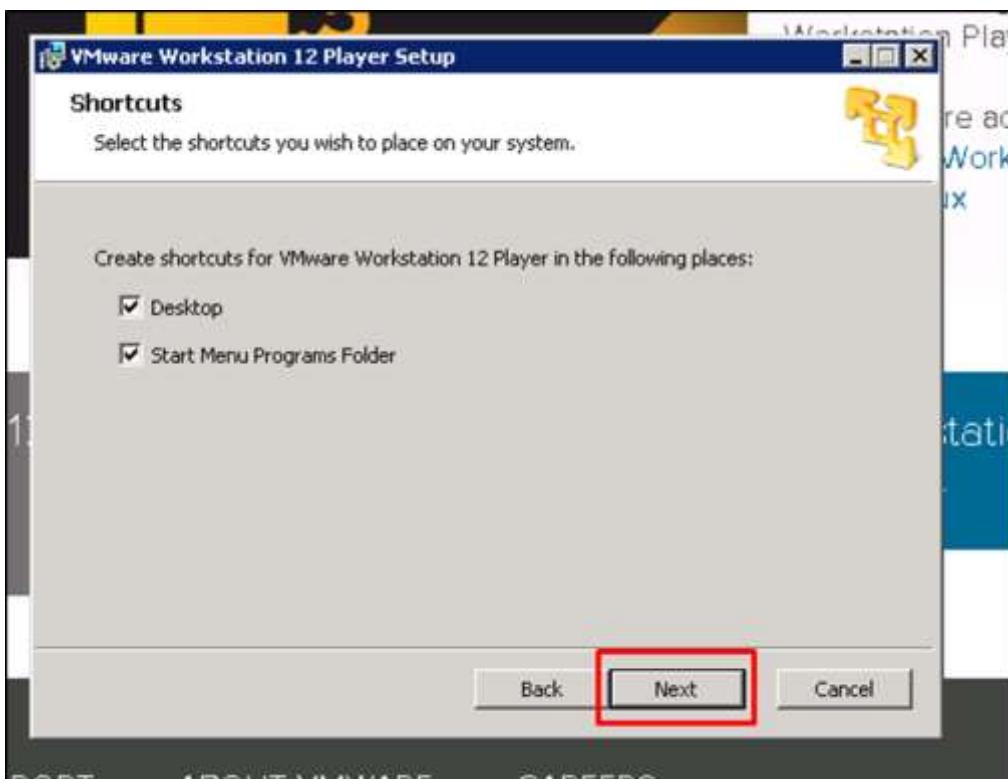
**Step 5:** Once again, click on the "Next" button.



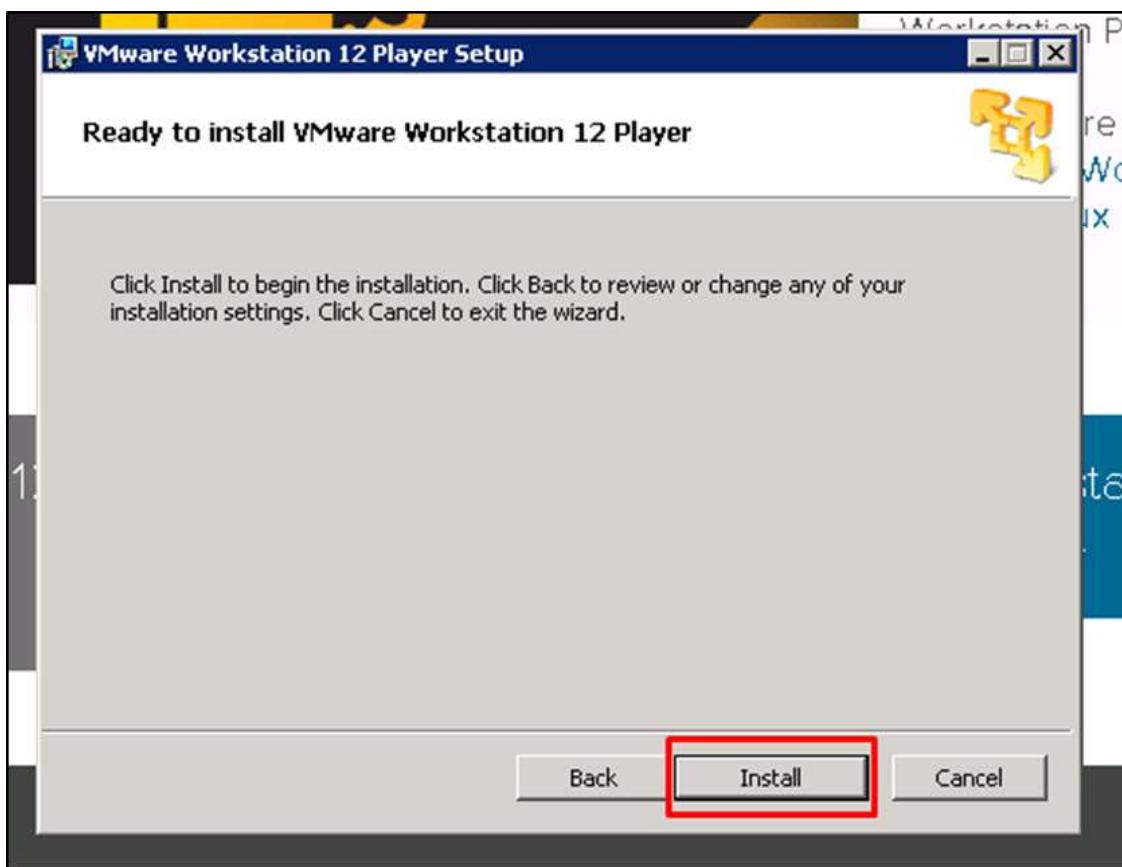
**Step 6:** Leave the default values and click on "Next".



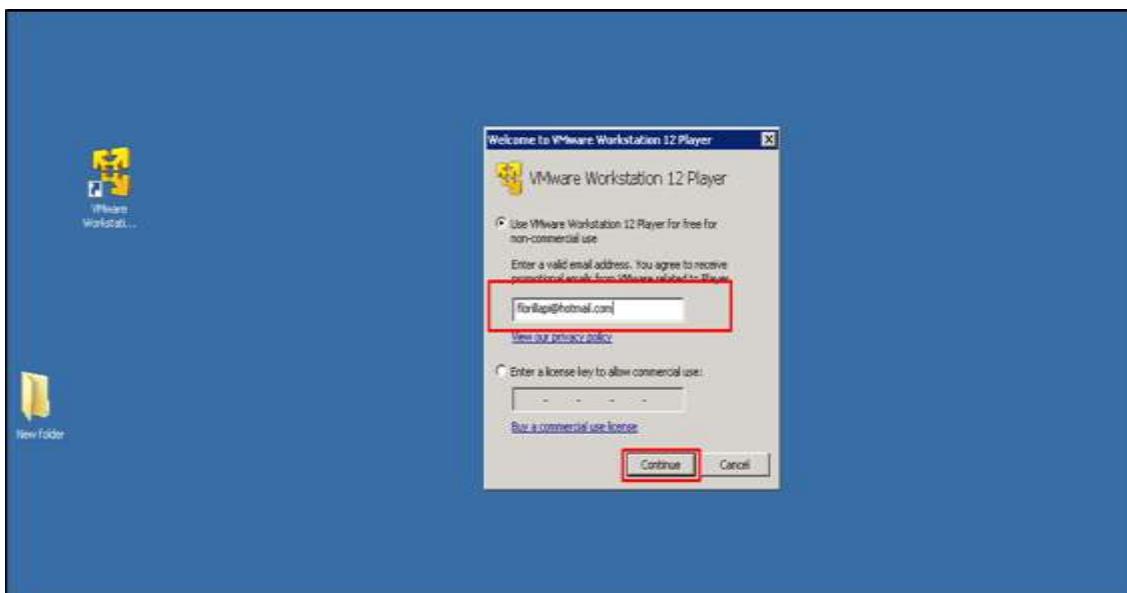
**Step 7:** Once again, click on "Next".



**Step 8:** Click on "Install".



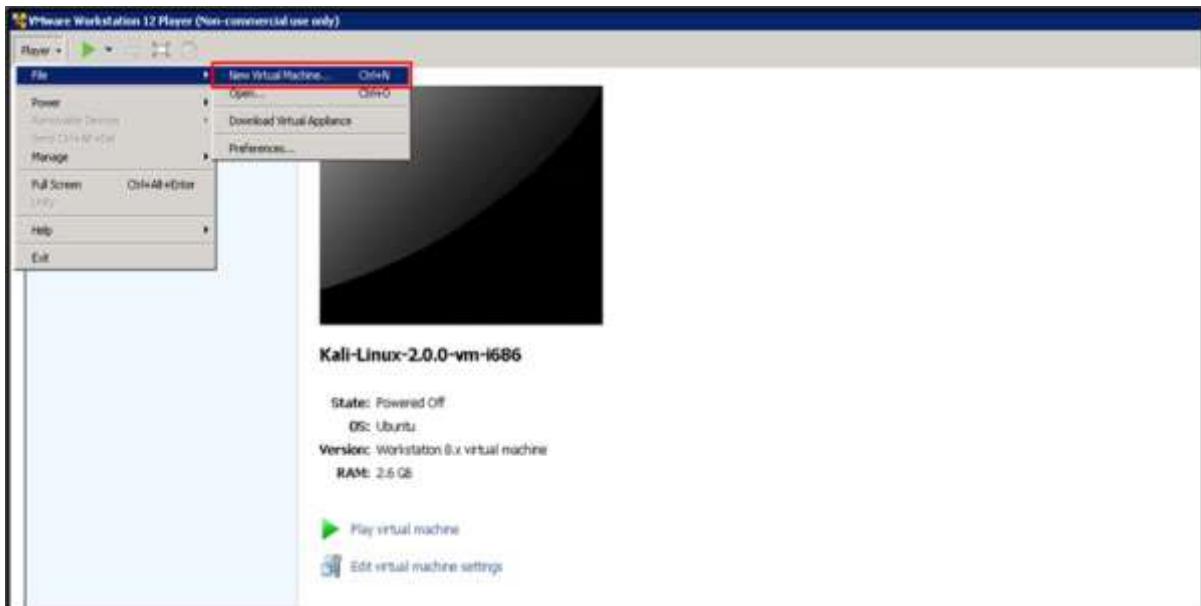
**Step 9:** An icon will be created on the desktop. Click on it and a table will pop-up, where you have two possibilities: If you want to use it as a non-commercial version, just enter your email address. If you want to use it as a commercial version, check the second option and enter your serial key.



## Creating a VM with VMware Workstation

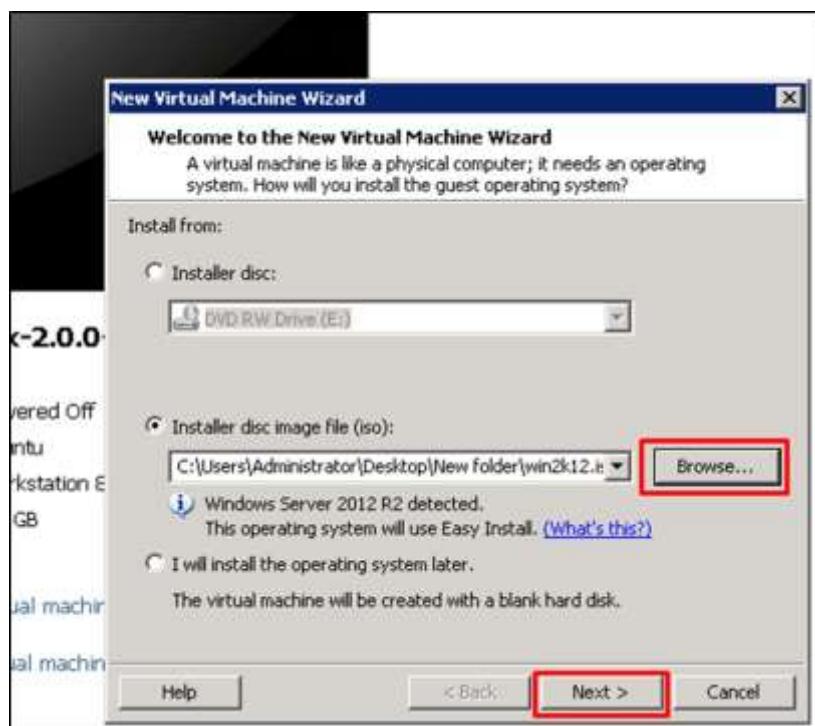
To create a virtual machine, we have to follow the steps given below.

**Step 1:** Click on “Player” → File → New Virtual Machine.



**Step 2:** A table will pop-up requesting you to find a “Boot disk”, “Boot Image” or to install OS at a later stage.

We will choose the second option and click on “Browse”. Then we have to click on the ISO image, which we want to install. Once all this is done, click on “Next”.



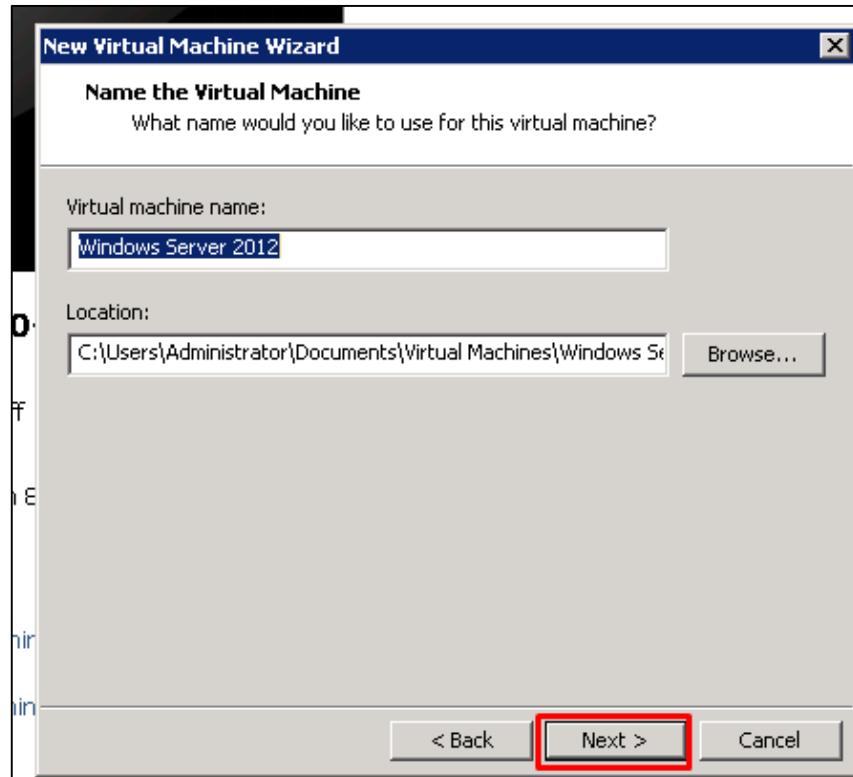
**Step 3:** As I am installing windows server 2012, it will pop-up a table requesting to enter the serial key → click directly on "Next", if you want to activate the non-commercial version for Windows.



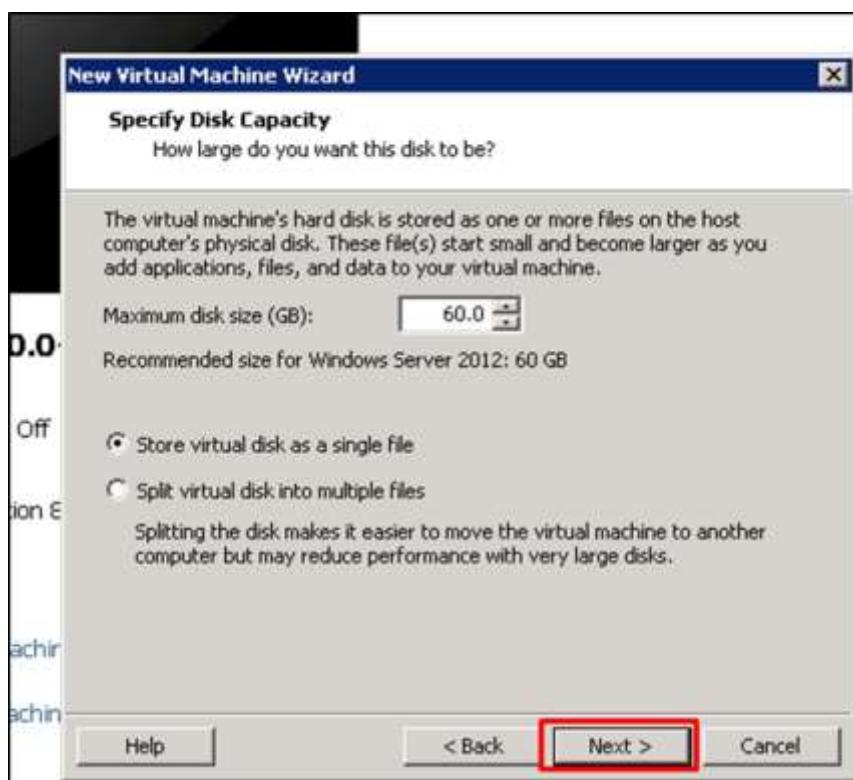
**Step 4:** After the above step is complete, a dialogue box opens. Click "Yes".



**Step 5:** Click "Next".



**Step 6:** In the "Maximum size disk" box, enter the value of your virtual Hard disk, which in our case is 60GB. Then click on "Next".



**Step 7:** Click on "Finish".

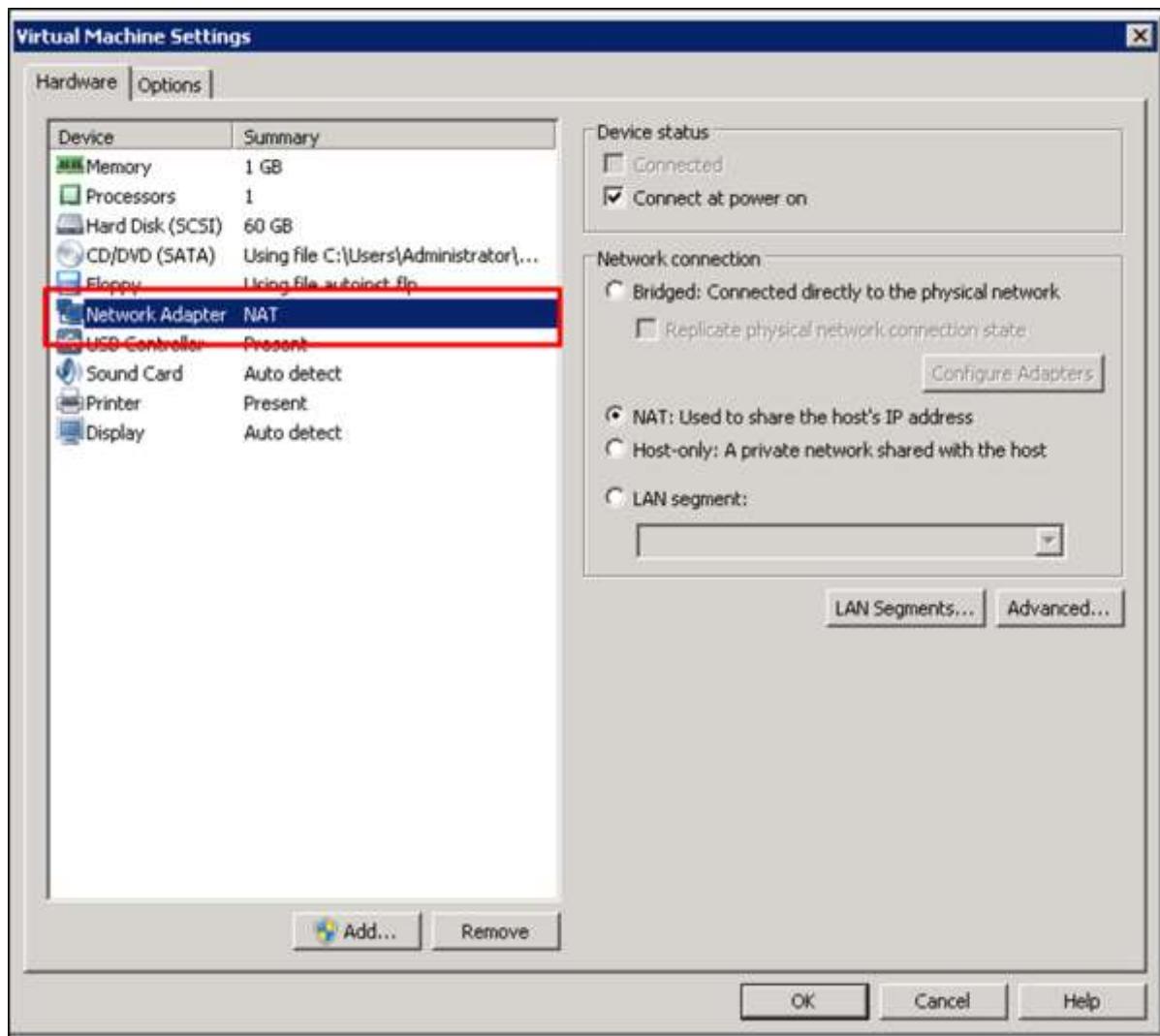


## Setting up Networking with VMware Workstation

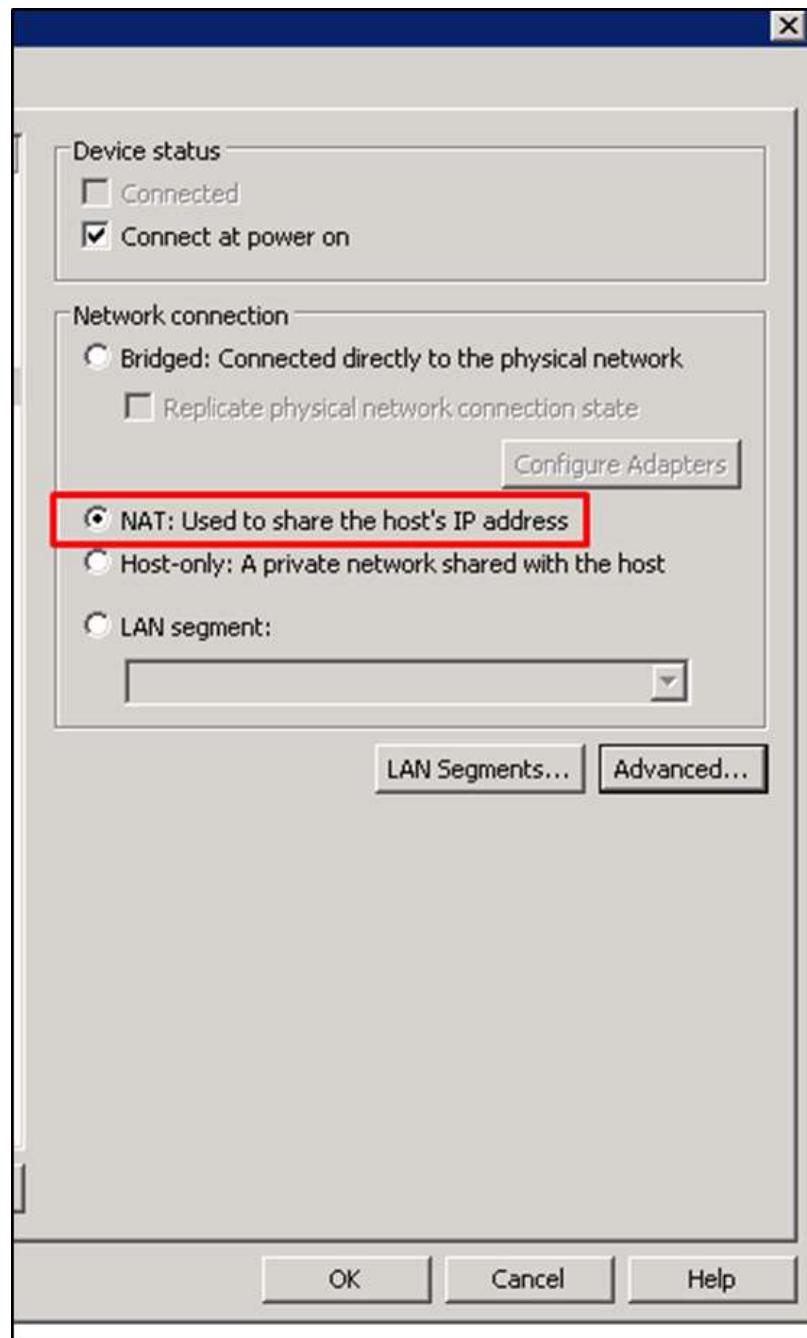
To set up the networking modes of a virtual machine in a VMware Workstation, we have to click on the "Edit virtual machine settings".



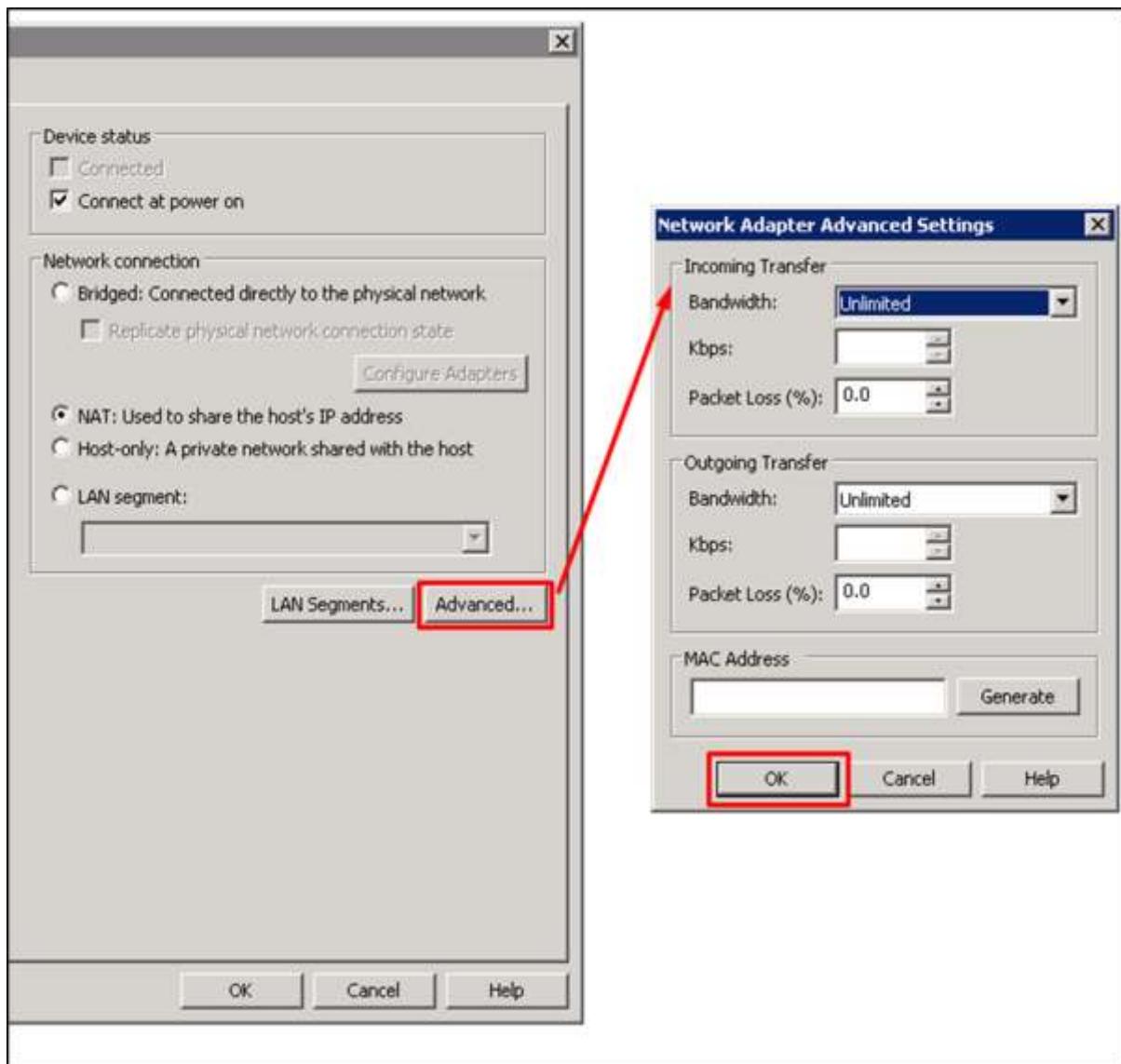
A table will be opened with the settings of networking and on the left hand side panel of this table click on "Network Adaptor".



On the left of this table, you can see the networking modes as shown in the following screenshots.



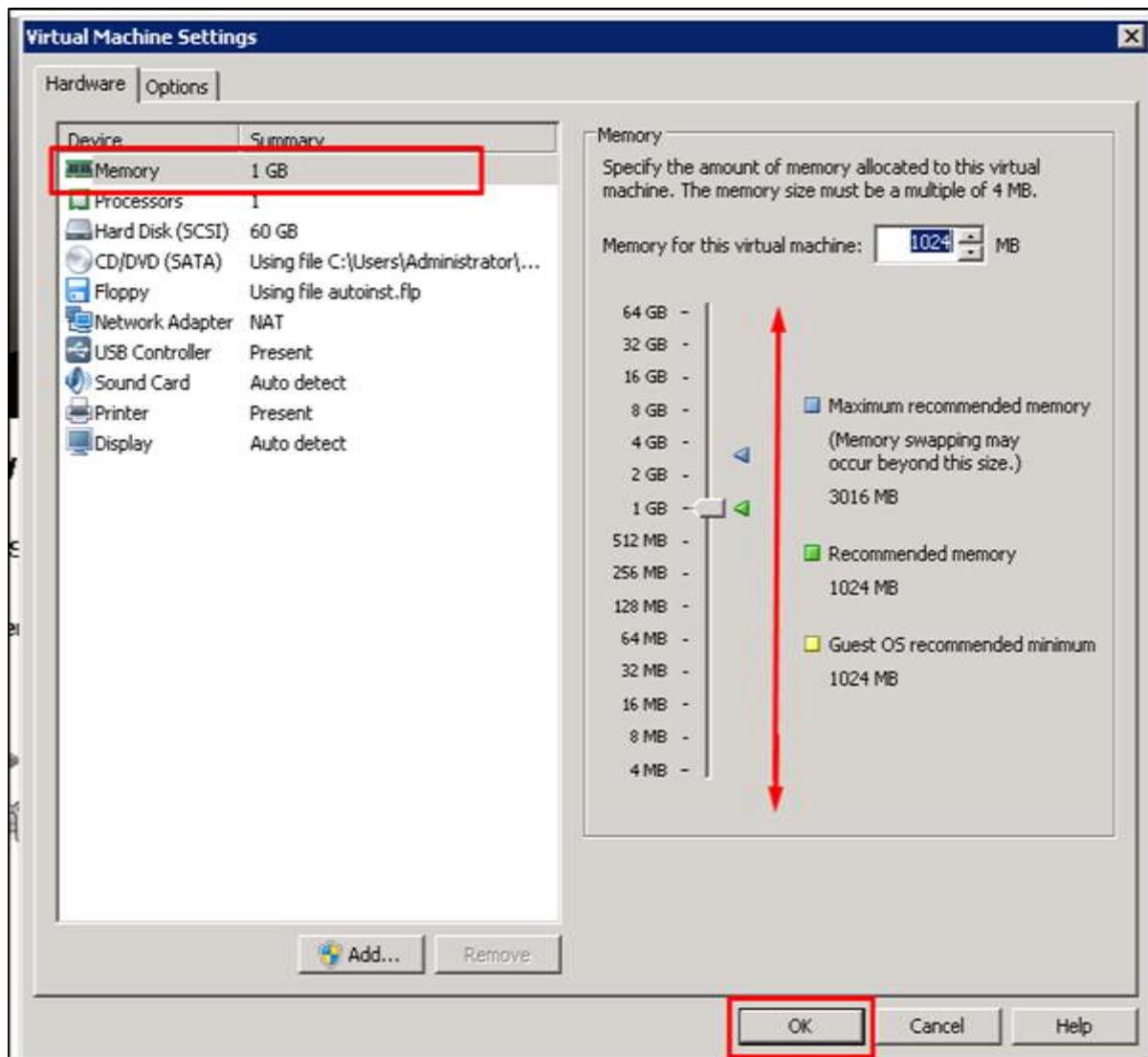
If we want to limit the bandwidth usage of a virtual machine, click on "Advance" and set the incoming and outgoing bandwidths.



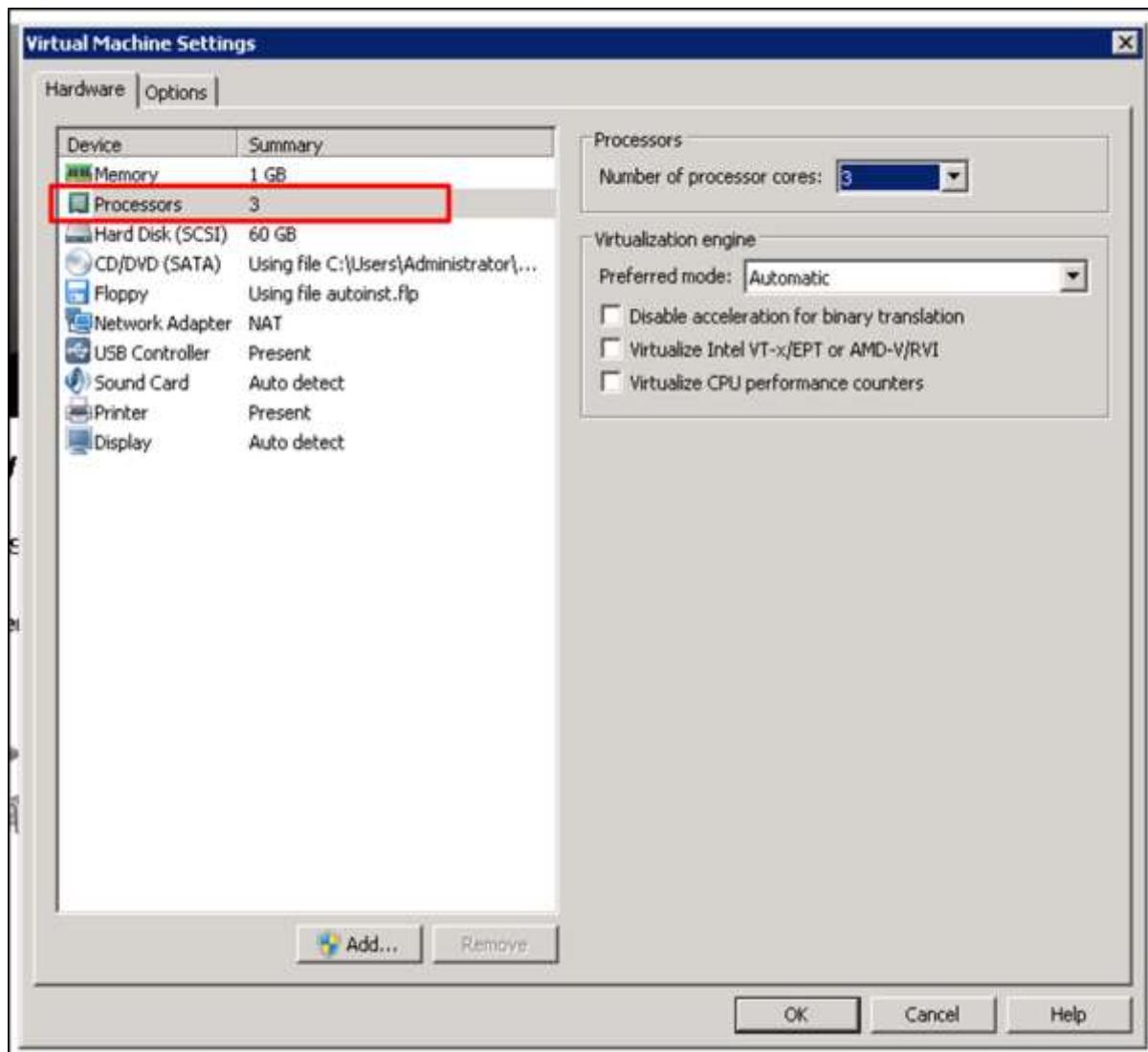
## Allocating Processors & Memory to a VM using VMware Workstation

To allocate memory to a virtual machine in a VMware Workstation, we have to click on "Edit virtual machine settings". A table will be opened and we will have to click on "Memory".

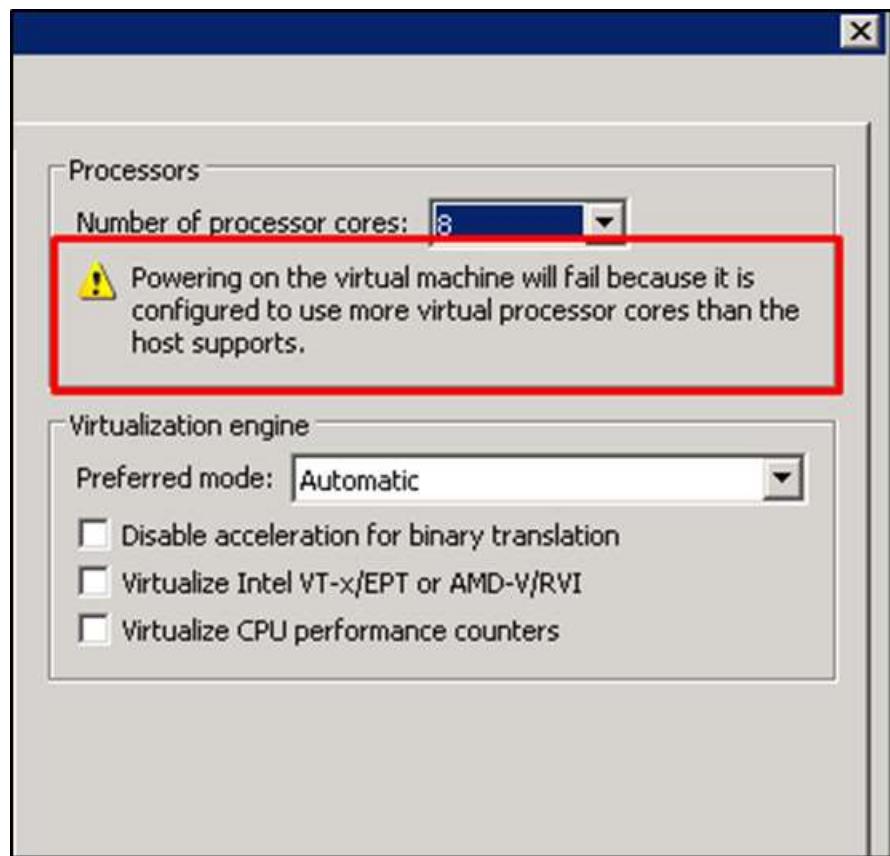
On the left hand side panel, we have to enter the amount of memory manually or by moving the arrow up and down as shown in the following screenshot.



If you click on "Processors". On the left hand side panel, we have to enter the amount of vCPU as shown in the screenshot below.



**Note:** If you put more vCPU-s than what the host supports, it will fail to power on the VM.



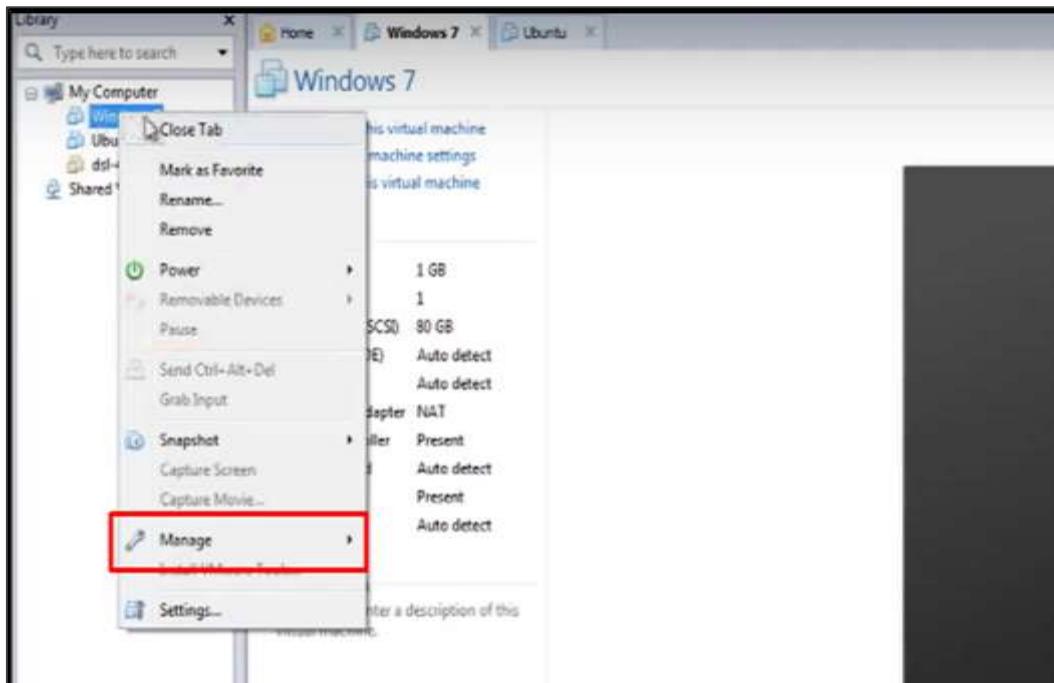
## Duplicating a VM Using VMware Workstation

---

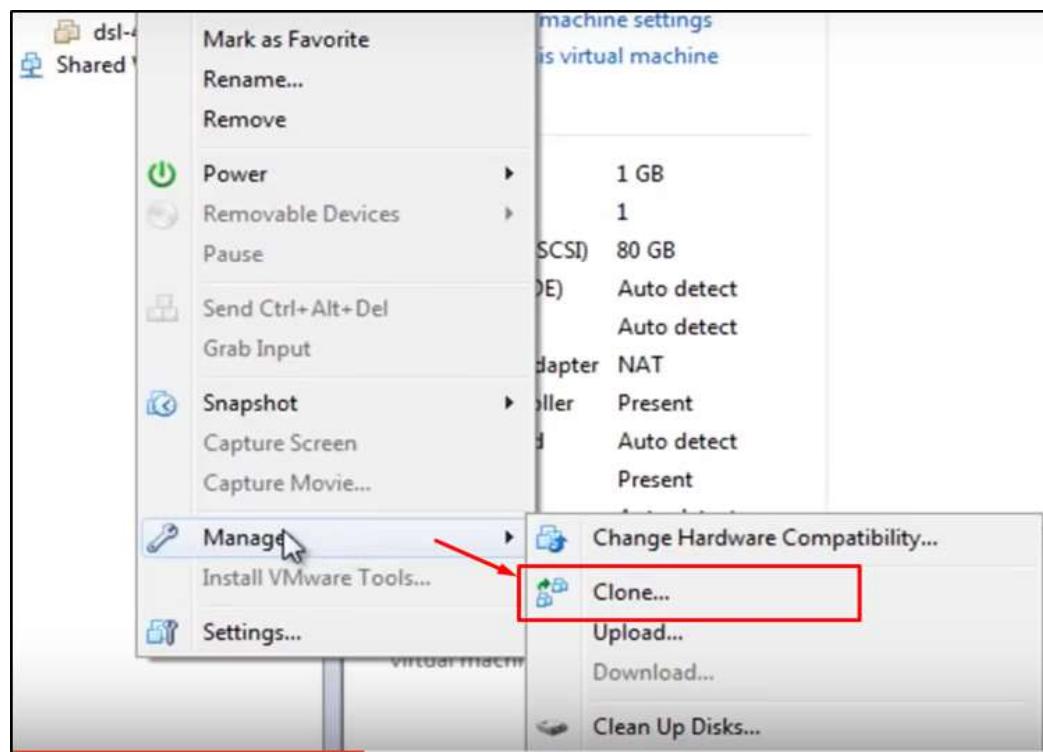
To create duplicates of VM machines, we have to use the VMware Workstation Commercial Version.

Let us see how to do it in practice by following the steps given below.

**Step 1:** Open the VMware managing console and right click on a VM that you want to duplicate. Click on "Manage".



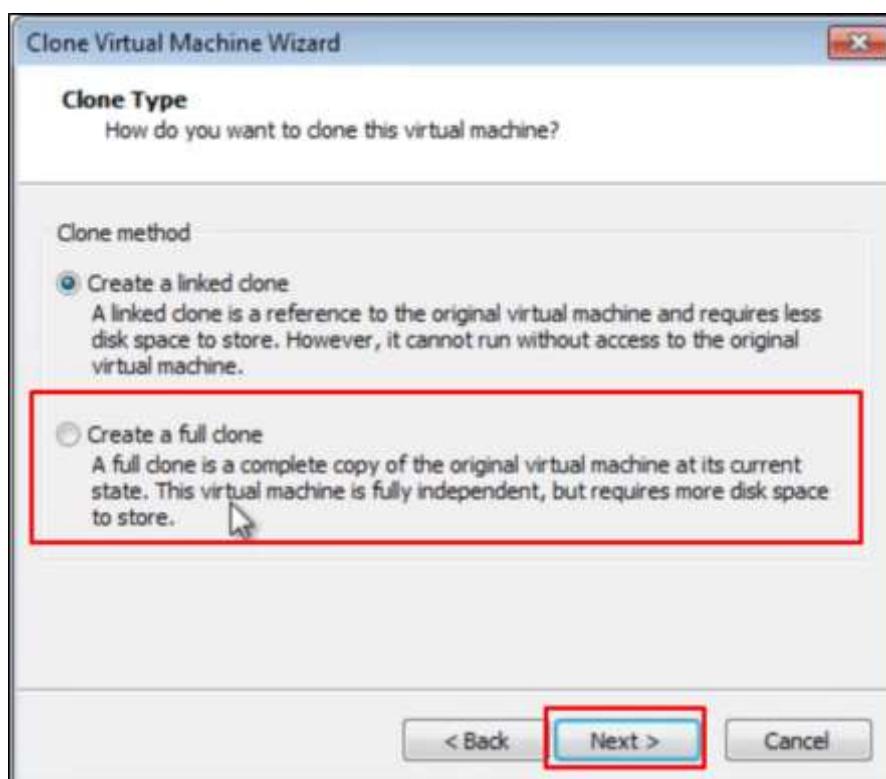
**Step 2:** Click on "Clone..." and a wizard will be open.



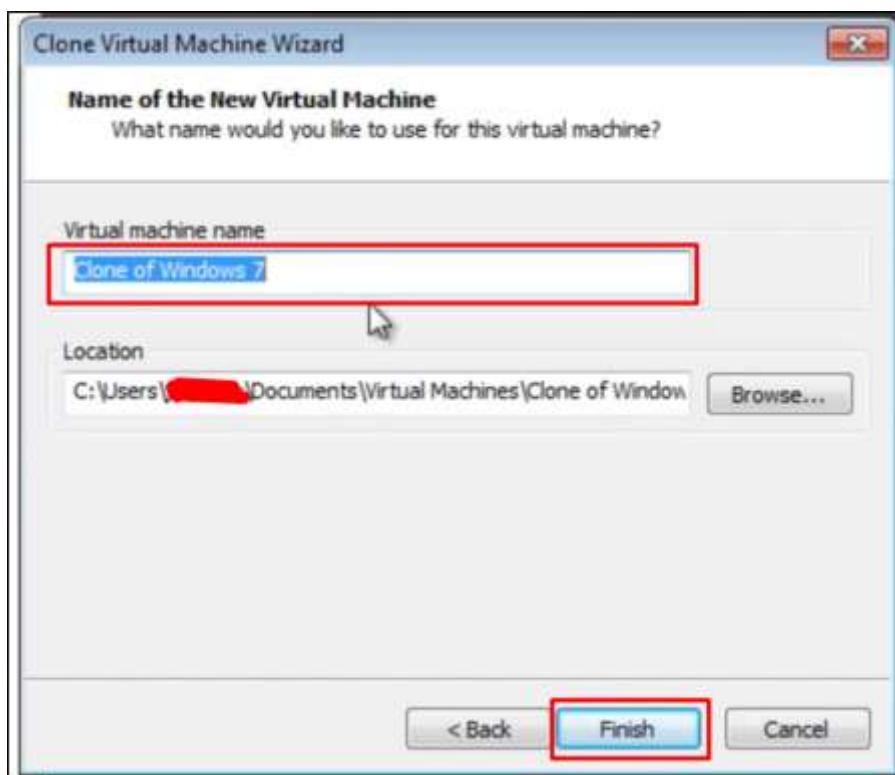
**Step 3:** Click on "Next".



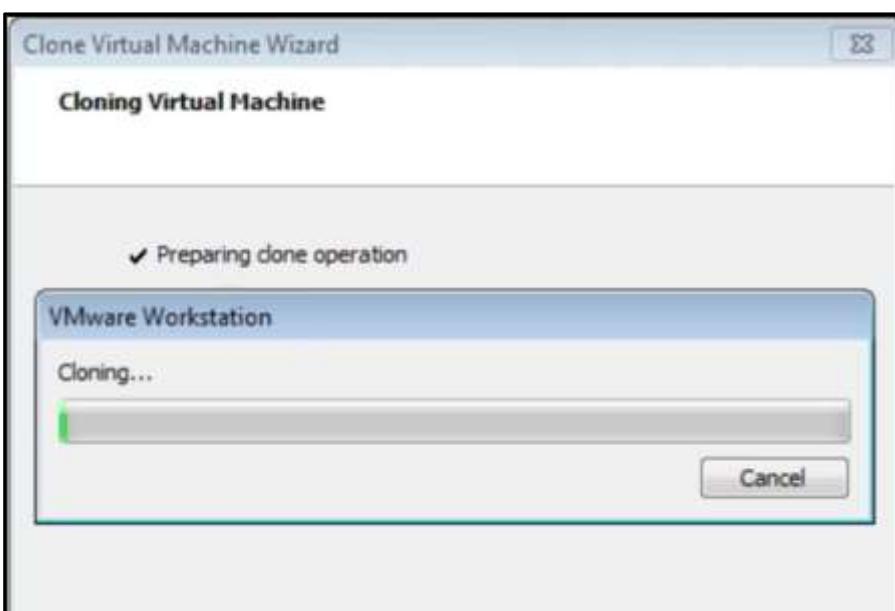
**Step 4:** Click on "Create a Full Clone" and "Next".



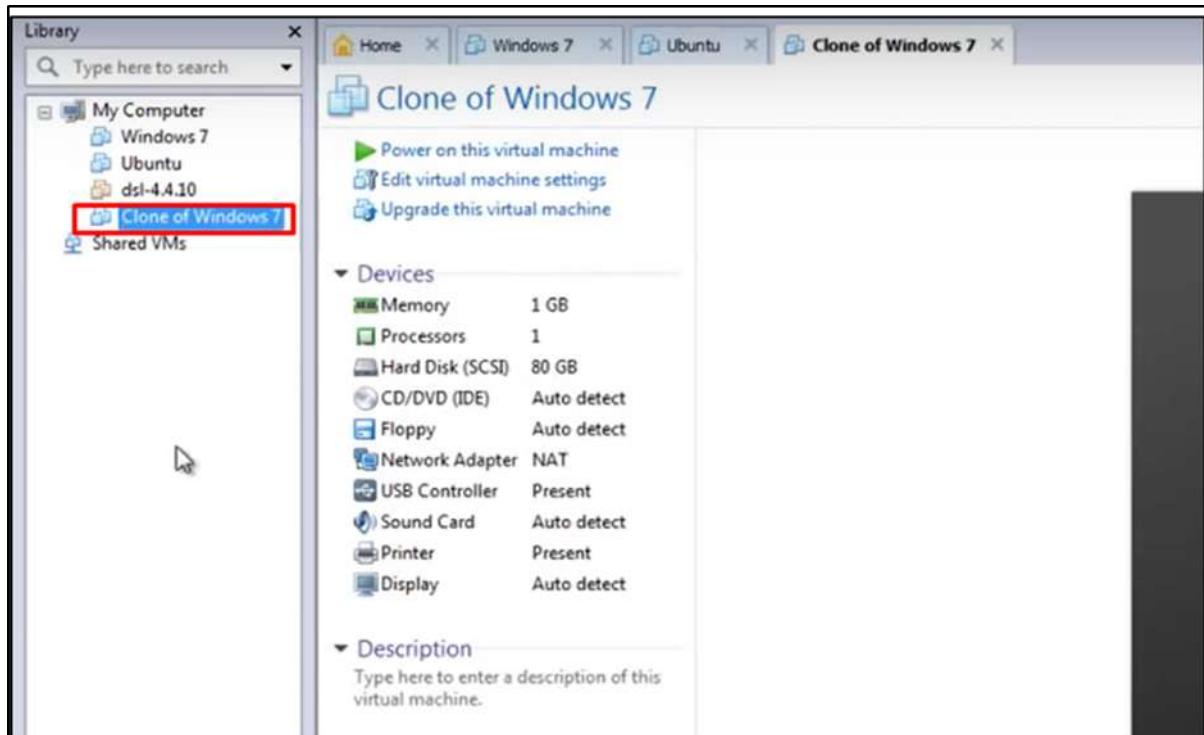
**Step 5:** Put a name for the clone that will be created and "Finish".



The following screenshots describe the process of cloning.



Once the cloning process is complete, the following window will open.



## 7. Virtualization – VirtualBox

In this chapter, we will understand what a VirtualBox is and discuss in detail the various components it has.

### Installing VirtualBox

To start with, we will download VirtualBox and install it. We should follow the steps given below for the installation.

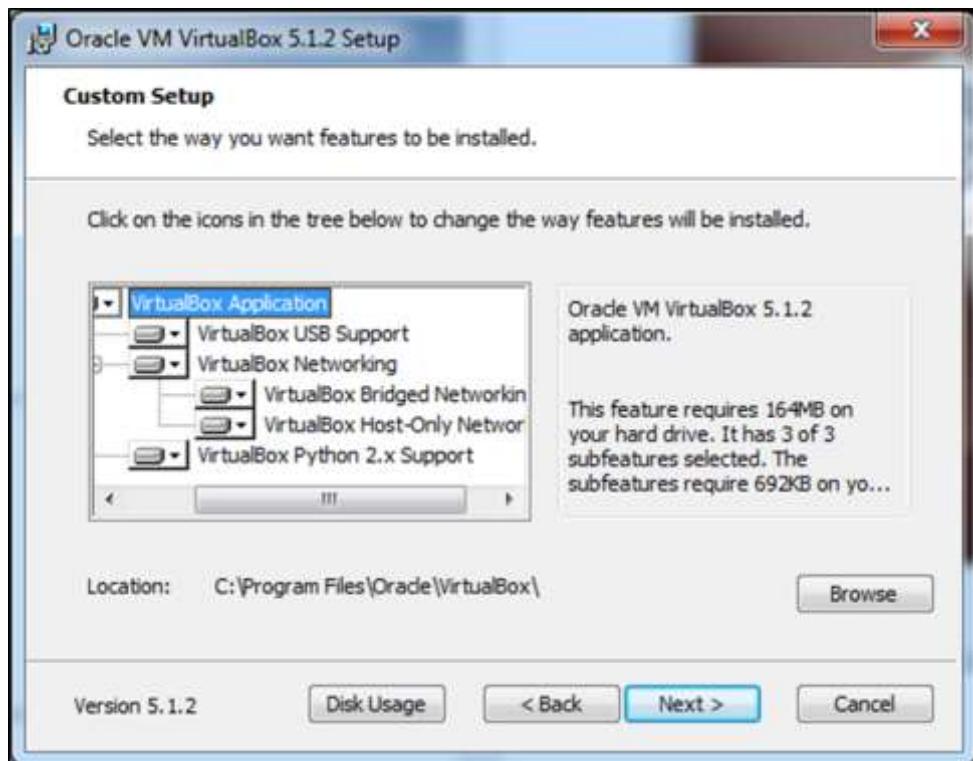
**Step 1:** To download VirtualBox, click on the following link – <https://www.virtualbox.org/wiki/Downloads>. Now, depending on your OS, select which version to install. In our case, it will be the first one (Windows host).

The screenshot shows the 'VirtualBox' download page. The main heading is 'Download VirtualBox'. Below it, a sub-section titled 'VirtualBox binaries' contains a note: 'Here, you will find links to VirtualBox binaries and its source code.' Under this, there is a section for 'VirtualBox platform packages'. It lists several options, with the first one, 'VirtualBox 5.1.2 for Windows hosts (x86/amd64)', highlighted by a red border. Other options listed include 'VirtualBox 5.1.2 for OS X hosts (amd64)', 'VirtualBox 5.1.2 for Linux hosts (amd64)', and 'VirtualBox 5.1.2 for Solaris hosts (amd64)'. Below this list, there is a section for 'VirtualBox 5.1.2 Oracle VM VirtualBox Extension Pack' with a note about PUEL and links to download the extension pack for different versions.

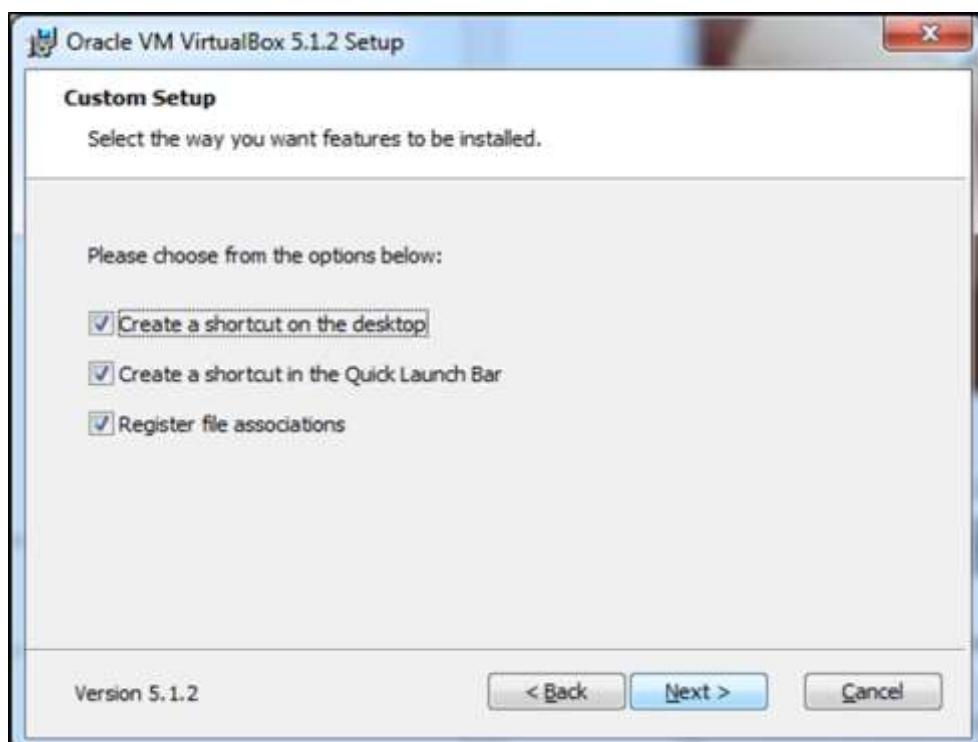
**Step 2:** Once the option is selected, click on “Next”.



**Step 3:** You have the option asking where to install the application. We can leave it as default and click on "Next".



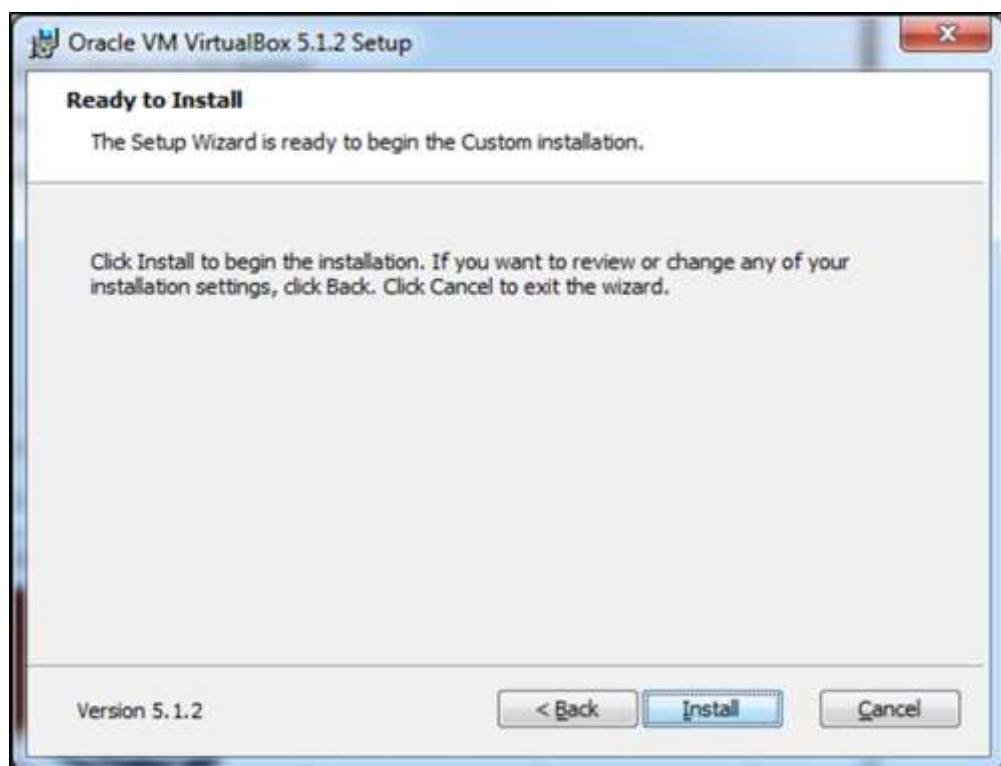
**Step 4:** Once the options are selected as shown in the following screenshot, click on Next.



**Step 5:** A dialog box will come up asking whether to proceed with the installation. Click "Yes".



**Step 6:** In the next step, click on "Install".



**Step 7:** Tick the start VirtualBox check box and click on "Finish".



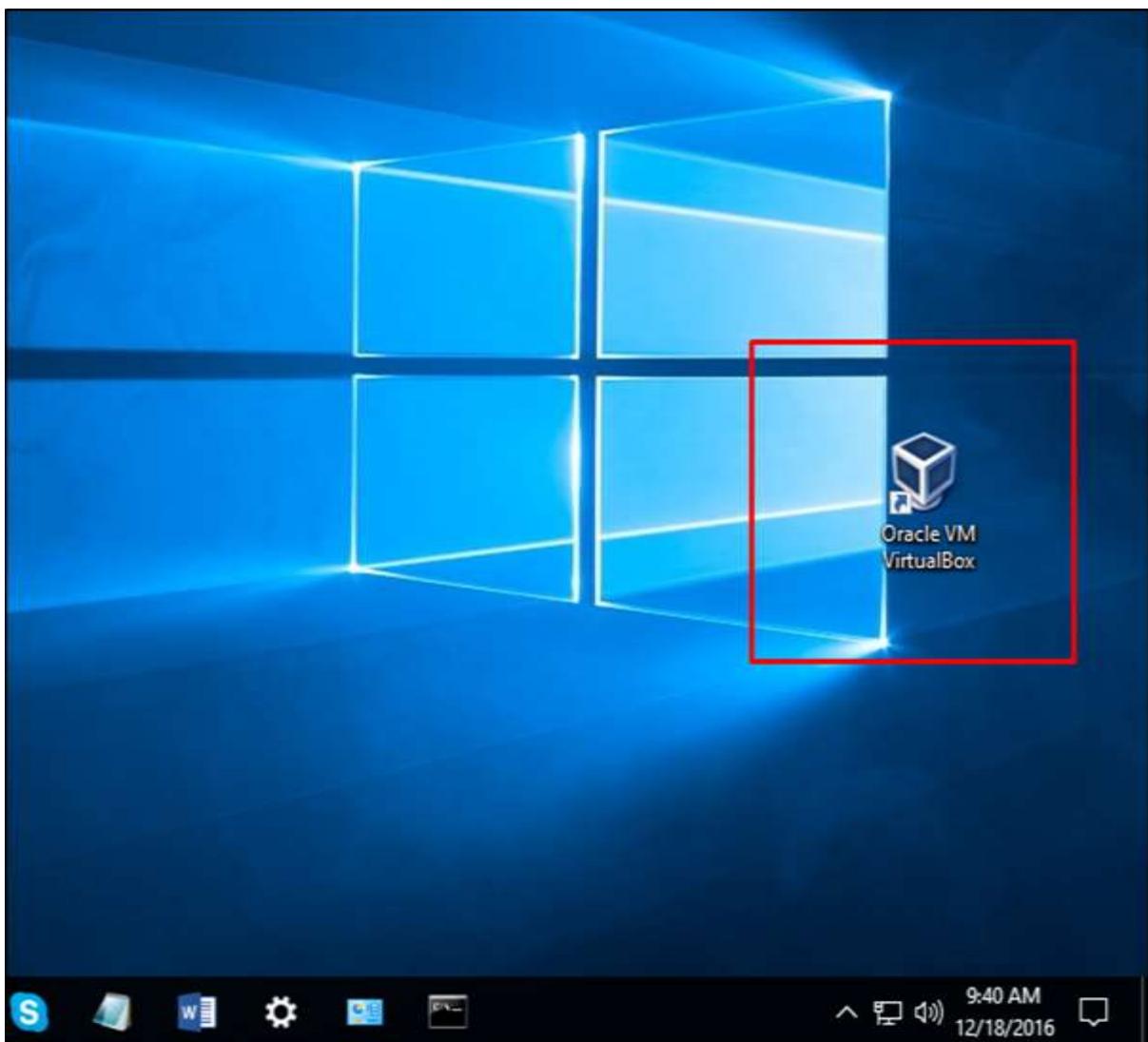
**Step 8:** VirtualBox application will now open as shown in the following screenshot. Now, we are ready to install the virtual machines.



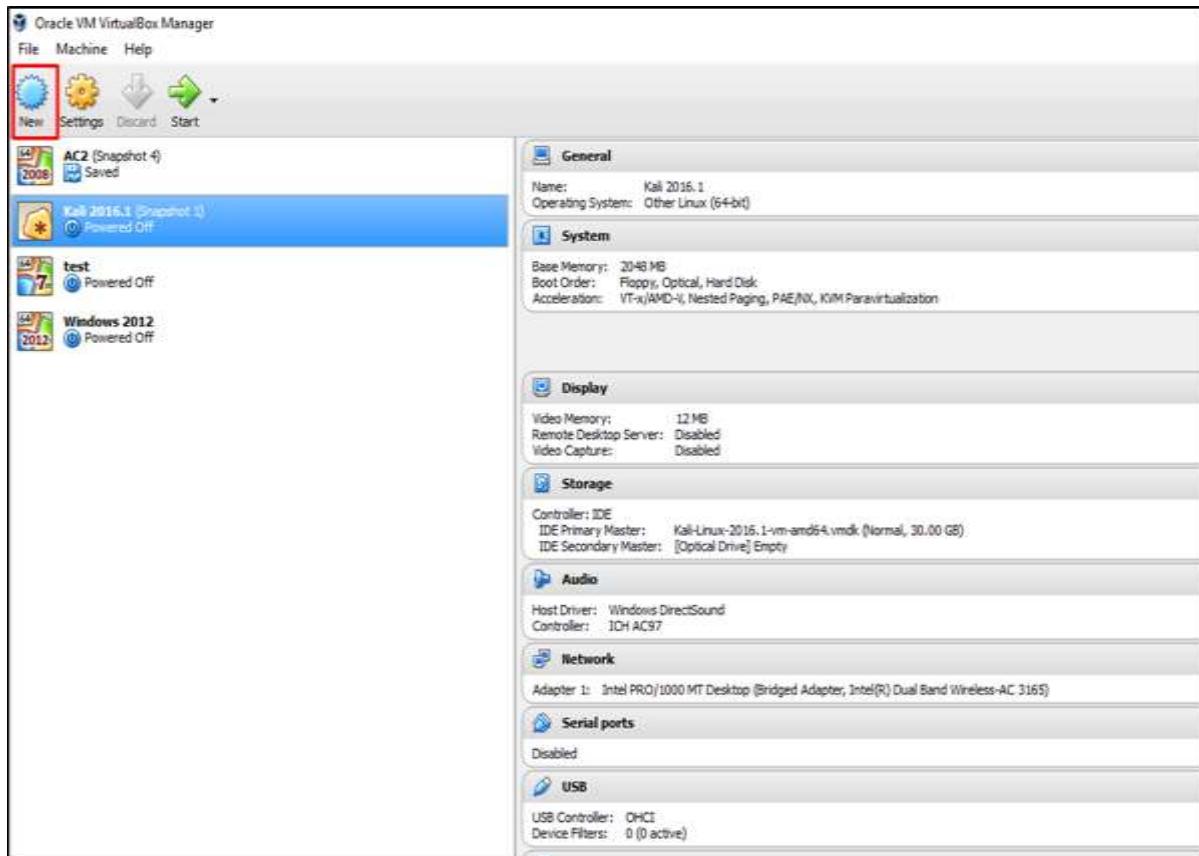
## Creating a VM with VirtualBox

To create a virtual machine with Oracle VirtualBox, we should follow the steps given below.

**Step 1:** To begin with, click on the “Oracle VM VirtualBox” icon on the desktop as shown in the screenshot below.



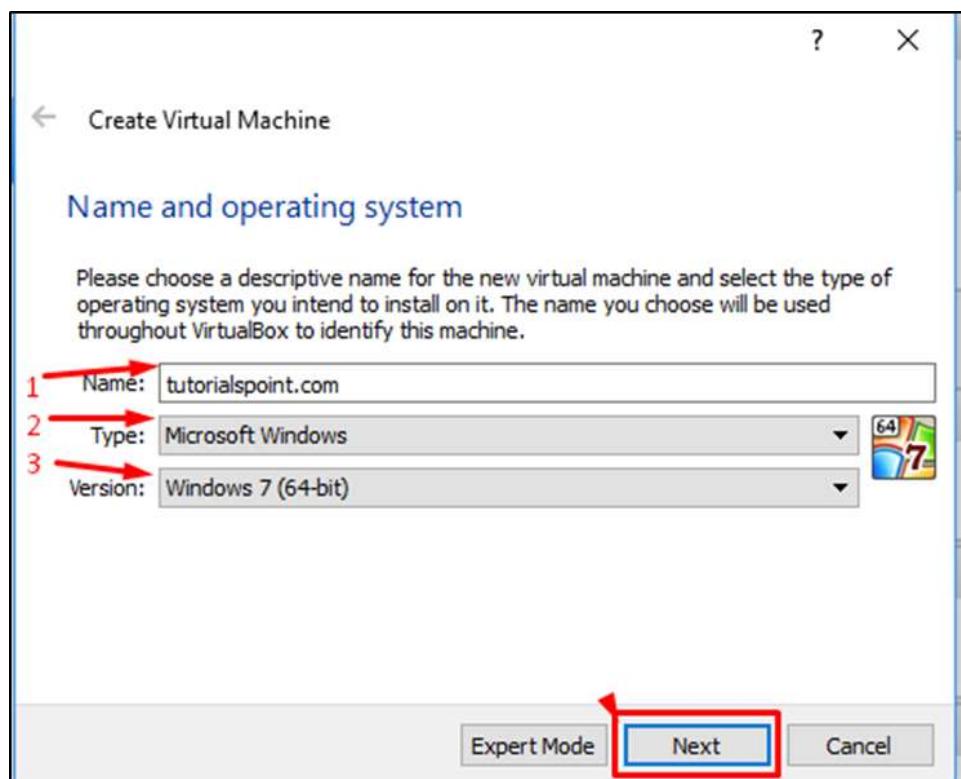
**Step 2:** The next step is to click on “New” button, which is in the top left hand side of the screen.



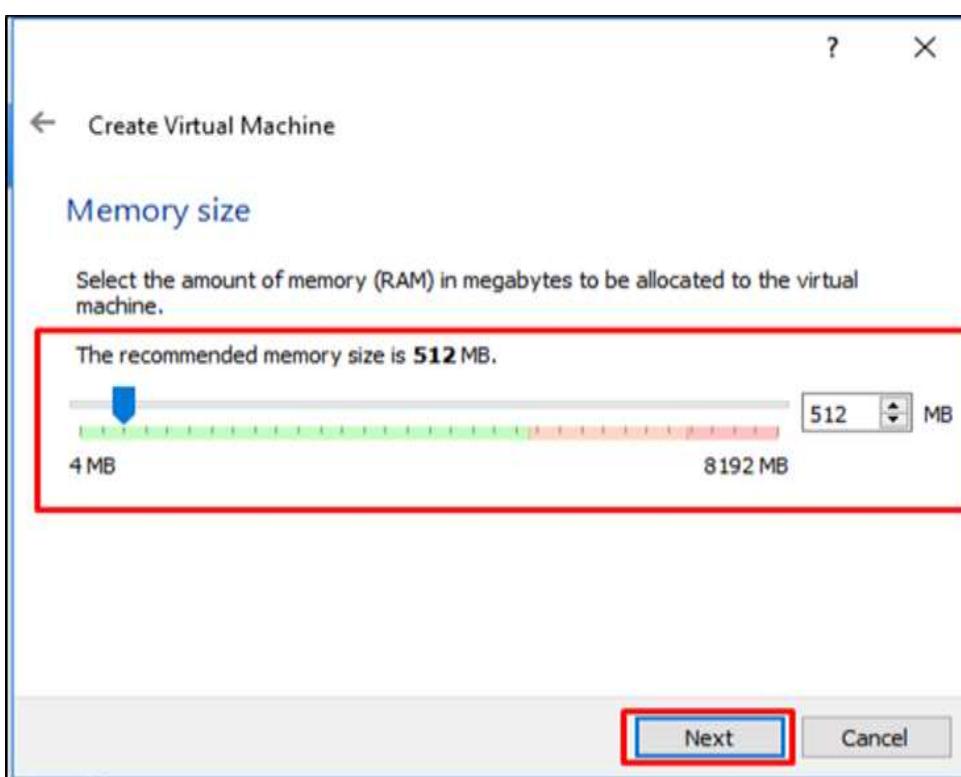
**Step 3:** A table will pop-up requesting you the parameters for the virtual machine. These will be –

- **Name:** We have to put a friendly name for this Virtual Machine.
- **Type:** Enter the OS that is going to be installed on it.
- **Version:** Enter the specific version for that OS, which we have selected earlier.

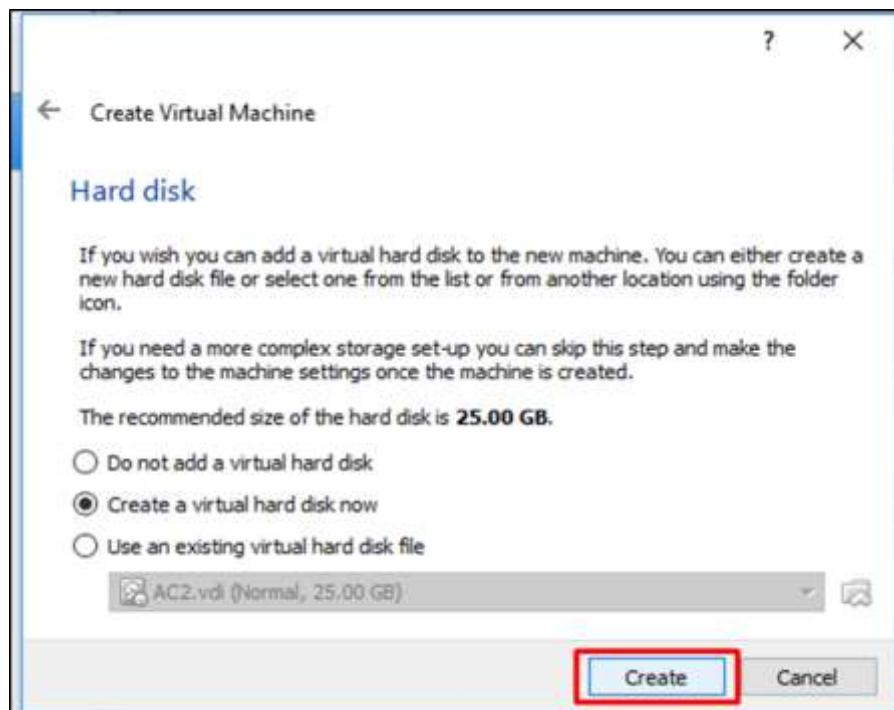
Once all the above parameters are filled, click on "Next".



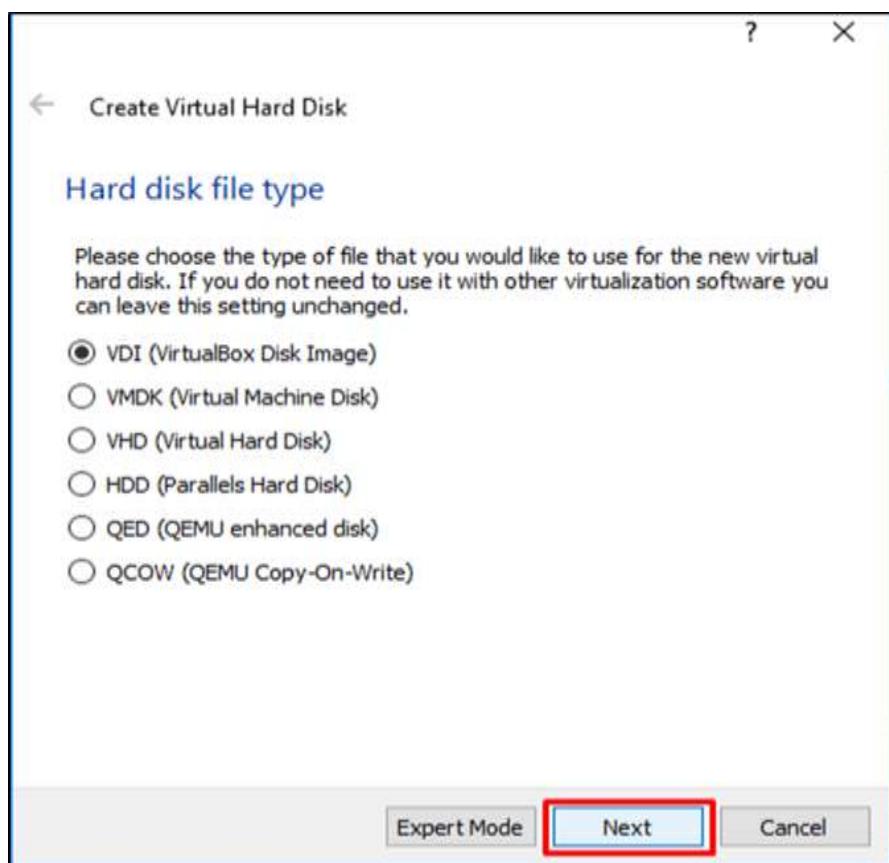
**Step 4:** Select the amount of memory that you need to allocate in this VM → Click on "Next".



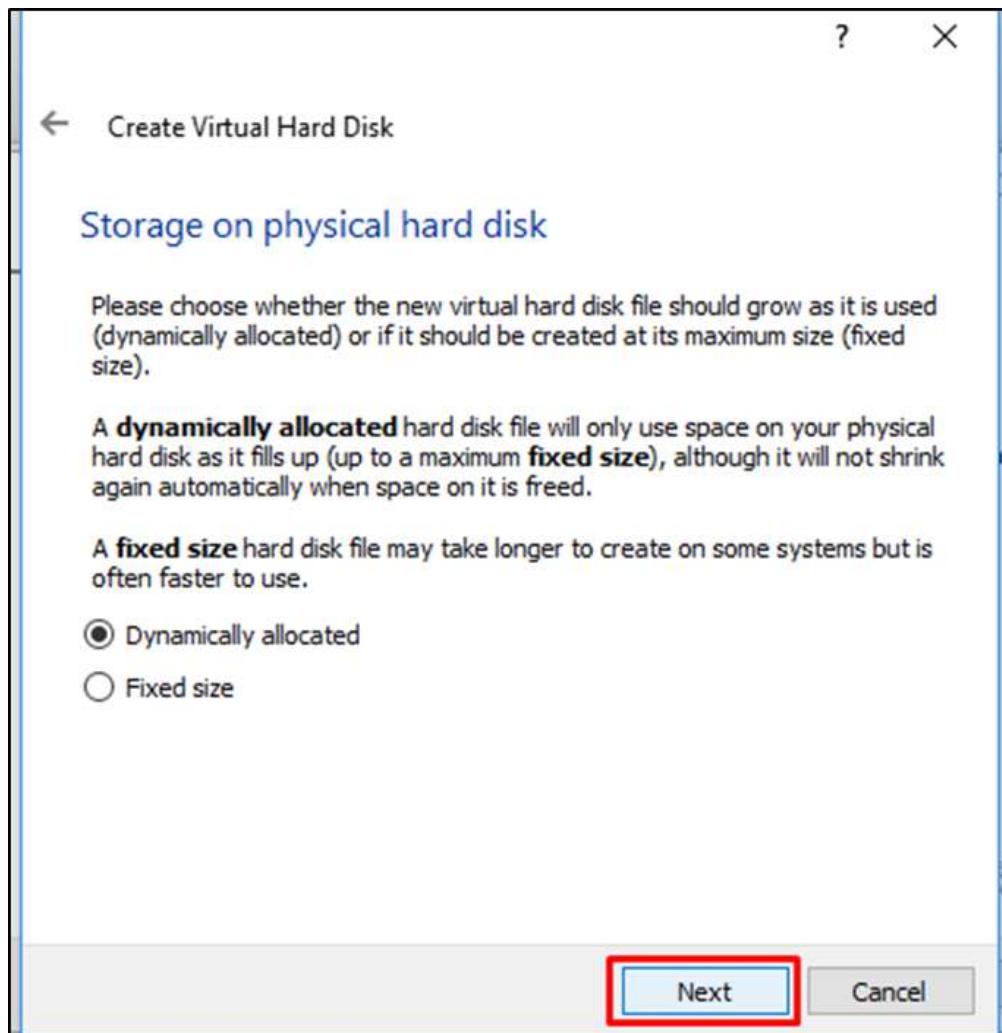
**Step 5:** Check one of the three options for the HDD and click on "Create".



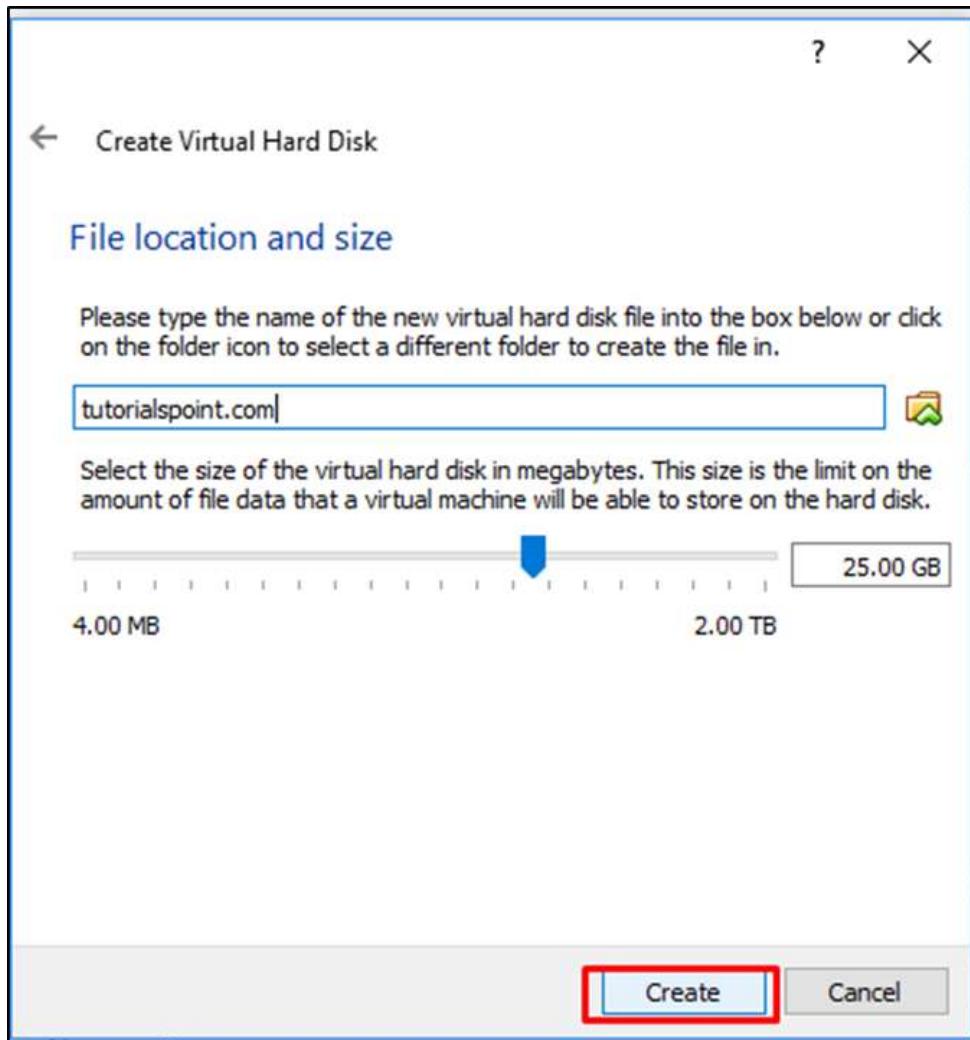
**Step 6:** Select a file extension for your virtual HDD (It is recommended to use a common file extension that most of the hypervisors use like VHD) → click on "Next".



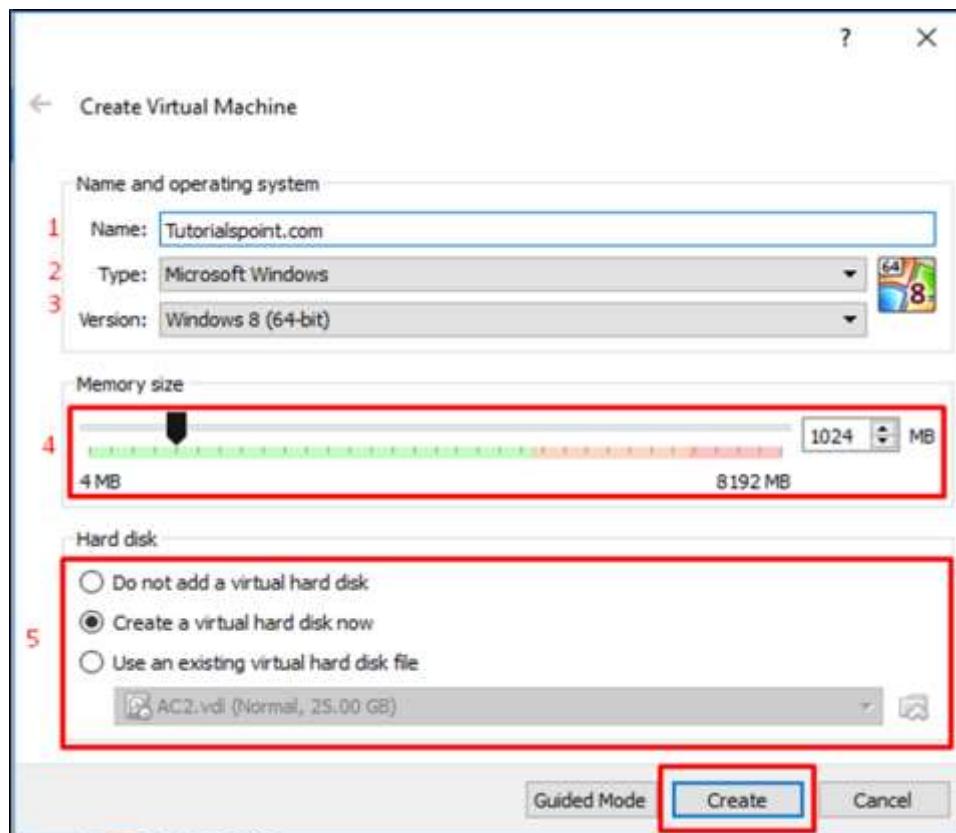
**Step 7:** Choose whether you want the Virtual HDD as dynamic or fixed. This is based on your needs → Click on "Next".



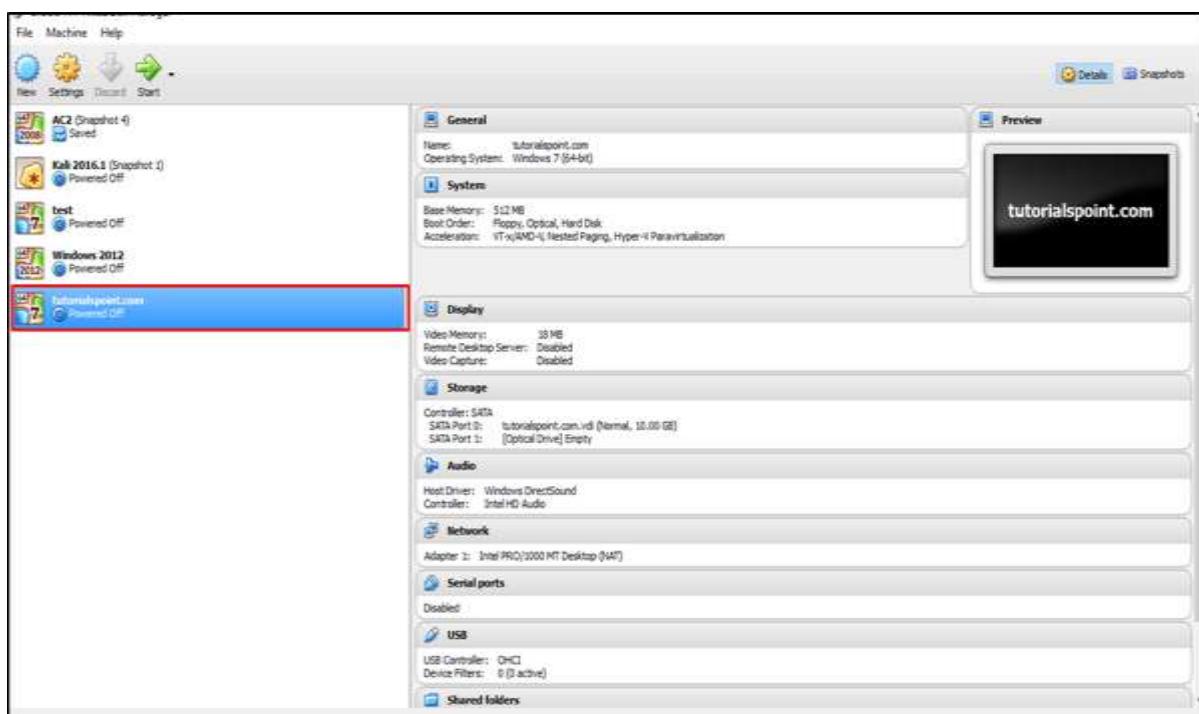
**Step 8:** Put a name for your virtual HDD file and select the disk size for your VM → Click on “Create”.



All the above steps can be done in one shot by selecting the "Expert mode".



The virtual machine created will be as shown in the screenshot below.



## Setting up Networking with VirtualBox

There are two types of networking modes in VirtualBox, which are –

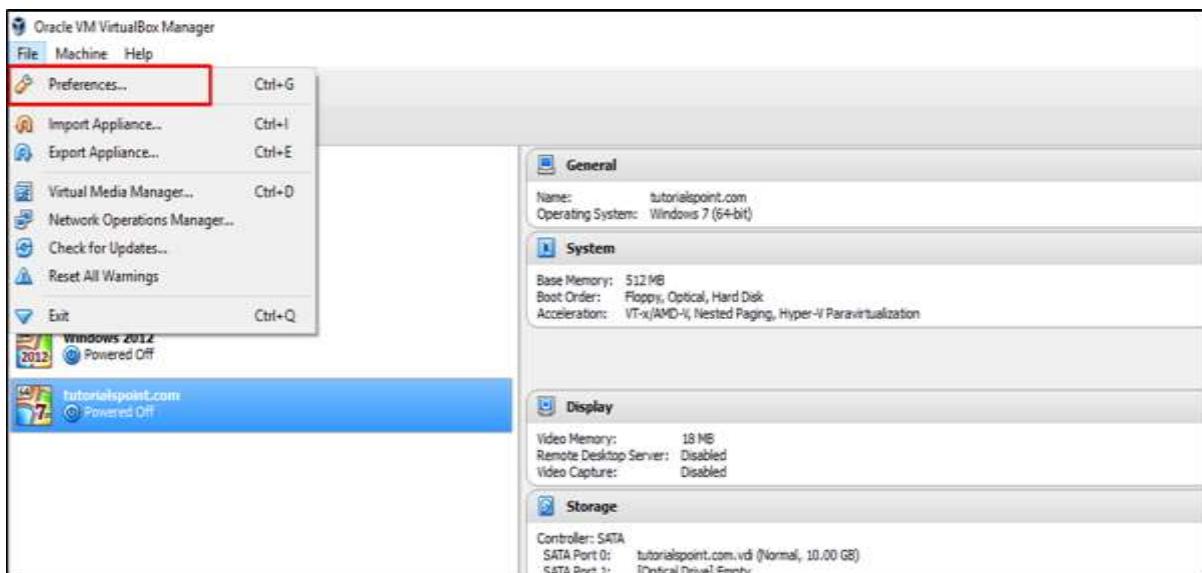
- Nat Networks and
- Host-only Networks.

Both of these are explained in detail below.

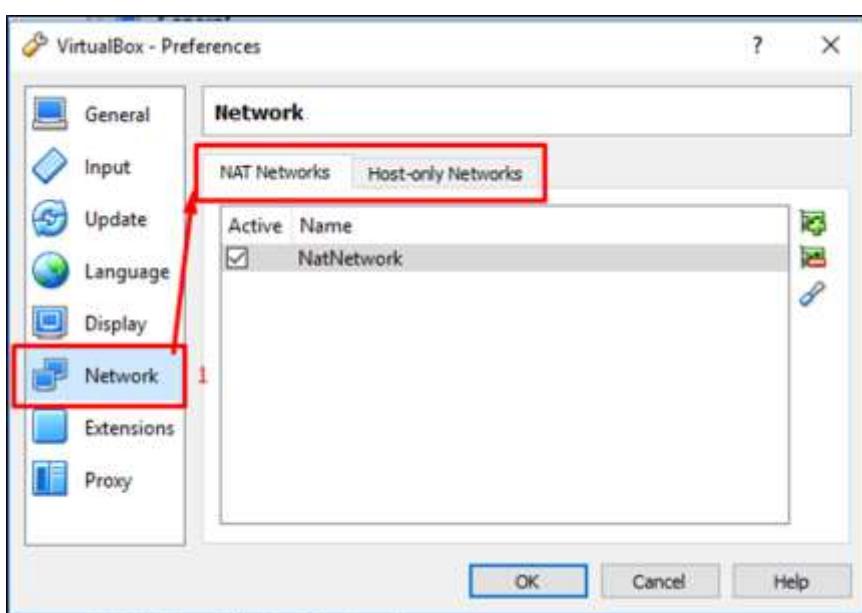
### Nat Networks

For setting up Nat Networks, we should follow the steps given below.

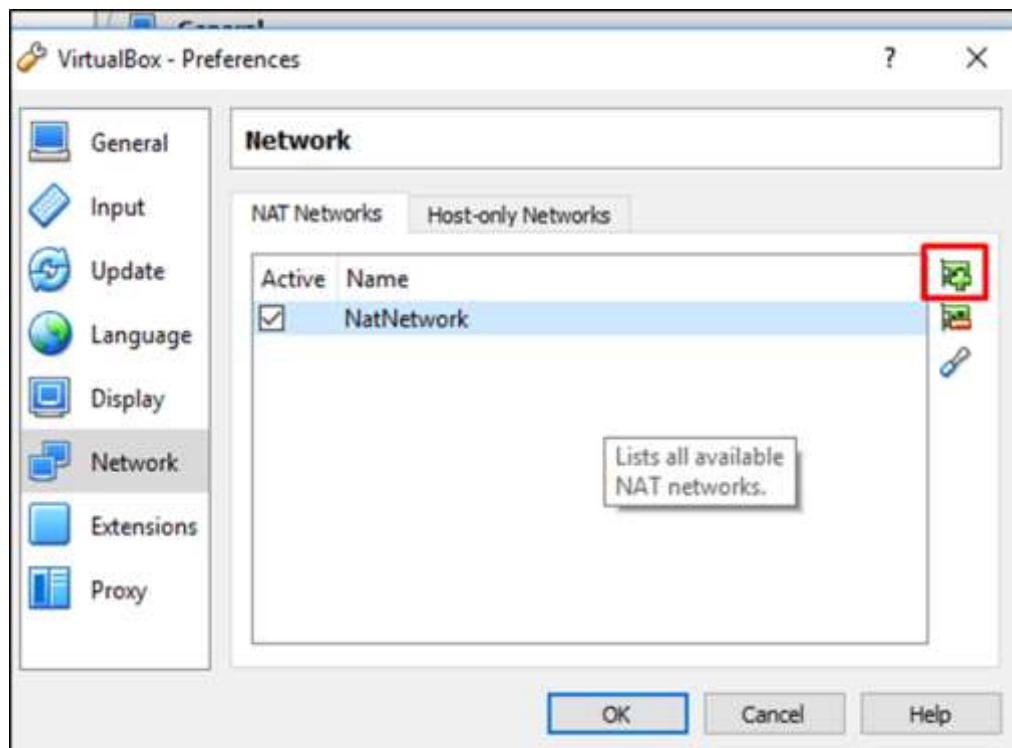
**Step 1:** Go to Oracle VM VirtualBox Manager → Click on “Preferences...”



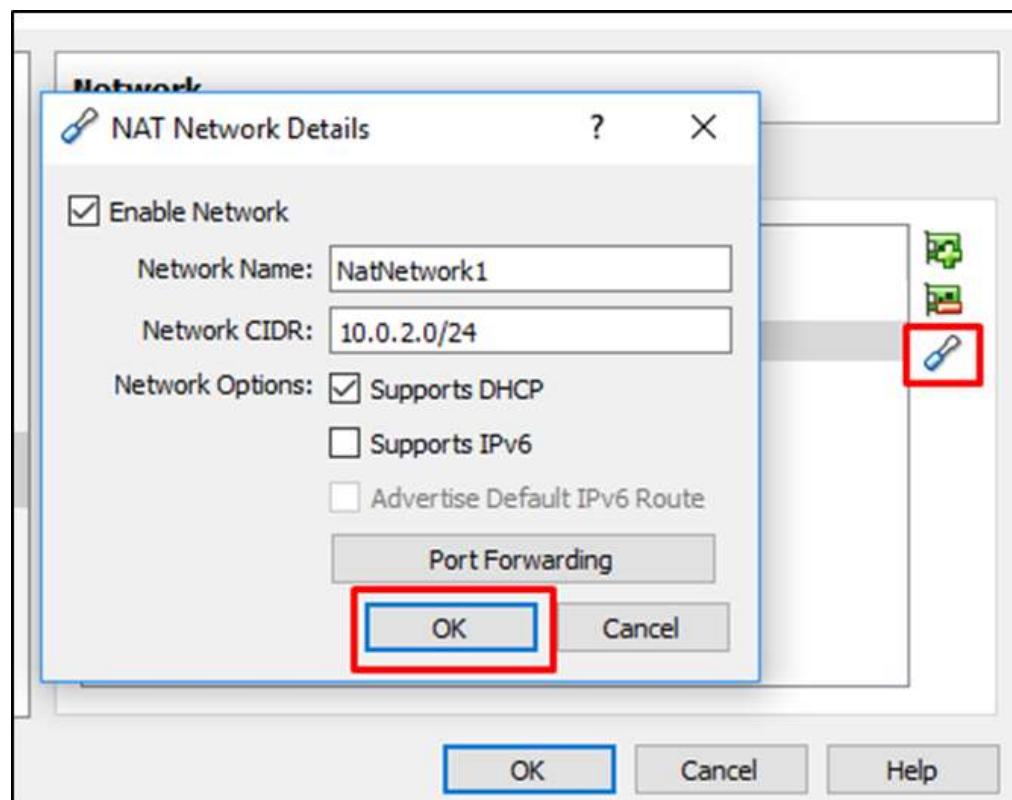
**Step 2:** Click on “Network” and then on the left panel click on the “NAT Networks” tab.



**Step 3:** Click on the “+” button, which is highlighted in the screenshot below.



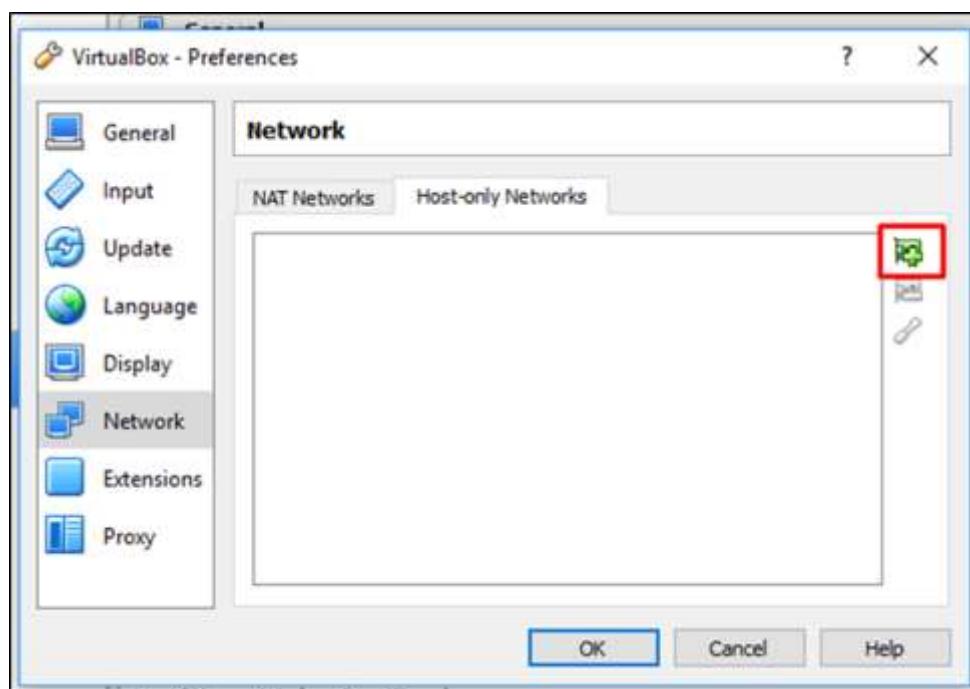
**Step 4:** Here, we have to put the “Network Name” and the IP range for this network that will be NAT-ed, in order to have access to internet and to other networks.



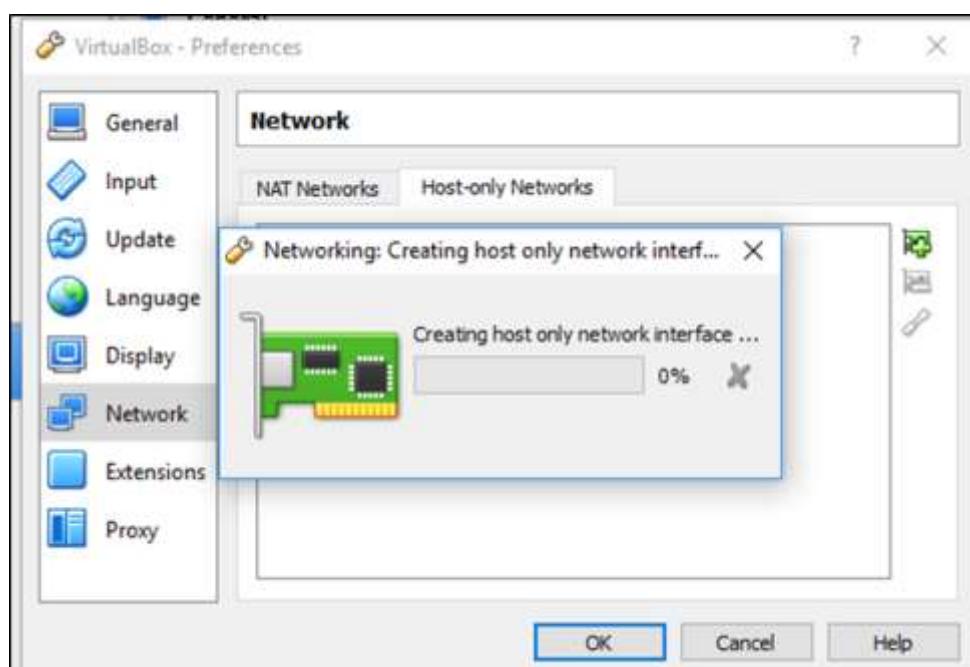
## Host-only Networks

For setting up Host-only Networks, we should follow the steps given below.

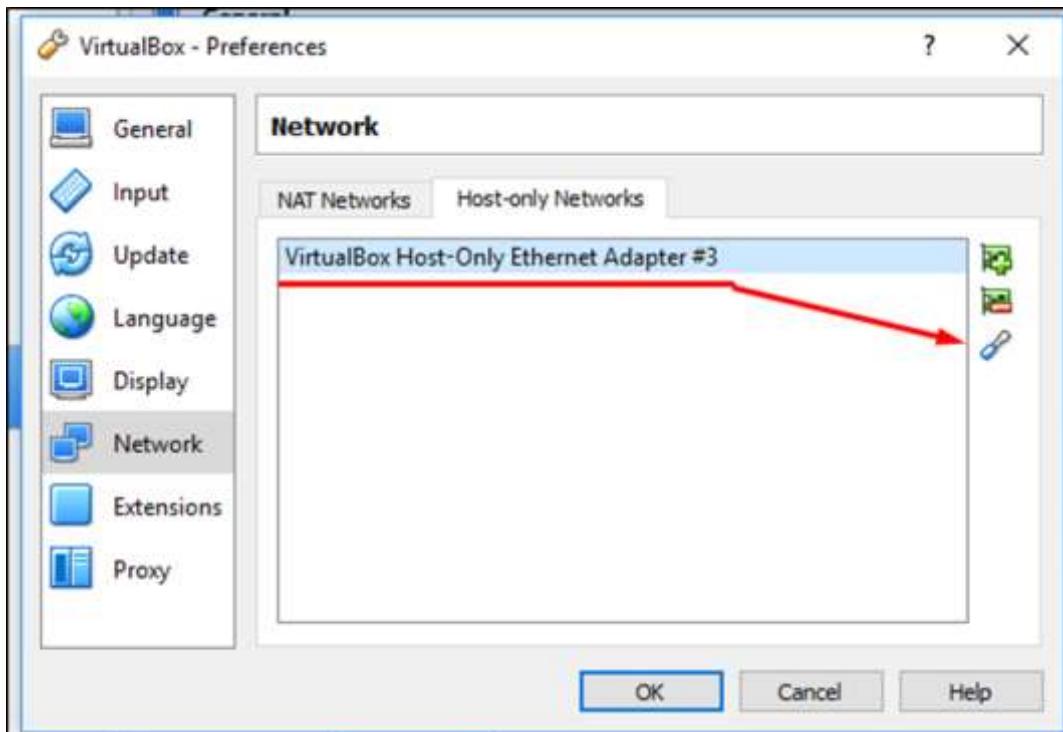
**Step 1:** If you click on the “Host-only Networks” tab, you can create networks that are isolated from the other networks. However, VM hosts communicate with each other and the Hypervisor machine. Click on the “+” sign.



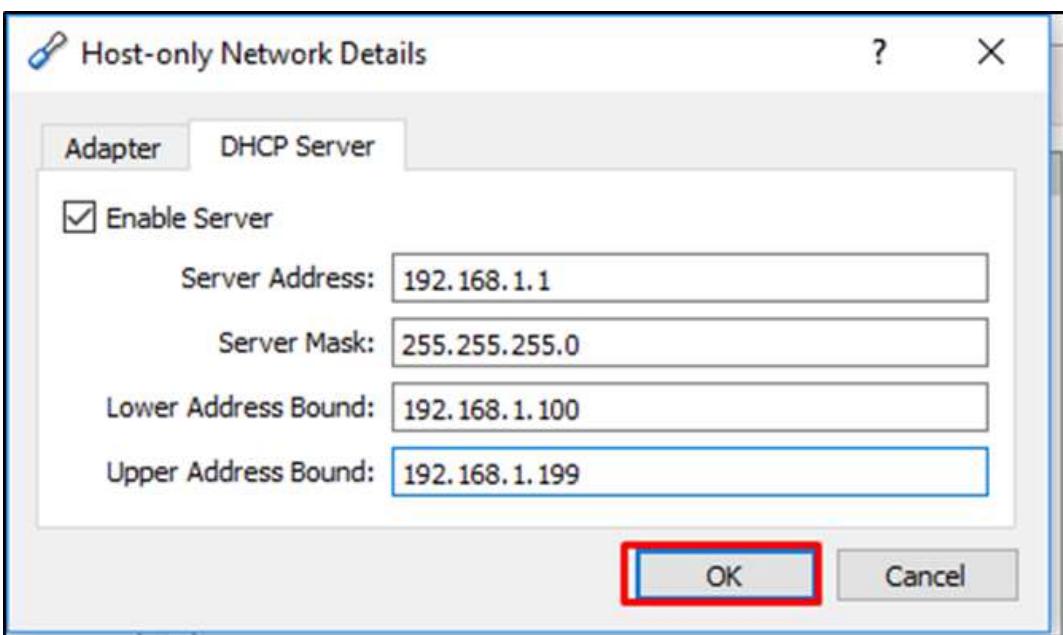
**Step 2:** The host interface will continue to be created as shown in the screenshot below.



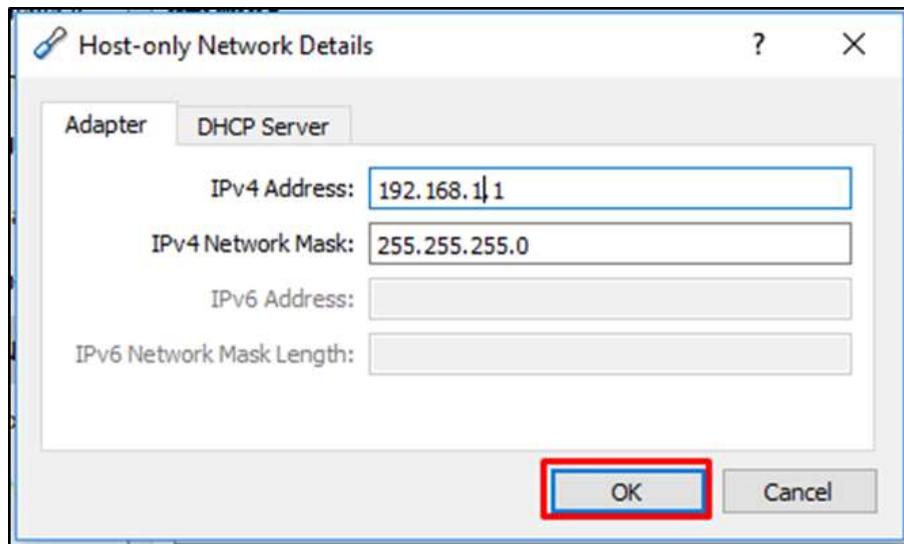
**Step 3:** If you click on  button, you can edit the settings.



**Step 4:** If you want your host machines to take "DHCP IP", click on the "DHCP Server" tab and check the box "Enable Server" → Click "OK".

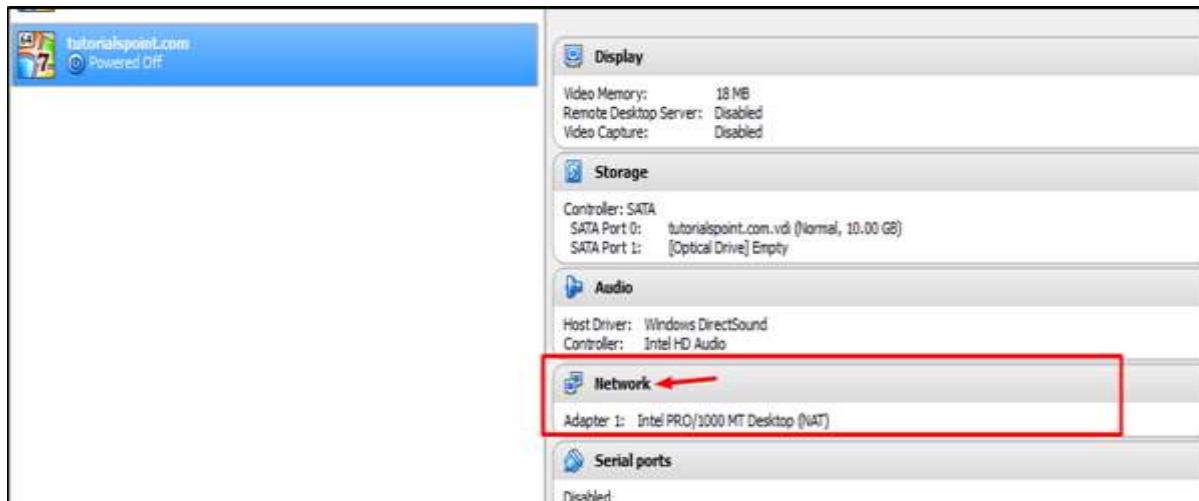


**Step 5:** In the “Adapter” tab, put the IP of the hypervisor.



After all these preparations for setting up the network modes is complete. It is now time to assign a network to our VMs.

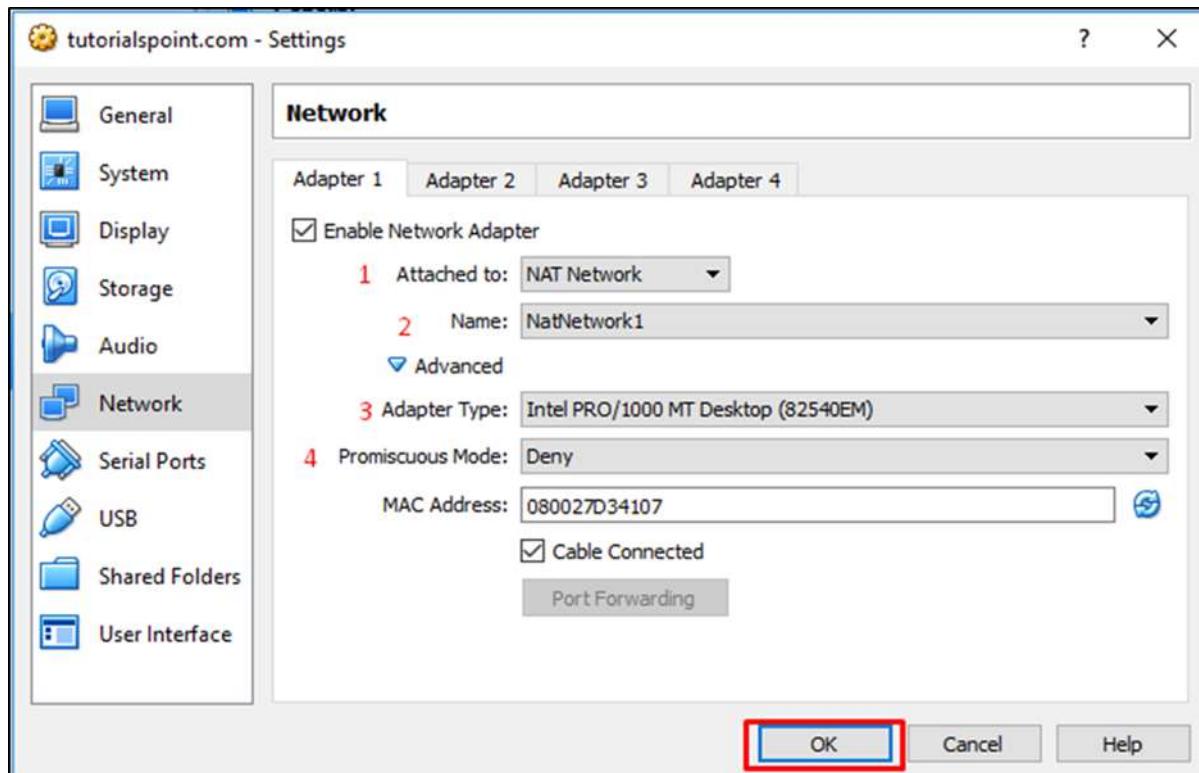
To do this, Click on the VMs on the left side of the panel, then right click on the “Network” option and a table will be open.



You can have up to four Ethernet adaptors per machine. The following image has four sections highlighted, which are explained below.

1. Check the box “Enable Network Adapter” to enable the vNIC on the VM and attach it to one network.
2. You can have many networks created, so we have to select one of them in the “Name” dropdown box.
3. In the adapter type dropdown-box, we have to select a physical NIC that the hypervisor has.
4. Promiscuous Mode: Here, we can select “Deny”, if we do not want the VMs to communicate with each other.

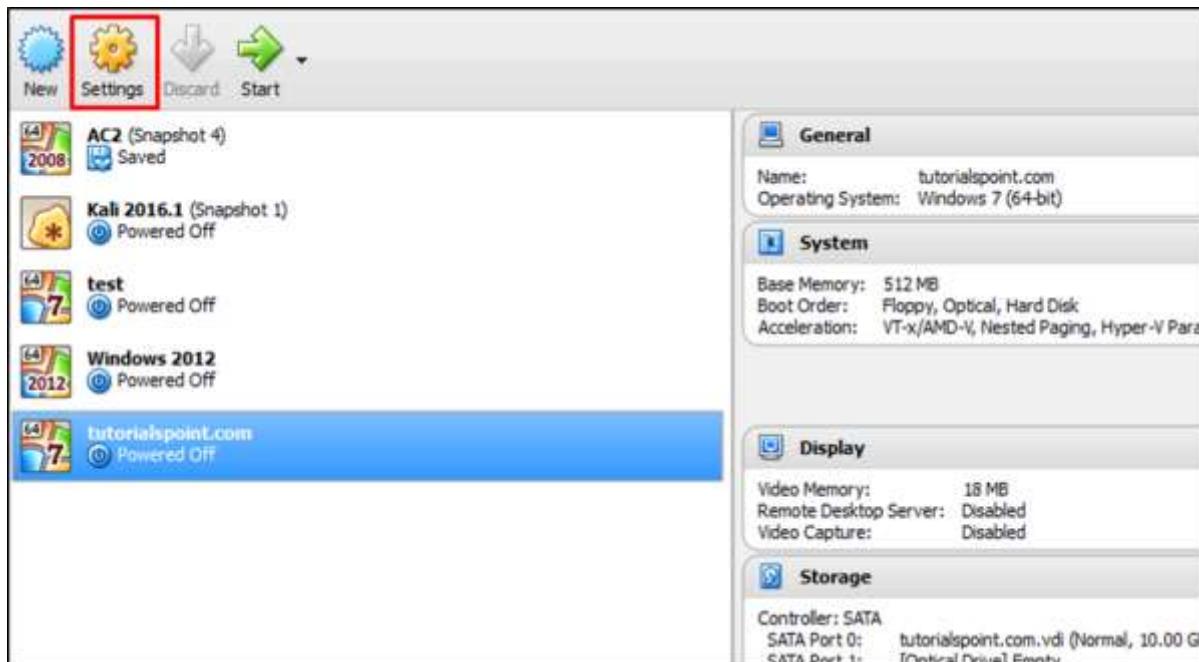
Once all the above parameters are completed. Click on "OK".



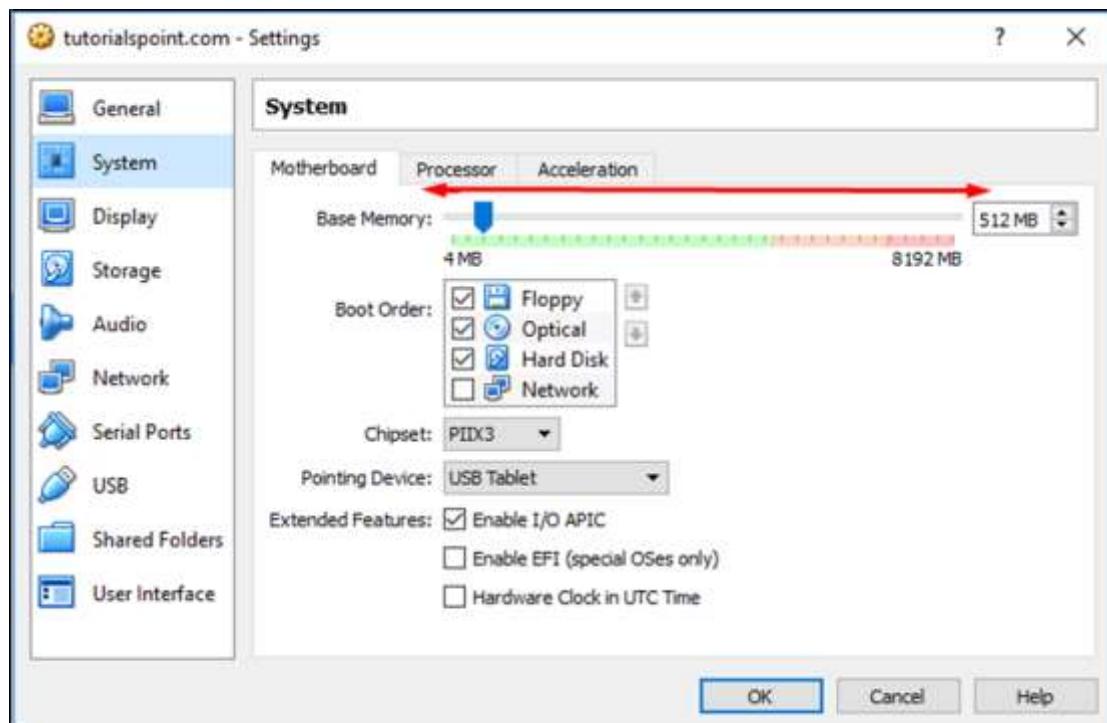
## Allocating Processors & Memory to a VM

To allocate processors and memory to a virtual machine using VirtualBox, we should follow the steps given below.

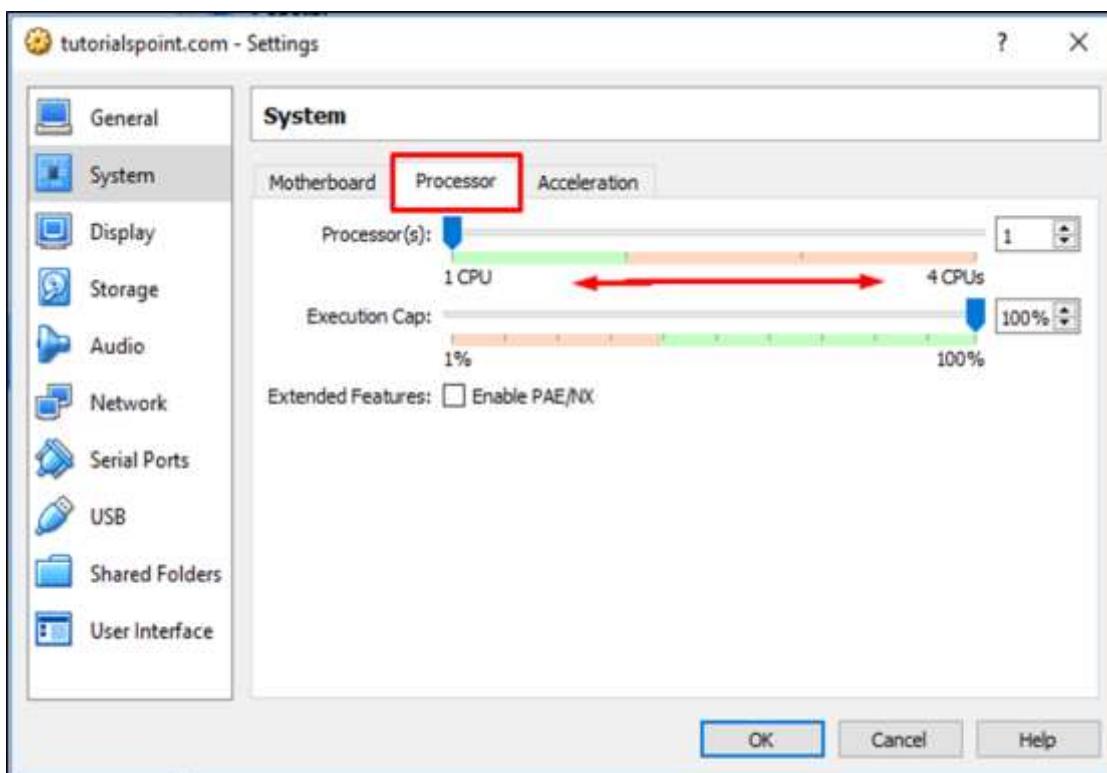
**Step 1:** To allocate a processor and memory, you have to click on "Settings" after you have selected the VM.



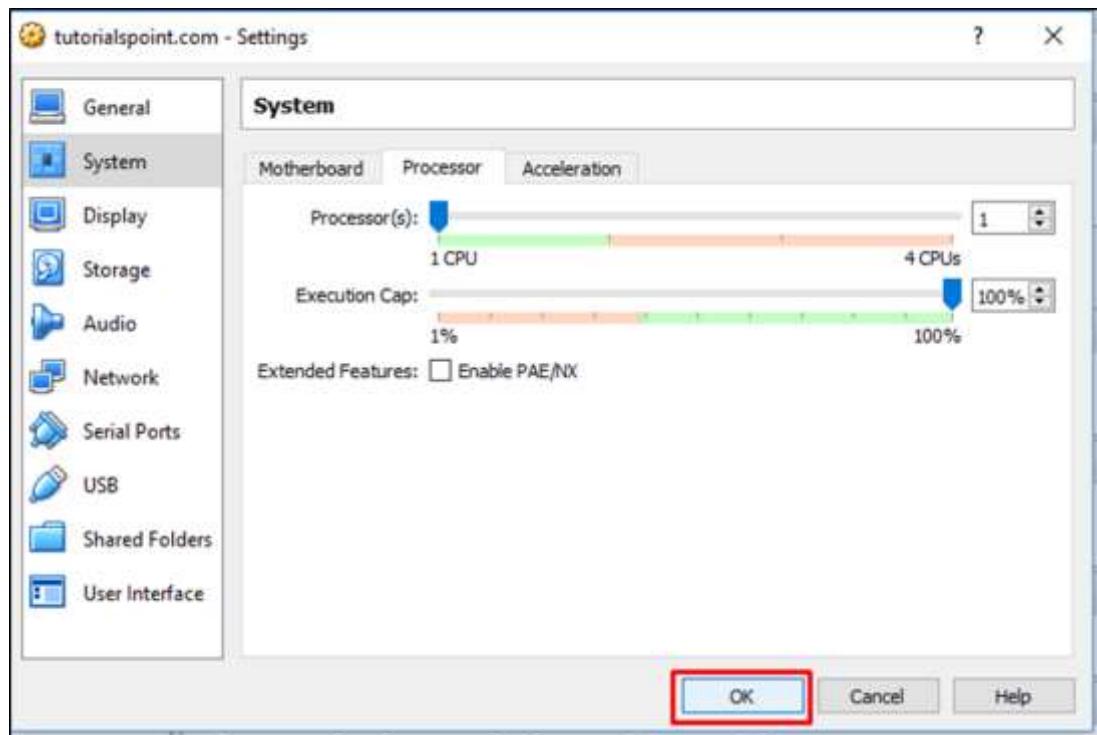
**Step 2:** Click on “System” on the left side tab, then click on the “Motherboard” tab. Move the arrow left or right to allocate the memory as shown in the screenshot below.



**Step 3:** To allocate processors, click on the “Processor” tab. Move the arrow left or right to allocate the number of processors as shown in the screenshot below.



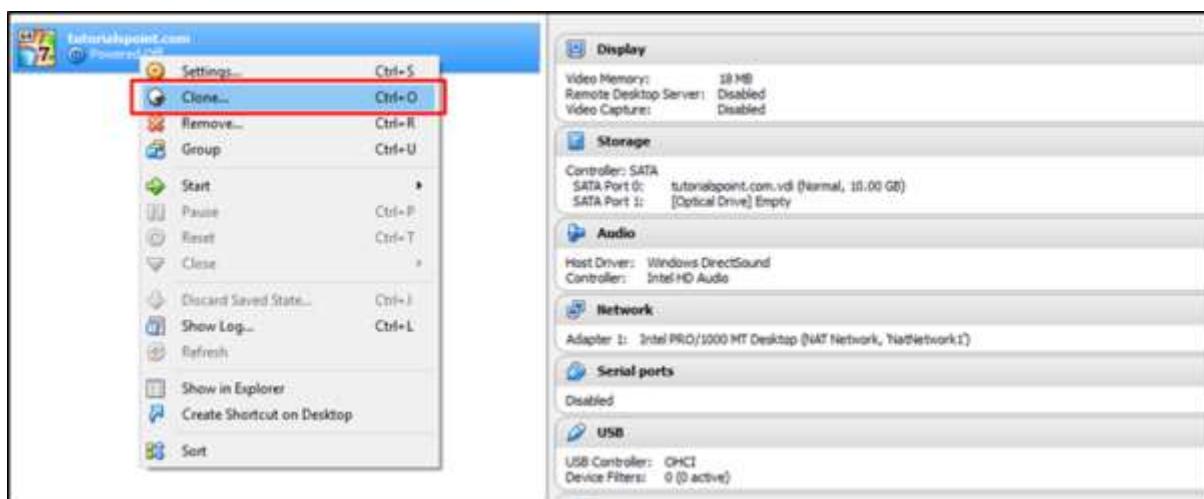
**Step 4:** After all those changes are done → click on “OK”.



## Duplicating a VM Using VirtualBox

To duplicate a virtual machine using VirtualBox, we should follow the steps given below.

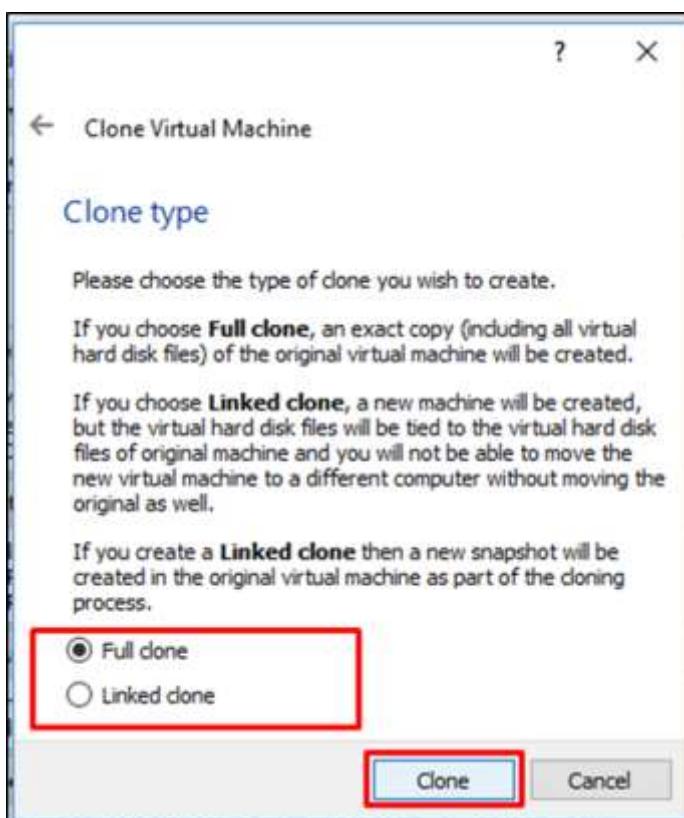
**Step 1:** To duplicate a VM that we created before, right click on the VM and select “Clone”. A wizard will open.



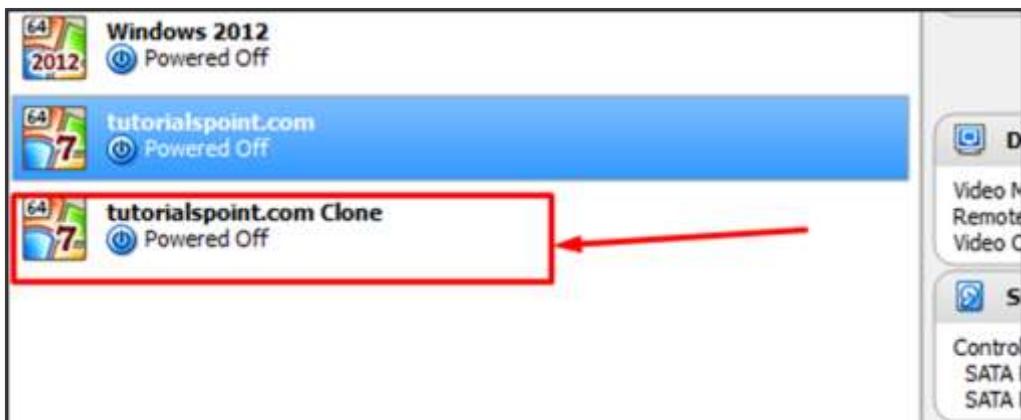
**Step 2:** Write the name of the cloning machine and click on “Next”.



**Step 3:** Select one of the options and Click on “Clone”.



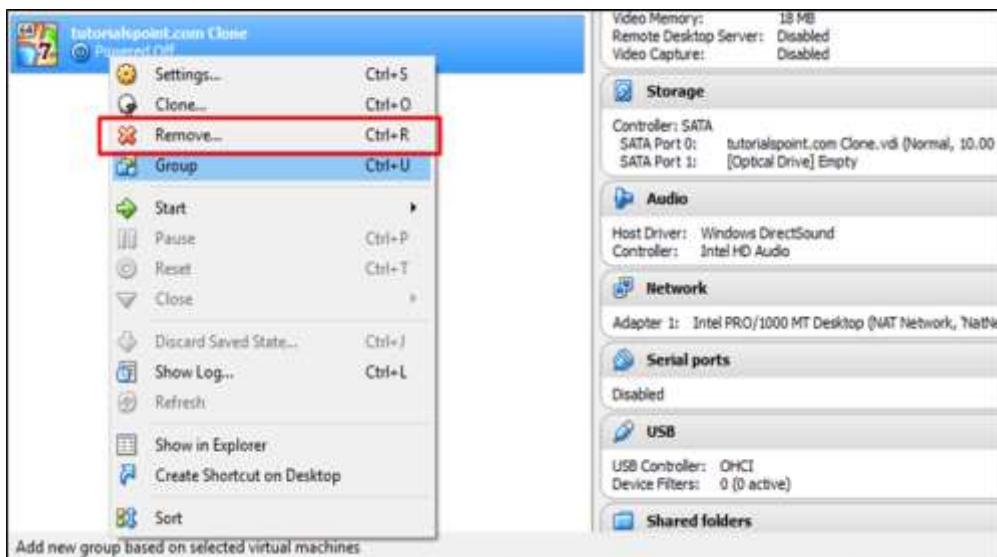
**Step 4:** The newly created VM will be as shown in the following screenshot.



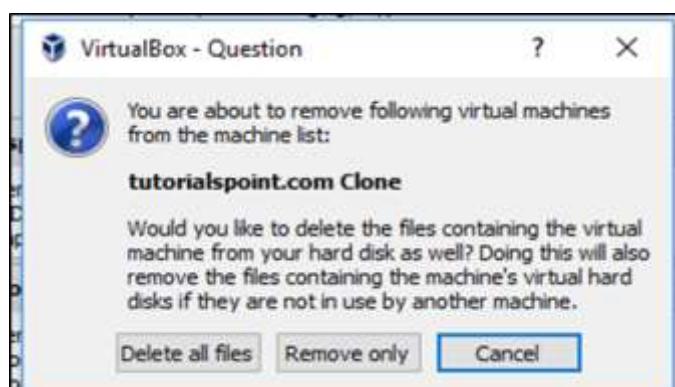
## Deleting a VM on VirtualBox

To delete a virtual machine on VirtualBox, we should follow the steps given below.

**Step 1:** To start with, we have to right click on the VM that we want to delete and then click on "Remove".



**Step 2:** To delete a virtual machine completely, select "Delete all files".



## 8. Virtualization – Openstack

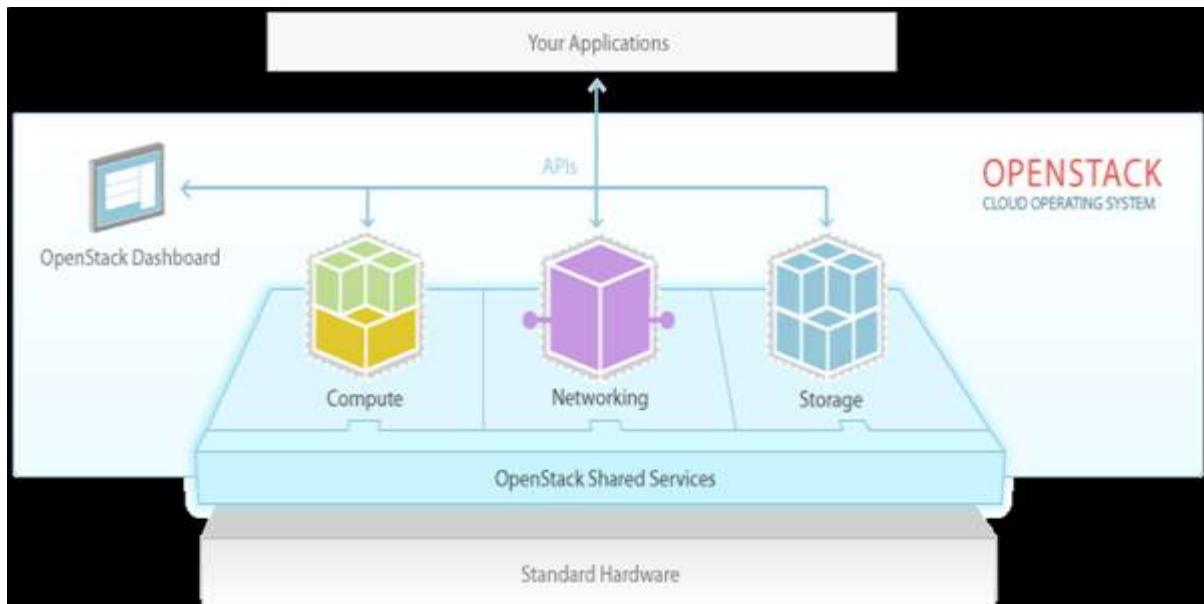
In this chapter, we will discuss regarding Openstack and its role in Virtualization.

### Understanding Openstack

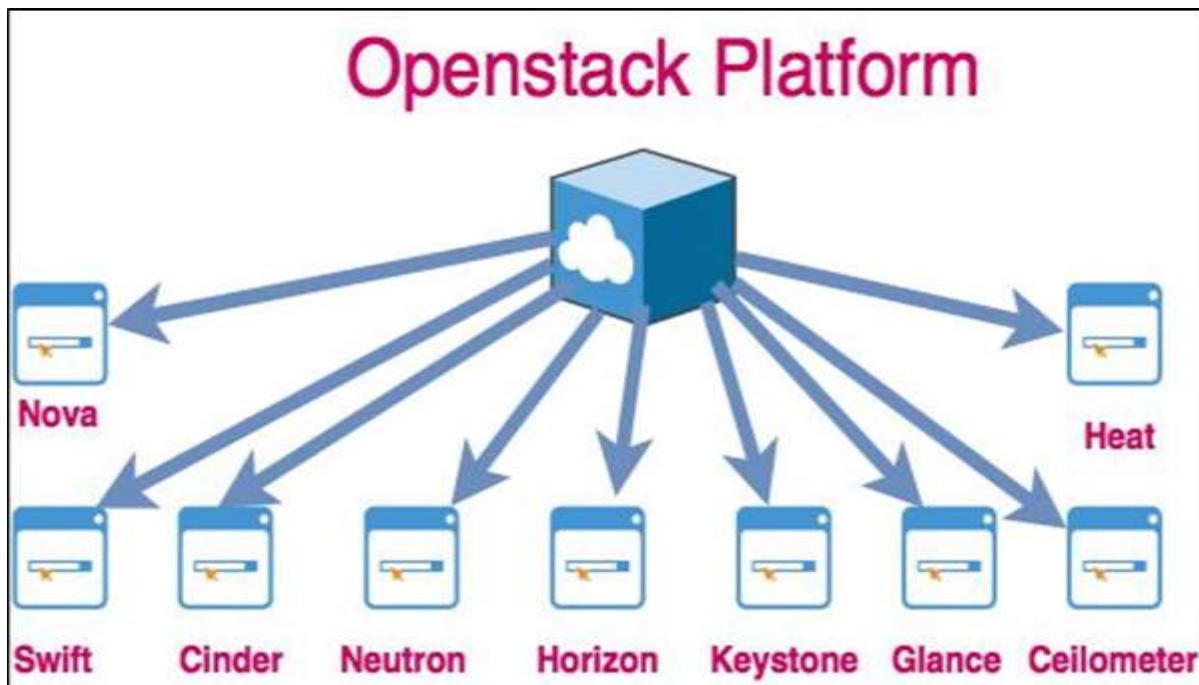
OpenStack is a software for building and managing cloud-computing platforms for public and private clouds. Openstack has one of the biggest communities. It is managed by the [OpenStack Foundation](#), a non-profit organization that oversees both development and community building. Its official webpage is – <https://www.openstack.org/>.

OpenStack is used to deploy virtual machines and other instances that handle different tasks for managing a cloud environment. OpenStack is considered as – Infrastructure as a Service (IaaS). Based on Openstack, please see the following schematic illustration, which describes how it is managed.

**Source:** <https://www.openstack.org/>



Openstack platform is managed by a web UI dashboard. It comprises of nine Core key components.



These key concepts are described in detail as follows:

- **Nova** is a computing engine. It is used for deploying and managing large numbers of virtual machines.
- **Swift** is a storage system for objects and files.
- **Cinder** is a block storage component. It accesses specific locations on a disk drive.
- **Neutron** provides the networking capability.
- **Horizon** is the dashboard of Openstack. It is the only graphical interface (WEB UI).
- **Keystone** provides identity services. It is essentially a central list of all the users.
- **Glance** provides image services to OpenStack. In this case, "images" refers to images (or virtual copies) of hard disks.
- **Ceilometer** provides telemetry services, which allow the cloud to provide billing services to individual users of the cloud.
- **Heat** allows developers to store the requirements of a cloud application in a file that defines what resources are necessary for that application.

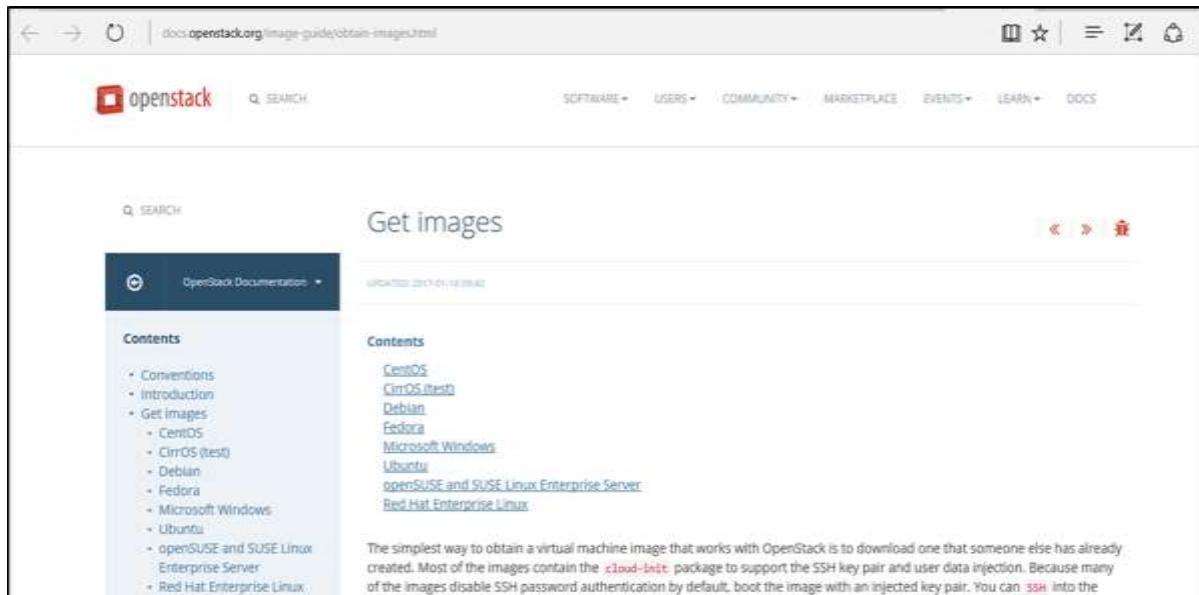
The release versions of Openstack are explained in the following illustration.

Release Name	Release Date	Components Included
Austin	October 2010	Nova, Swift
Bexar	February 2011	Nova, Glance, Swift
Cactus	April 2011	Nova, Glance, Swift
Diablo	September 2011	Nova, Glance, Swift
Essex	April 2012	Nova, Glance, Swift, Horizon, Keystone
Folsom	September 2012	Nova, Glance, Swift, Horizon, Keystone, Quantum, Cinder
Grizzly	April 2013	Nova, Glance, Swift, Horizon, Keystone, Quantum, Cinder
Havana	October 2013	Nova, Glance, Swift, Horizon, Keystone, Quantum, Cinder, Ceilometer, Heat
Icehouse	April 2014	Nova, Glance, Swift, Horizon, Keystone, Quantum, Cinder, Ceilometer, Heat, Trove
Juno	October 2014	Nova, Glance, Swift, Horizon, Keystone, Quantum, Cinder, Ceilometer, Heat, Trove, Sahara
Kilo	April 2015	Nova, Glance, Swift, Horizon, Keystone, Quantum, Cinder, Ceilometer, Heat, Trove, Sahara, Ironic

## Installing Openstack

As Openstack is an open source platform, there are many ways to install and deploy it through different software distributions. Each one of them adds their own value to the cloud operating system.

For a new system administrator, who wants to play with Openstack will get pre-installed ISO images of the Openstack platform. They can be downloaded from the following link – <http://docs.openstack.org/image-guide/obtain-images.html>



The screenshot shows a web browser displaying the OpenStack documentation at [docs.openstack.org/image-guide/obtain-images.html](http://docs.openstack.org/image-guide/obtain-images.html). The page title is "Get images". The left sidebar has a "Contents" section with links to "Conventions", "Introduction", "Get images" (which is expanded to show "CentOS", "CirrOS (test)", "Debian", "Fedora", "Microsoft Windows", "Ubuntu", "openSUSE and SUSE Linux Enterprise Server", and "Red Hat Enterprise Linux"), and "Red Hat Enterprise Linux". The main content area also has a "Contents" section with the same list of images. A note at the bottom states: "The simplest way to obtain a virtual machine image that works with OpenStack is to download one that someone else has already created. Most of the images contain the `cloud-init` package to support the SSH key pair and user data injection. Because many of the images disable SSH password authentication by default, boot the image with an injected key pair. You can `ssh` into the".

To install them manually, the main distributors are:

- **Ubuntu:** <https://www.ubuntu.com/cloud/openstack>
- **Red Hat:** <https://www.rdoproject.org>
- **Suse:** <https://www.suse.com/products/suse-openstack-cloud/>

We must understand that Openstack is for enterprise environment and to install it we should have the following hardware requirements.

## Installation Requirements

The installation requirements for Openstack are as follows:

- A minimum of 5 machines with the following roles:
  - 1 machine for the MAAS server.
  - 1 machine for the Autopilot.
  - 3 or more machines for the cloud:
    - At least one must have 2 NICs.
    - At least 3 must have 2 disks.
- A dedicated switch to create a private cloud LAN.
- Internet access through a router on that LAN.

For this laboratory, we have a virtual machine and we will install **Devstack**. It is designed for installation on a single laptop, PC or VM. It includes the raw upstream code for development evaluation. It is supported on –

- Ubuntu 14.04/16.04
- Fedora 23/24
- CentOS/RHEL 7
- Debian and
- OpenSUSE.

The link for this version is – <http://docs.openstack.org/developer/devstack/#install-linux>

## Installing Openstack on Ubuntu 14.04

---

For installing Openstack on Ubuntu 14.04, we should follow the steps given below.

**Step 1:** Download the installation script from the following link – <https://git.openstack.org/cgit/openstack-dev/devstack>.

Use this command – **git clone https://git.openstack.org/openstack-dev/devstack**

```
:-$ git clone https://git.openstack.org/openstack-dev/devstack

cloning into 'devstack'...
remote: Counting objects: 26862, done.
remote: Compressing objects: 100% (12384/12384), done.
remote: Total 26862 (delta 19136), reused 21637 (delta 14215)
Receiving objects: 100% (26862/26862), 5.30 MiB | 1.22 MiB/s, done.
Resolving deltas: 100% (19136/19136), done.
Checking connectivity... done. ◀
```

**Step 2:** Browse the folder Devstack by keying in the following command – **\$ cd devstack.**

```
-6300-MT:~$ cd devstack/
```

**Step 3:** Execute the file **stack.sh** with the following command **./stack.sh** and installation process will continue as shown below:

```
++ set +o
* XTRACE='set -o xtrace'
* set +o xtrace

#####
ENTER A PASSWORD TO USE FOR THE DATABASE.
#####
This value will be written to your localrc file so you don't have to enter it
again. Use only alphanumeric characters.
If you leave this blank, a random default value will be used.
Enter a password now:
shivam
++ get_database_type
++ [[ -n '' ]]
++ echo mysql
+ BASE_SQL_CONN=mysql://root:shivam@127.0.0.1
+ return 0
+ echo 'Using mysql database backend'
Using mysql database backend
+ RABBIT_USERID=stackrabbit
+ is_service_enabled rabbit
+ set +o
+ grep xtrace
+ local 'xtrace=set -o xtrace'
+ set +o xtrace
+ return 0
+ RABBIT_HOST=192.168.2.3
+ read_password RABBIT_PASSWORD 'ENTER A PASSWORD TO USE FOR RABBIT.'
+ grep xtrace
+ set +o
+ XTRACE='set -o xtrace'
+ set +o xtrace

#####
ENTER A PASSWORD TO USE FOR RABBIT.
#####
This value will be written to your localrc file so you don't have to enter it
again. Use only alphanumeric characters.
```

**Step 4:** Enter your password.

```
#####
ENTER A PASSWORD TO USE FOR RABBIT.
#####
This value will be written to your localrc file so you don't have to enter it
again. Use only alphanumeric characters.
If you leave this blank, a random default value will be used.
Enter a password now:
[REDACTED]
```

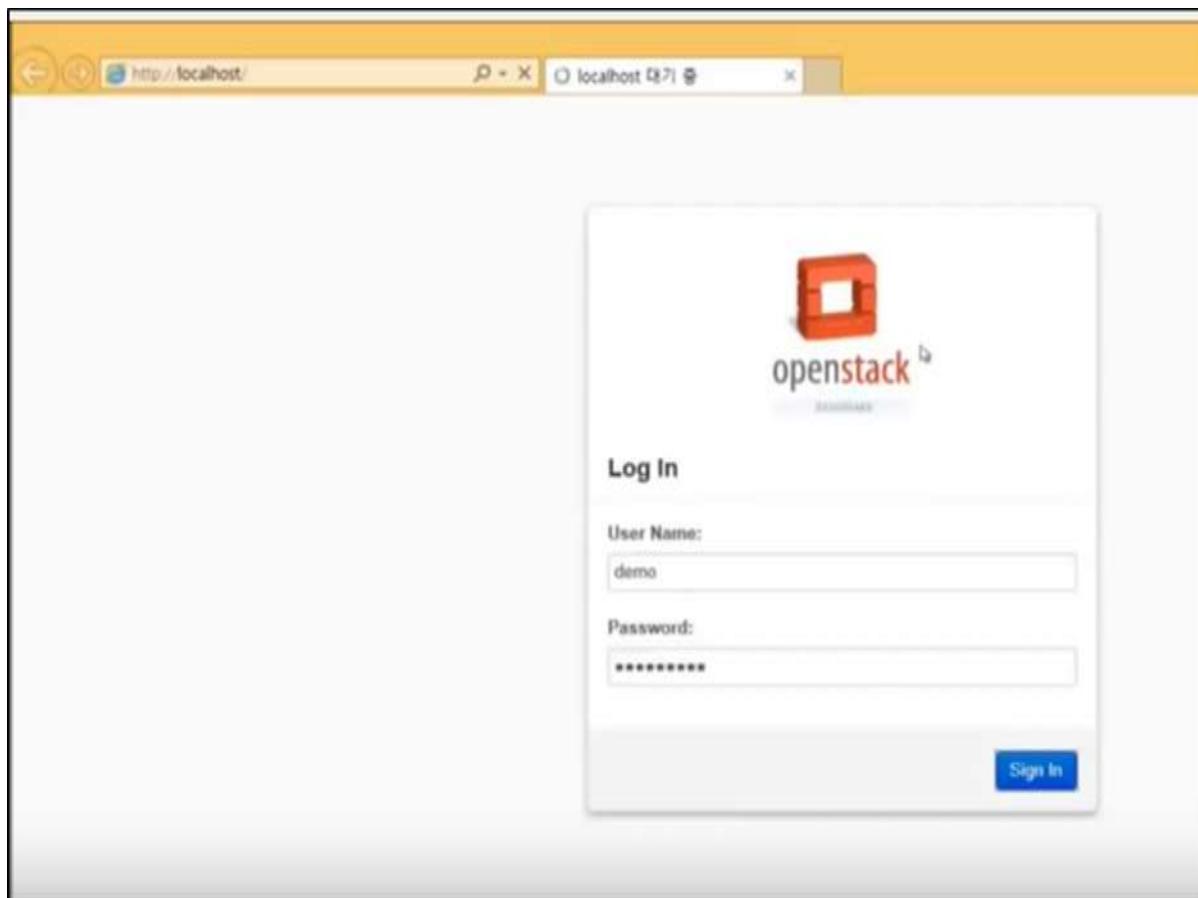
**Step 5:** Now it will take about 15-20 minutes to finish the installation process, while it downloads and installs all the libraries.

```
Get:2 http://in.archive.ubuntu.com/ubuntu/ trusty/main libgbroker3 amd64 4.8.2-19ubuntu1 [250 kB]
Get:3 http://in.archive.ubuntu.com/ubuntu/ trusty-updates/main libgbroker4 amd64 1:1.2+dfsg-2ubuntu5.1 [531 kB]
Get:4 http://in.archive.ubuntu.com/ubuntu/ trusty-updates/main libkadm5clnt-mit9 amd64 1:1.2+dfsg-2ubuntu5.1 [36.1 kB]
Get:5 http://in.archive.ubuntu.com/ubuntu/ trusty-updates/main libkdb5-7 amd64 1:1.2+dfsg-2ubuntu5.1 [36.2 kB]
Get:6 http://in.archive.ubuntu.com/ubuntu/ trusty-updates/main libkadm5srv-mit9 amd64 1:1.2+dfsg-2ubuntu5.1 [56.3 kB]
Get:7 http://in.archive.ubuntu.com/ubuntu/ trusty-updates/main mysql-common-all 5:5.5.43-0ubuntu0.14.04.1 [19.8 kB]
Get:8 http://in.archive.ubuntu.com/ubuntu/ trusty-updates/main libmysqclient18 amd64 5:5.5.43-0ubuntu0.14.04.1 [59.1 kB]
Get:9 http://in.archive.ubuntu.com/ubuntu/ trusty/main libpcrecpp0 amd64 1:8.31-2ubuntu2 [14.5 kB]
Get:10 http://in.archive.ubuntu.com/ubuntu/ trusty/main libexpat1-dev amd64 2:1.0.0-4ubuntu1 [155 kB]
Get:11 http://in.archive.ubuntu.com/ubuntu/ trusty/main libpython2.7-dev amd64 2:27.1.0-8 [22.0 kB] + 84.1 [1422 kB]
Get:12 http://in.archive.ubuntu.com/ubuntu/ trusty-updates/main librados2 amd64 0.8.9-0ubuntu0.14.04.1 [1427 kB]
Get:13 http://in.archive.ubuntu.com/ubuntu/ trusty-updates/main libsysfs2 amd64 2.1.0+repack-3ubuntu3 [19.3 kB]
Get:14 http://in.archive.ubuntu.com/ubuntu/ trusty/main libtbyaml-0-2 amd64 0.1.4-3ubuntu3.1 [48.1 kB]
Get:15 http://in.archive.ubuntu.com/ubuntu/ trusty/main libpython2.7 amd64 2.7.10-0ubuntu1 [156 kB]
Get:16 http://in.archive.ubuntu.com/ubuntu/ trusty-updates/main librbd1 amd64 0.8.9-0ubuntu0.14.04.1 [318 kB]
Get:17 http://in.archive.ubuntu.com/ubuntu/ trusty-updates/main libsysfs2 amd64 2.1.0+repack-3ubuntu3 [19.3 kB]
Get:18 http://in.archive.ubuntu.com/ubuntu/ trusty-updates/main libtbyaml-0-2 amd64 0.1.4-3ubuntu3.1 [48.1 kB]
Get:19 http://in.archive.ubuntu.com/ubuntu/ trusty/main libtbyaml-0-2 amd64 0.1.4-3ubuntu3.1 [48.1 kB]
Get:20 http://in.archive.ubuntu.com/ubuntu/ trusty-updates/main libtcmpp0 amd64 4:4.5+1-2ubuntu1.2 [355 kB]
Get:21 http://in.archive.ubuntu.com/ubuntu/ trusty/main libbridge-utils amd64 1:5+0ubuntu2 [29.2 kB]
Get:22 http://in.archive.ubuntu.com/ubuntu/ trusty-updates/main libtcmpp0 amd64 4:4.5+1-2ubuntu1.2 [355 kB]
Get:23 http://in.archive.ubuntu.com/ubuntu/ trusty/main libuniverse-conctrack amd64 1:1.4.1-1ubuntu1 [25.1 kB]
Get:24 http://in.archive.ubuntu.com/ubuntu/ trusty/main dnsmasq-utils amd64 2:0.88-1 [7972 kB]
Get:25 http://in.archive.ubuntu.com/ubuntu/ trusty/main liblvm2api0 amd64 2:2.0.8-1 [77.5 kB]
Get:26 http://in.archive.ubuntu.com/ubuntu/ trusty/universe-dstat-all 0.7.2-3build1 [6812 kB]
Get:27 http://in.archive.ubuntu.com/ubuntu/ trusty/main libctables amd64 2.0.10-4ubuntu1 [77.5 kB]
Get:28 http://in.archive.ubuntu.com/ubuntu/ trusty/universe fping amd64 3.8-5 [38.7 kB] + 1050 kB
Get:29 http://in.archive.ubuntu.com/ubuntu/ trusty/main libbstdc++-4.8-dev amd64 4:4.8.2-19ubuntu1 [1050 kB]
Get:30 http://in.archive.ubuntu.com/ubuntu/ trusty/main g++-4.8 amd64 4.8.2-19ubuntu1 [7038 kB]
Get:31 http://in.archive.ubuntu.com/ubuntu/ trusty/main g++-4.8-2-1ubuntu0.1 [1490 kB]
Get:32 http://in.archive.ubuntu.com/ubuntu/ trusty-updates/main libcdt5 amd64 2:36.0-0ubuntu3.1 [23.3 kB]
Get:33 http://in.archive.ubuntu.com/ubuntu/ trusty-updates/main libcgrope amd64 2:36.6-0ubuntu3.1 [44.1 kB]
Get:34 http://in.archive.ubuntu.com/ubuntu/ trusty-updates/main libpathplan4 amd64 2:36.0-0ubuntu3.1 [26.3 kB]
Get:35 http://in.archive.ubuntu.com/ubuntu/ trusty-updates/main libgvce amd64 2:36.0-0ubuntu3.1 [576 kB]
Get:36 http://in.archive.ubuntu.com/ubuntu/ trusty-updates/main libgvpr2 amd64 2:36.0-0ubuntu3.1 [169 kB]
Get:37 http://in.archive.ubuntu.com/ubuntu/ trusty-updates/main graphviz amd64 2:36.0-0ubuntu3.1 [461 kB]
Get:38 http://in.archive.ubuntu.com/ubuntu/ trusty/main javascript-common-all 1.1 [6066 kB] + 38.3 kB
Get:39 http://in.archive.ubuntu.com/ubuntu/ trusty-updates/main concerr-dev amd64 2:1.1.42.9-3ubuntu1.2 [38.3 kB]
Get:40 http://in.archive.ubuntu.com/ubuntu/ trusty-updates/main krb5-multidev amd64 1.12+dfsg-2ubuntu5.1 [111 kB]
Get:41 http://in.archive.ubuntu.com/ubuntu/ trusty/main libblas3 amd64 1:2.20110419-7 [215 kB]
Get:42 http://in.archive.ubuntu.com/ubuntu/ trusty/main libjs-jquery-all 1.7.2+dfsg-2ubuntu1 [78.8 kB]
Get:43 http://in.archive.ubuntu.com/ubuntu/ trusty/main libjs-jquery-metadata-all 0.2-0 [8856 kB]
Get:44 http://in.archive.ubuntu.com/ubuntu/ trusty/main libjs-jquery-tablesorter-all 0.2-2 [64.0 kB]
Get:45 http://in.archive.ubuntu.com/ubuntu/ trusty/main liblapack3 amd64 3:5.0.2-2ubuntu1 [1730 kB]
Get:46 http://in.archive.ubuntu.com/ubuntu/ trusty/main zlib1g-dev amd64 1:1.2.8+dfsg-1ubuntu1 [183 kB] + 84.1 [1422 kB]
```

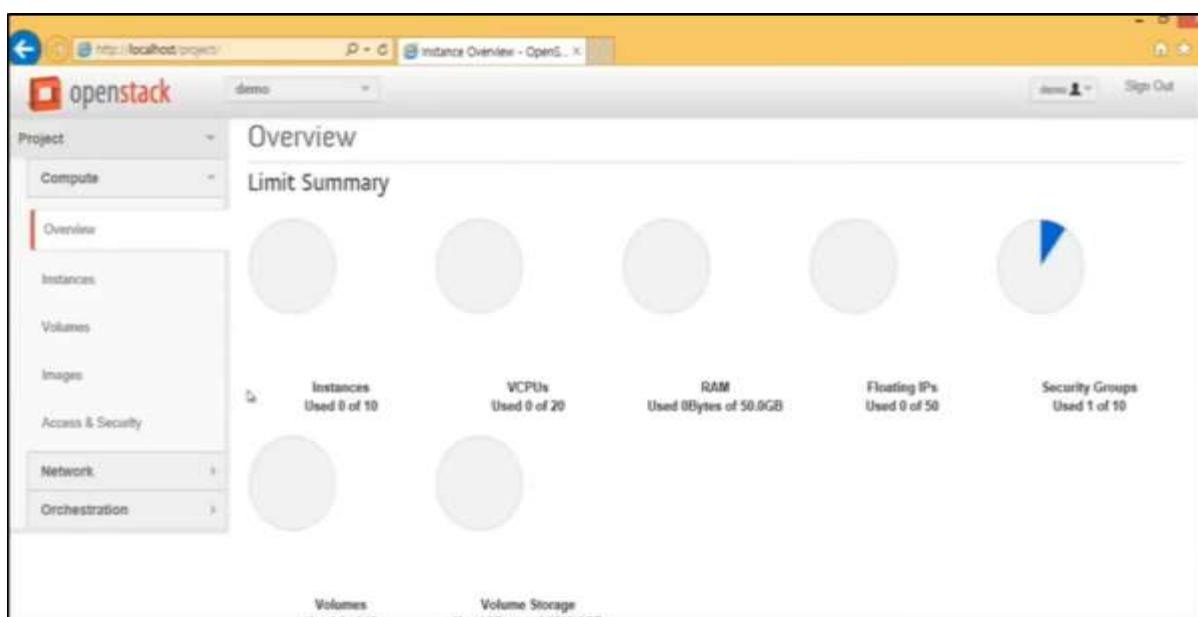
**Step 6:** At the end of the installation, we will see the IP of the host, URL for managing it, username and password to administrate it.

```
This is your host ip: 192.168.2.3
Horizon is now available at http://192.168.2.3/
Keystone is serving at http://192.168.2.3:5000/
The default users are: admin and demo
The password: [REDACTED]
2015-04-28 14:33:41.286 | stack.sh completed in 1825 seconds.
```

**Step 7:** To sign in, you have to type in the browser: Host IP, username and password that we entered during the installation.



**Step 8:** In the main dashboard, you will see “Instances” overview.



**Step 9:** To create new instances or in simple words to create Virtual machines click on "Instances".



**Step 10.** Click on "Launch Instances".



**Step 11:** Fill in all the fields as shown below.

**Launch Instance**

Details \* Access & Security \* Networking \* Post-Creation Advanced Options

Availability Zone: nova

Instance Name: test

Flavor: m1.micro

Instance Count: 1

Instance Boot Source: Boot from image

Image Name: cirros-0.3.1-x86\_64-uec (24.0 MB)

**Flavor Details**

Name	m1.micro
VCPUs	1
Root Disk	0 GB
Ephemeral Disk	0 GB
Total Disk	0 GB
RAM	128 MB

**Project Limits**

Number of Instances	Green
Number of VCPUs	Green

**Step 12:** We will see the instance created as shown in the following screenshot.

demo

Instances

Instances Filter  Filter  + Launch Instance Set Default Instances Remove Instances

Instance Name	Image Name	IP Address	Size	Key Pair	Status	Availability Zone	Task	Power State	Uptime	Actions
test	cirros-0.3.1-x86_64-uec		m1.micro (128MB)		Build	nova	Scheduling	No State	0 minutes	<input type="button" value="Associate Floating IP"/> More

Displaying 1 item

# 9. Virtualization – Preparing the Infrastructure

Virtualization, at least at the server level, has been in use for some time. Since, the concept has expanded to the User, Application, Network, Security, Storage, and of course to Desktop Virtualization (VDI) as well. Therefore, to roll out in all these concepts, we have to prepare an infrastructure before, which is divided in some key components.

The First one is **Hypervisor** and we have discussed it in the earlier chapters. Some important specifications needed for this job are – Processors, Memory and Networking Modes, etc.

## Networking - LAN and WAN Optimization

VDI can be very resource intensive – this includes traffic over the wire. Having a good **core-switching infrastructure** will help alleviate this pain by allowing the administrators to create certain rules and policies revolving around traffic flow.

Setting up **QoS metrics** for VDI-specific traffic can help remove congestion and ensure that the right traffic has the proper amount of priority. As for the traffic that is leaving the data center, knowing where the user is located and optimizing their experience based on certain criteria, becomes very important.

Another significant aspect is storage. Large organizations will have numerous storage controllers. At the same time, some smaller organizations will be using only one. Regardless of the amount of storage controllers available, they need to be sized properly for VDI.

To prevent boot and processing storms, organizations must look at IOPS requirements for their images. To alleviate processing pains, administrators can look at **Flash Technologies (NetApp, Fusion-IO, XtremIO)** or **SSD Technologies (Violin, Nimbus)** to help offload that kind of workload. Furthermore, intermediary platforms like **Atlantis ILIO** run on top of a virtual machine that utilizes massive amounts of RAM as the key storage repository.

## Understanding Different File Systems

---

File systems are varied depending on their functions. Some of the most common ones are listed below:

- Virtual Machine File System- VMFS
- Network File System- NFS
- New Technology File System- NTFS
- Raw Device Mapping- RDM

Let us discuss each of these in detail.

### Virtual Machine File System

**VMFS** is a file system proprietary to VMware. It is a clustered file system and it can be mounted on multiple servers simultaneously. This helps every host to connect to the any file system at the same time, which is expected from a proprietary system preferred by

VMware. Most of the VMware hypervisors will work with other file systems, but the default choice and the preferred choice is VMFS.

## **Network File System**

NFS is a system that was originally developed by Sun, but is now an Open Standard system. Used commonly in the UNIX and Linux world. It is a distributed file system. It can be mounted on one server and the network will be used to share information to multiple machines.

## **New Technology File System**

NTFS is the standard file structure for the Windows NT operating system. It is used for retrieving and storing files on the hard disk.

## **Raw Device Mapping**

RDM helps any file in a virtual machine file system to act as a proxy for any raw device. It allows a VM to access as well as use the storage device.

## **Choosing Between Different Types of Storage**

---

All the storage devices are divided into three categories, which are –

- Direct-Attached Storage
- Network-Attached Storage
- Storage-Area Network

Let us understand each of these in detail.

### **Direct-Attached Storage**

DAS is your local hard drive. We can have one or more local hard drives in every machine. It is mostly used for small-virtualized systems. This solution is appropriate for a small amount of guest machines. It is the cheapest and the easiest method of storage. You can connect an external hard drive too with this method, load any guest machines onto that external hard drive, and that is just a quick and easy way to get rolling.

For example – A server that has local hard discs on it. That type of hypervisor can hold not more than 10-20 VM machines. A sample device can be a HP Server with eight local hard discs as shown in the following illustration.



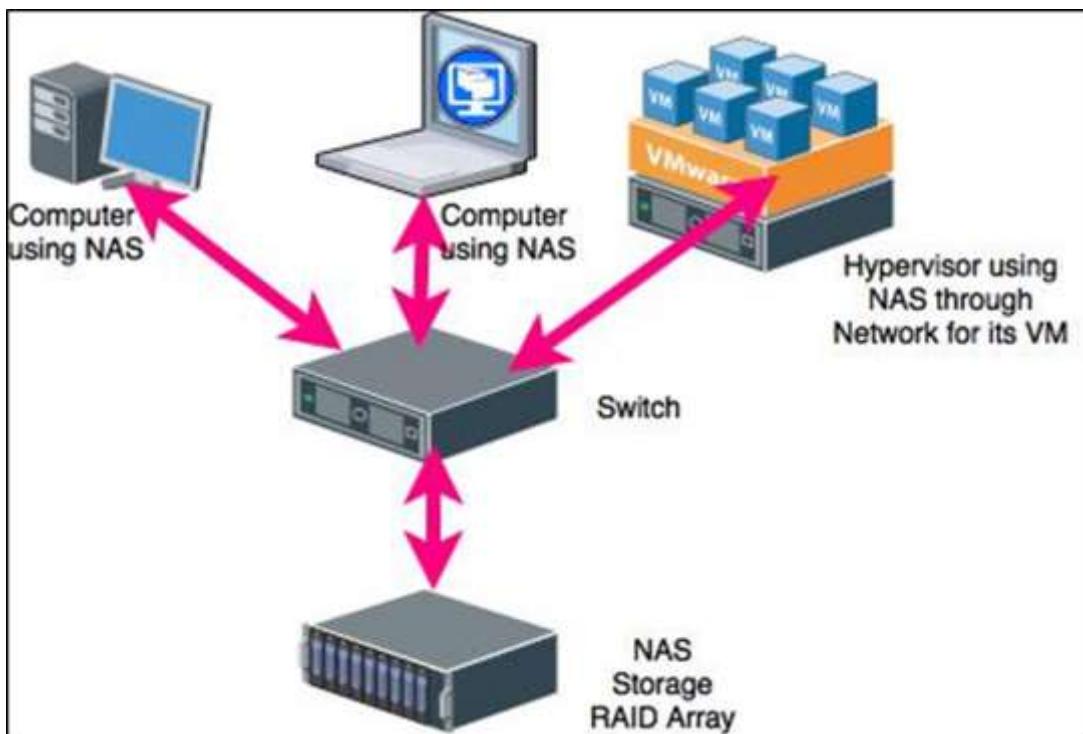
## Network-attached Storage

NAS is "Any server that shares its own storage with others on the network and acts as a file server in the simplest form". Network Attached Storage shares files over the network. Some of the most significant protocols used are **SMB**, **NFS**, **CIFS**, and **TCP/IP**. When you access files on a file server on your windows system, it is NAS.

NAS will be using an Ethernet connection for sharing files over the network. The NAS device will have an IP address and then will be accessible over the network through that IP address. Biggest providers of NAS are **QNAP** and **Lenovo**.



The following illustration shows how NAS works.

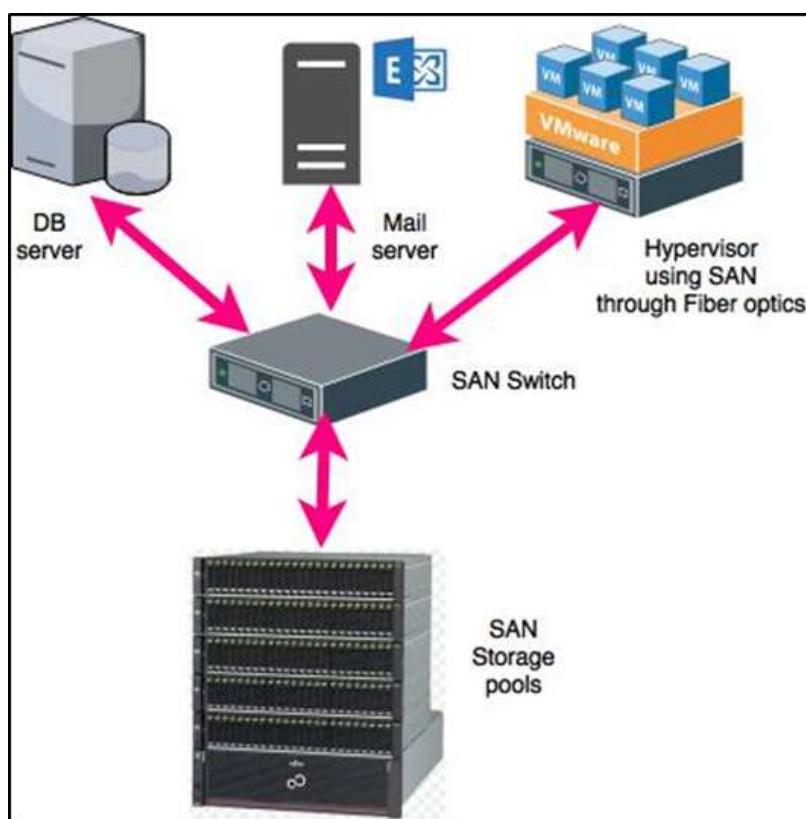


## Storage-area Network

SANs allow multiple servers to share a pool of storage; making it appear to the server as if it were local or directly attached storage. A dedicated networking standard, **Fibre Channel**, has been developed to allow blocks to be moved between servers and storage at high speed. It uses dedicated switches and a fiber-based cabling system, which separates it from the day-to-day traffic traversing the busy enterprise network. While the well-established SCSI protocol enables communication between the servers' host bus adaptors and the disk system.



The following illustration shows how a SAN switch operates.



# 10. Virtualization – Troubleshooting

In this chapter, we will discuss troubleshooting for network communication and for slow performance.

## Troubleshooting Network Communication

---

In a virtual environment, to troubleshoot a network connectivity, we should follow some steps in order to find a resolution.

### Network Communication Indications

Some of the symptoms that we have are as follows:

- You cannot connect to the Internet.
- There is no network connectivity to or from a single virtual machine.
- Virtual machines fail to connect to the network.
- Cannot get an IP.
- A TCP/IP connection fails to and from a single virtual machine.

### Network Communication Errors

You can get the following errors as well:

- Destination Host Unreachable
- Network error
- Connection Refused
- Network cable is unplugged
- Ping request could not find host
- Please check the name and try again
- Unable to resolve target system name, etc.

### Network Communication Resolutions

The resolutions for the above symptoms and errors are as follows –

- Make sure that the **Port Group Name** is associated with the virtual machine's network adapter, which exists in the switch or in the Virtual Distributed Switch. Also, ensure that it is spelt correctly.
- Make sure that there are enough **storage capacities** in your virtual machine sometimes it affects the connectivity.
- Verify that the **virtual network adapter** is present and connected.

- Verify that the networking within the virtual machine's **guest operating system** is correct.
- Verify that the **TCP/IP stack** is functioning correctly.
- If this virtual machine was converted from a physical system, verify that there are no **hidden network adapters** present. Because it can have hidden static routes.
- Verify that the **vSwitch** has enough ports for the virtual machine.
- Verify that the virtual machine is configured with two **vNICs** to eliminate a NIC or a physical configuration issue.
- Confirm that your virtual machine's **firewall** is not blocking the Internet access.
- Confirm that your virtual machine's **anti-virus program** is not blocking the Internet access.
- Ensure that the **network adapter** is enabled.
- Shut down the virtual machine and then restart your **Host Machine**.
- Removing and re-adding **virtual network card adaptor**.

## **Troubleshooting Slow Performance**

---

Check if your CPU load is high. You can click on "CPU". This will show you the amount of CPU the VM is consuming. If it is very high, you may consider adding some more vCPUs. This should be done after ensuring that the physical host has more cores available than what you are going to configure inside the VM. We should also consider whether the applications inside the VM are actually able to utilize multiple vCPUs or not.

### **Check the Memory**

Memory could be a serious limit on VM performance as well. If you do not configure enough memory, the VM will usually respond by starting to swap its memory pages to disk .If your virtual machine is using more than 2/3rd of the memory and then we should allocate more.

### **See Disk Alignment**

For any pre-Windows 7, pre-Windows 2008 Server or older Linux based systems, your disks may be misaligned. Misalignment may cause quite a performance hit, especially when your storage underneath does not have many IOPS to spare.

It is important to format the virtual disks to a specific format or block size according to the application needs. For example – The database of a Microsoft SQL 2005 server is generally put on an NTFS that has a block size of 64KB.

If the virtual machine has a performance issue at some point in time, you need to check the virus scanner. Not only on the impacted virtual machine, but also on the other virtual machines as well.

If you do a P2V – a virtual machine (meaning you convert a physical machine to a VM) and you do not "clean up" afterwards, there may be a lot of unused drivers and even applications.

# 11. Virtualization – Backing Up, Restoring & Migrating VM

In this chapter, we will discuss how to back up, restore and migrate a virtual machine.

## Duplicating a VM

To duplicate or clone a machine means making an exact copy of it. Most of the hypervisors support this feature. By duplicating a machine, we copy down every detail, including the name of the machine and the different network addresses attached to the machine.

Duplicating a machine and putting it in to function is not always the best option because a duplicate name or IP in network can be a problem. We make duplication generally for backup purposes.

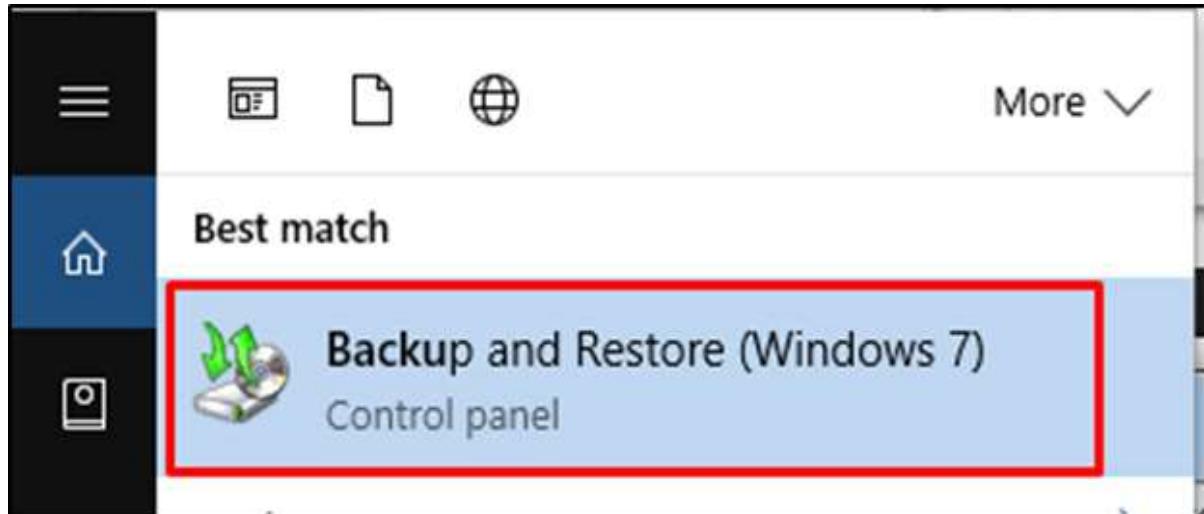
Most hypervisors can clone while the machine is turned off. If the hypervisor accepts to clone while it is on, it is recommended to turn it off, because the process can crash the machine. In practice, we have discussed “How cloning is done in VMware Workstation”, please refer to the previous chapters.

## Backing Up and Recovering a VM

There are three methods for backing up virtual machines.

### Method 1

The most common one is to install traditional backup software on the guest VM. If Windows OS is used on our VM, we can use “Backup and restore” to back up the machine, which is found in the “Control Panel”.

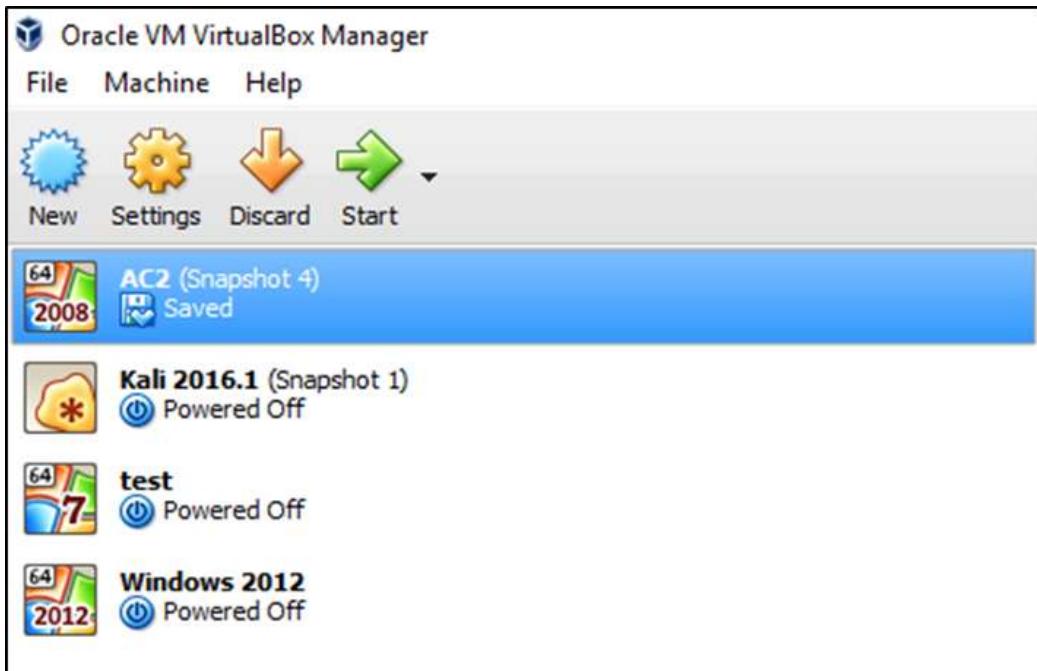


For Linux OS, we can use many open source tools depending on our needs, like “Bacula”, “rsync”, etc.

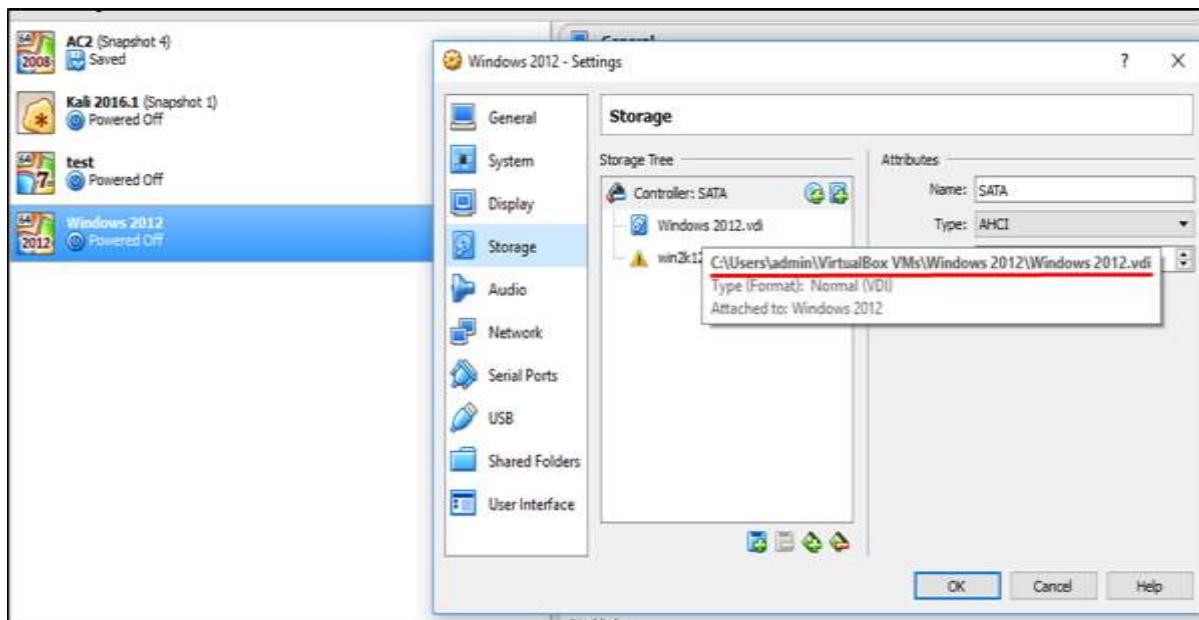
## Method 2

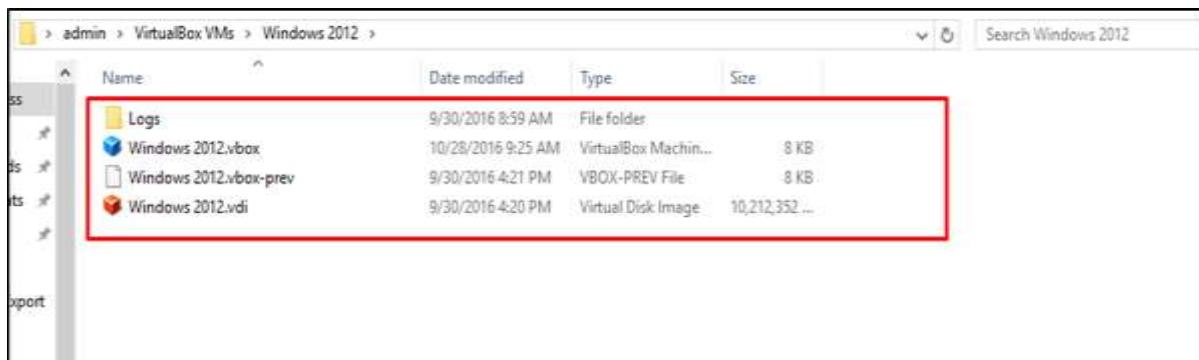
Another strategy or method is to copy all of the files that define a VM. Therefore, we will have to go out and find all of the individual files that define our virtual machine and copy them to an alternate location. Some of these files are going to be large.

Example: Here, we have created several VM machines with VirtualBox as shown in the following illustration. Their names are – “AC2”, “Kali2016.1”, “test”, “Windows 2012”.



To find the files that we have to copy or to backup, we have to right click on the VM machine. Go to “Storage” then move your mouse over the virtual HDD and it will show the full path where the VDI files are found.

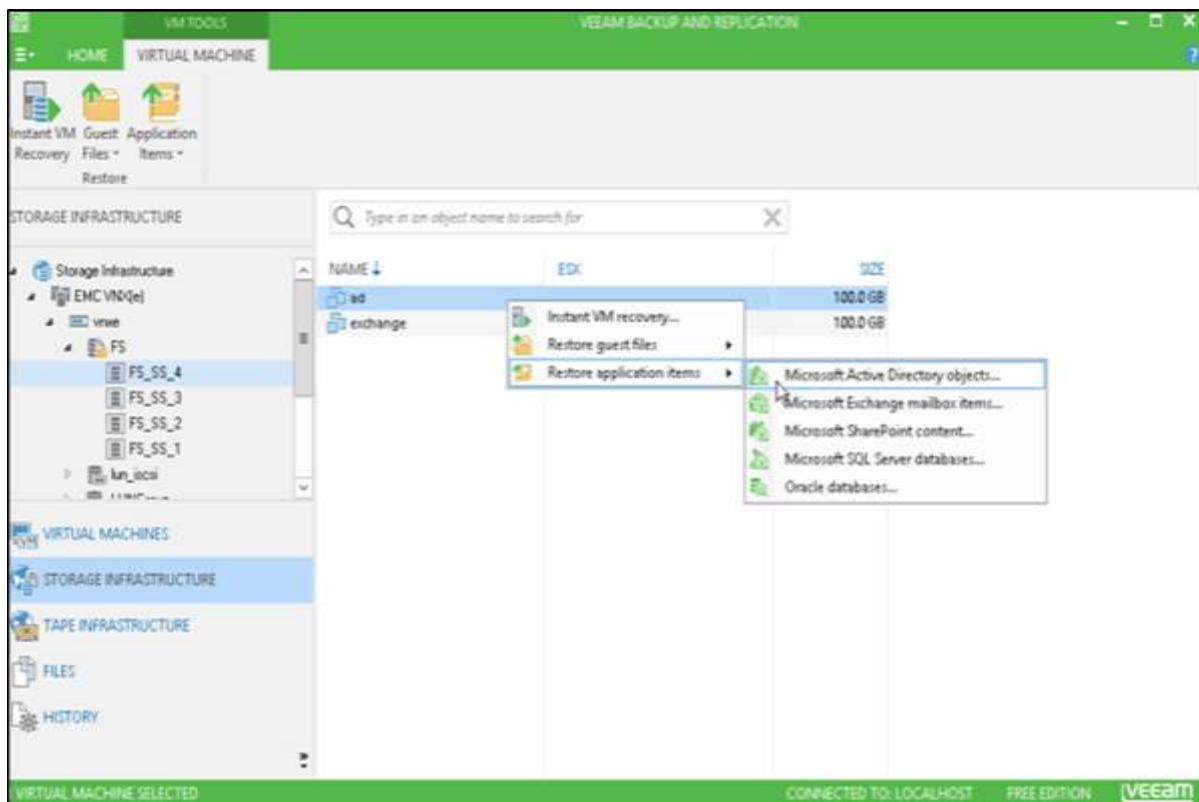




We will save all these files to another location.

### Method 3

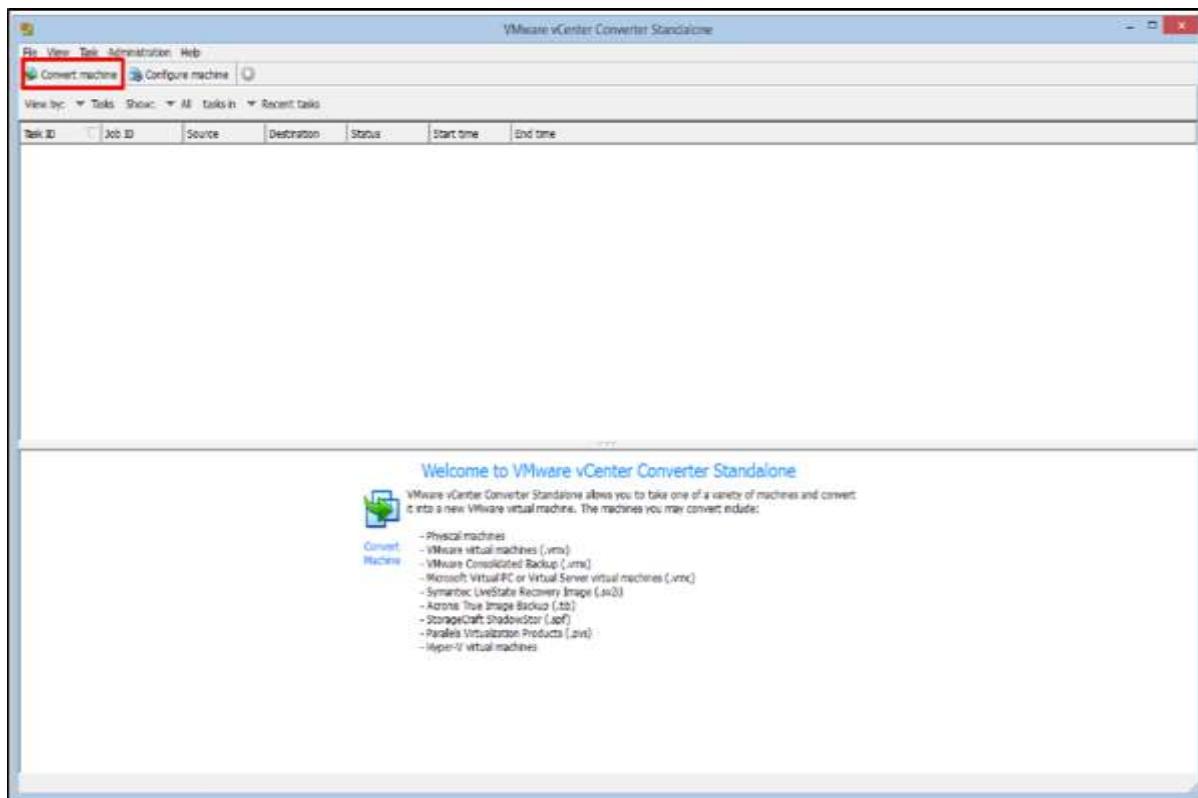
The third option to backup and restore VM machines is to use third party software. One of the best is **VEEAM**, which can be found on the following URL – <https://www.veeam.com/>



## Converting a Physical Server into a Virtual Server

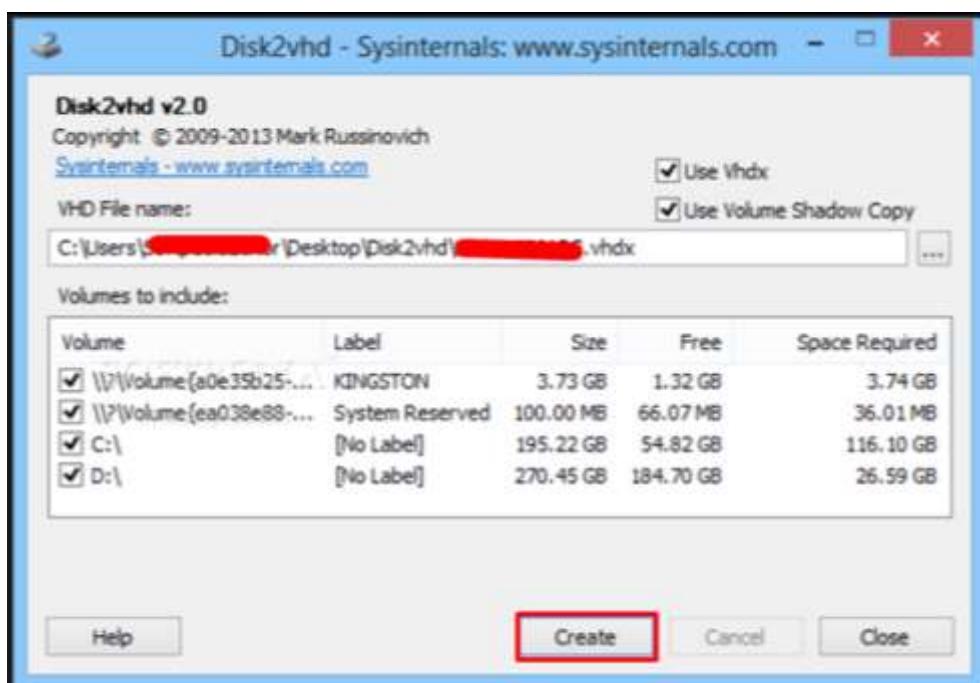
In this section, we will see how to convert a physical machine into a virtual machine. This is often called as **P2V** in many literatures.

VMware puts out a product that is called **vCenter Converter**, which will convert from a physical machine specifically into a VMware virtual machine. The software can be downloaded from – <http://www.vmware.com/products/converter.html>



Microsoft has a product called **Disk2vhd**, which will convert a physical hard drive into a VHD formatted virtual hard drive. It can be downloaded from the following link – <https://technet.microsoft.com/en-us/sysinternals/ee656415.aspx>

We just have to install the software on the physical server and click “Create” as shown in the screenshot below. A VHDX file will be created which could be imported in a Hypervisor.



Both of these products will convert machines, while the server is running and is free. All the vendors of hypervisors have some P2V tool and they are typically free. From the vendor's point of view, they would very much like you to convert your physical machines into virtual machines that are optimized for their hypervisor.

## **Converting a Virtual Server into a Physical Server**

---

To convert a virtual server to a physical server also commonly called as V2P is certainly less common than a P2V conversion. However sometimes, it is needed in development-based environments. It does happen where a product needs to be tested in the virtual server than to a physical server, or to clone a production machine and move it to test.

Hypervisor vendors do not offer such a tool. However, you have to request the hardware vendor, if they could offer such tools.



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  
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

# 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

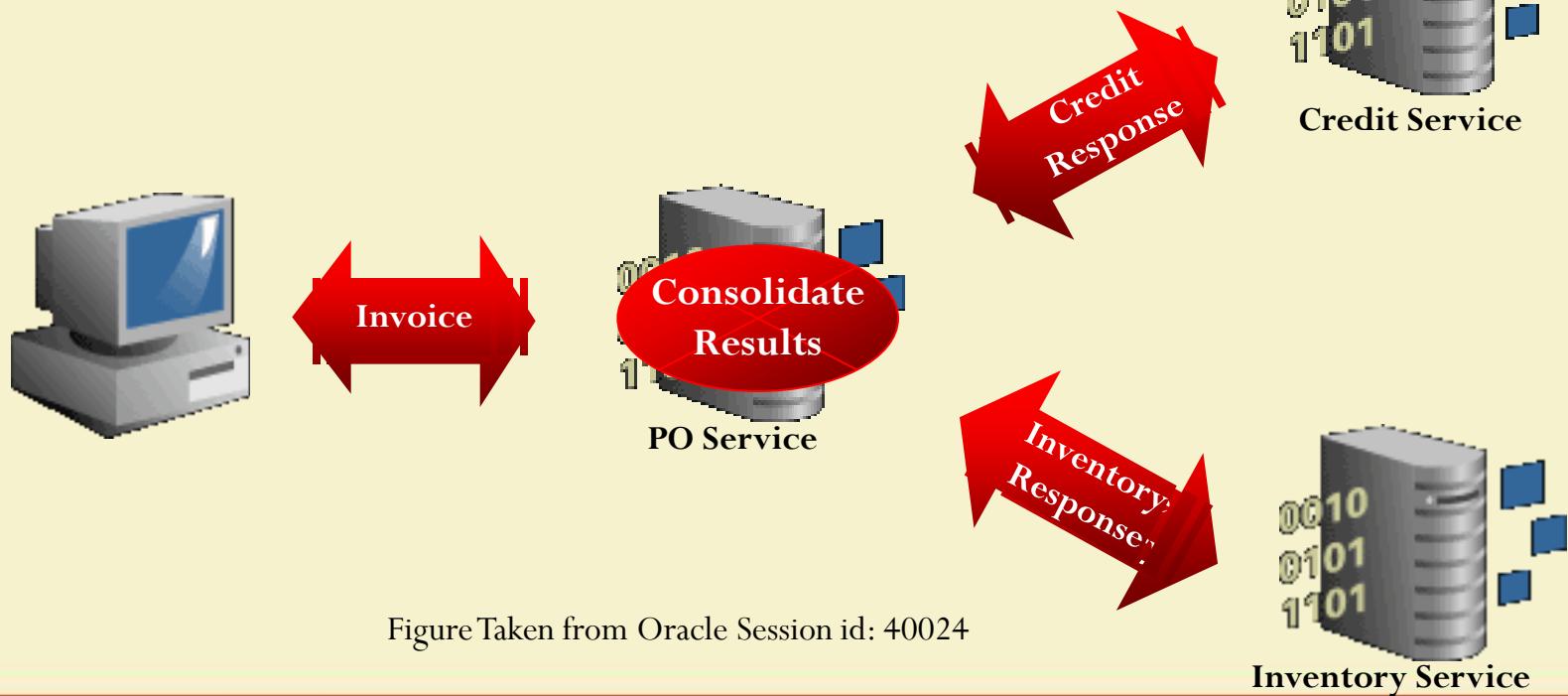
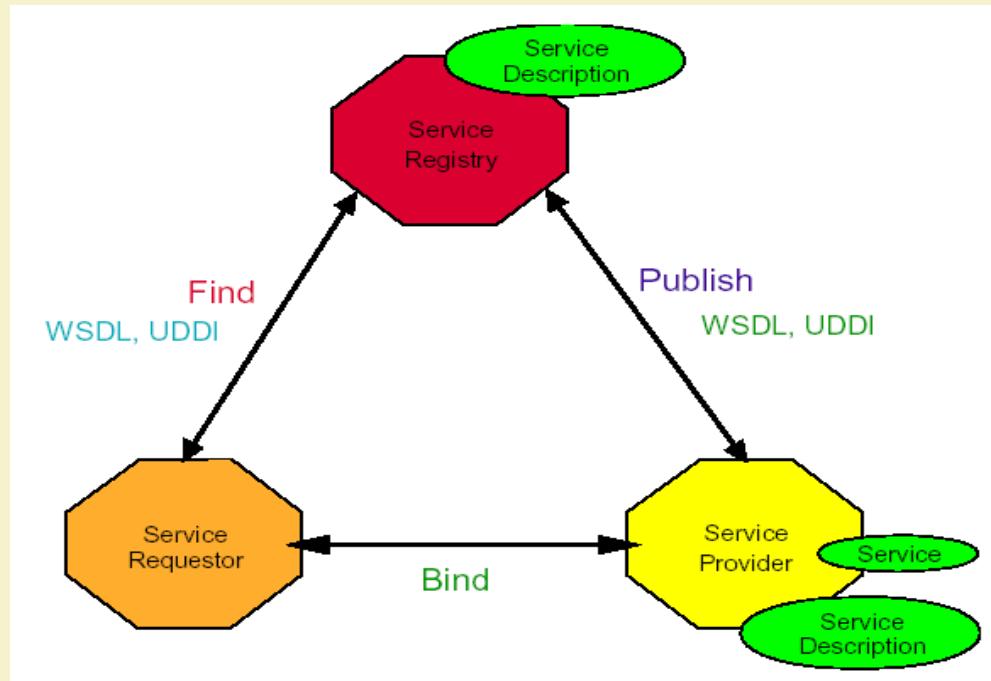


Figure Taken from Oracle Session id: 40024

# 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

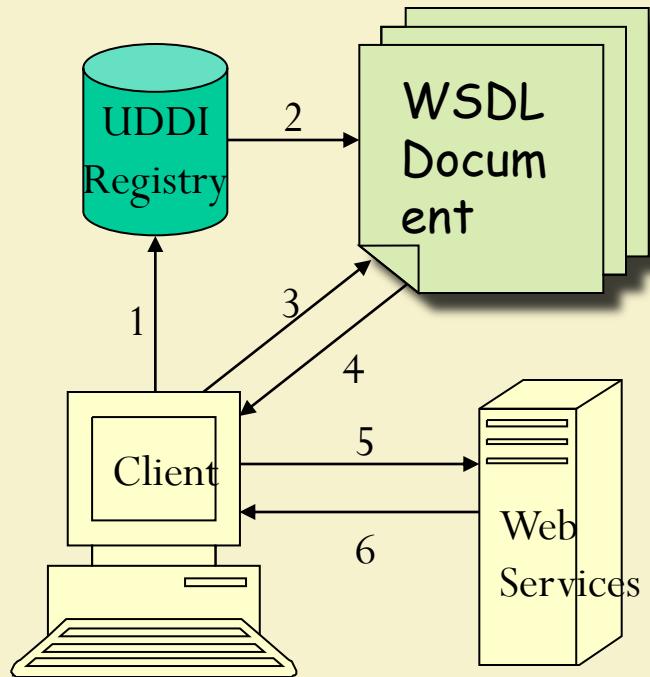
# Web Service Model (contd...)

- 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.

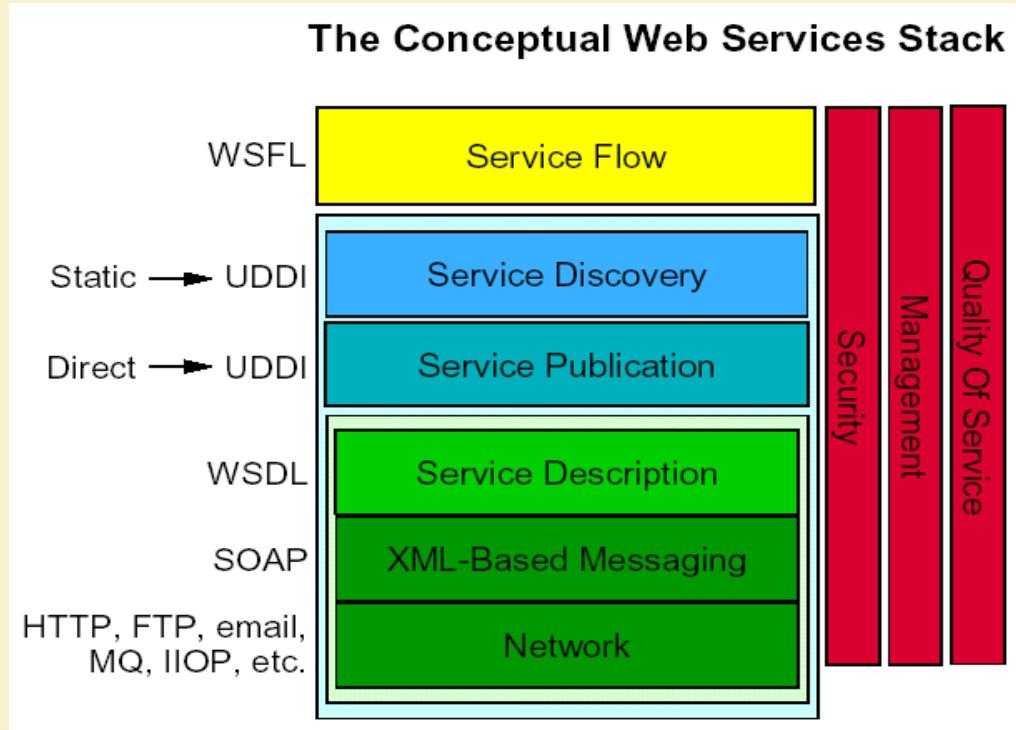


IIT KHARAGPUR



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*

## XML (contd...)

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



IIT KHARAGPUR



NPTEL  
NPTEL ONLINE  
CERTIFICATION COURSES

# 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>
```

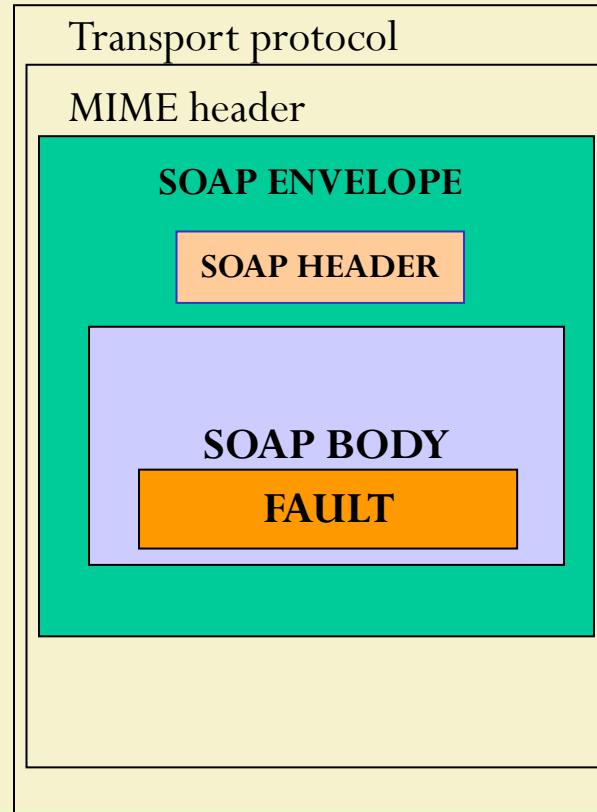
- In this case:
  - The information contained is being marked, but not for displaying.
  - Readable by both human and machines.

# 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>
```

# 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 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



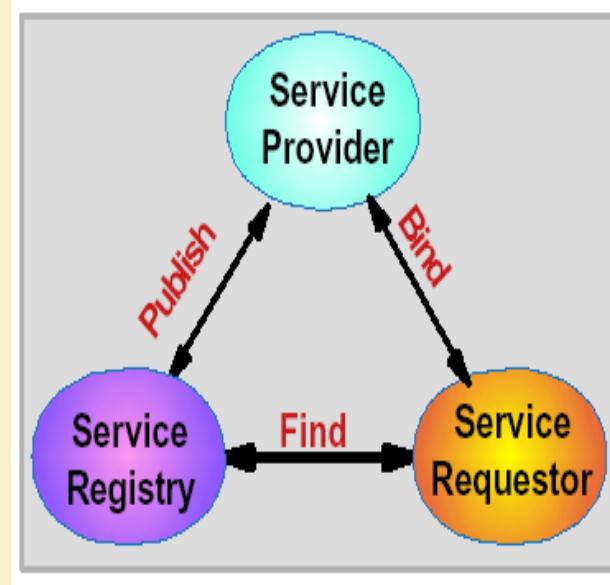
NPTEL ONLINE  
CERTIFICATION COURSES

# UDDI - Universal Description, Discovery, and Integration

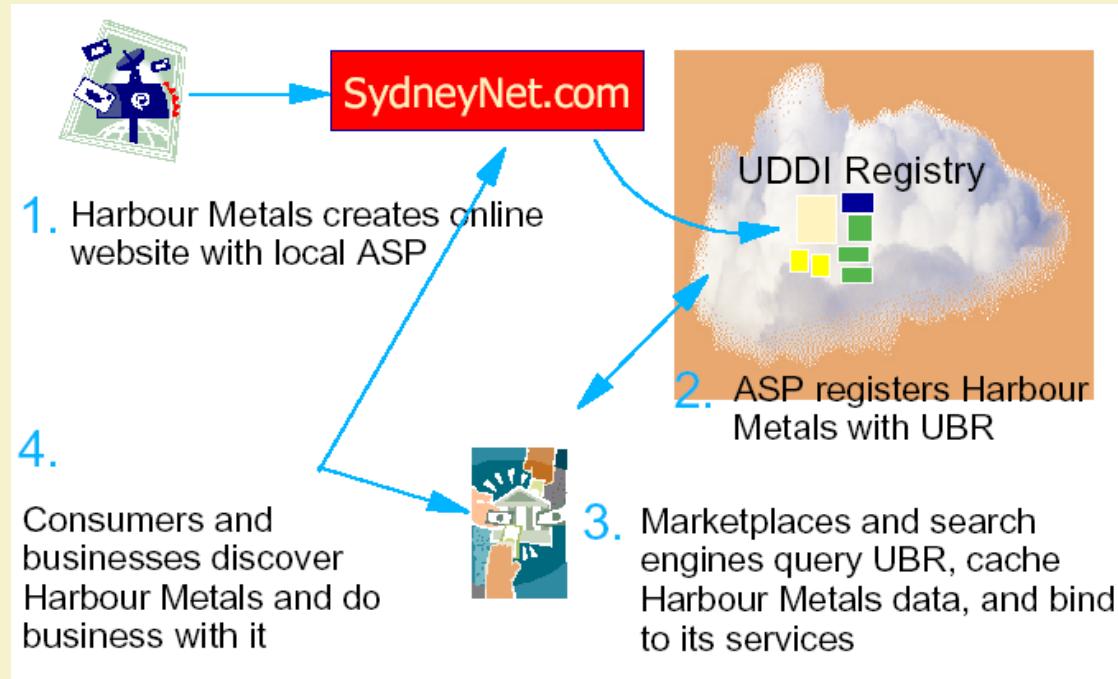
- A framework to define XML-based registries
- Registries are repositories that contain documents that describe business data and also provide search capabilities and programmatic access to remote applications
- Businesses can publish information about themselves and the services they offer
- Can be 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?



IIT KHARAGPUR

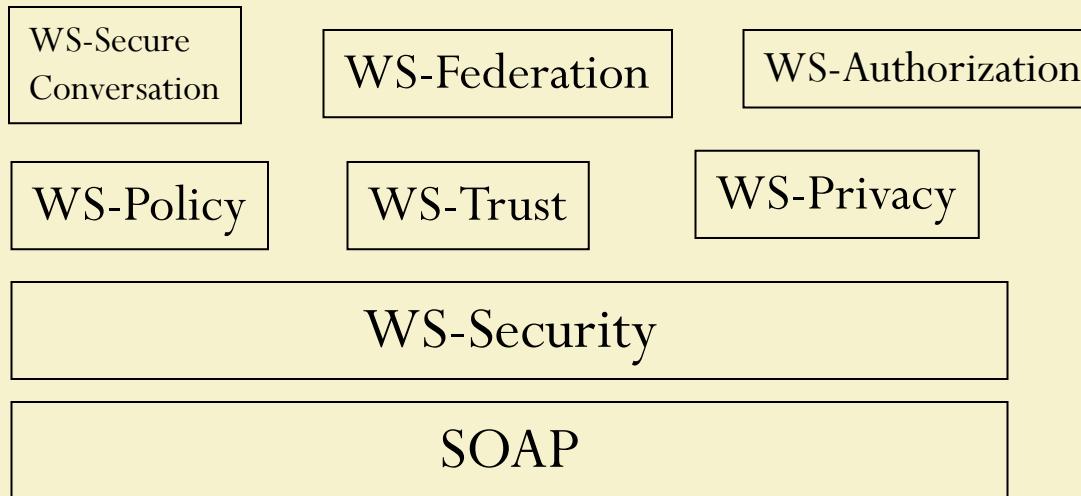


NPTEL  
NPTEL ONLINE  
CERTIFICATION COURSES

# 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!



IIT KHARAGPUR



NPTEL  
NPTEL ONLINE  
CERTIFICATION COURSES