Basic Concepts and Terminology

Cloud

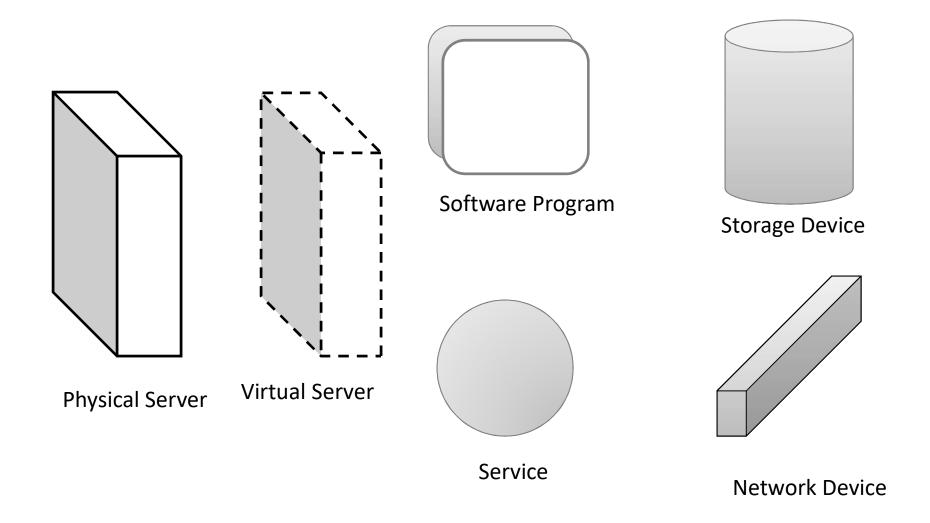
 A cloud refers to a distinct IT environment that is designed for the purpose of remotely provisioning scalable and measured IT resources.

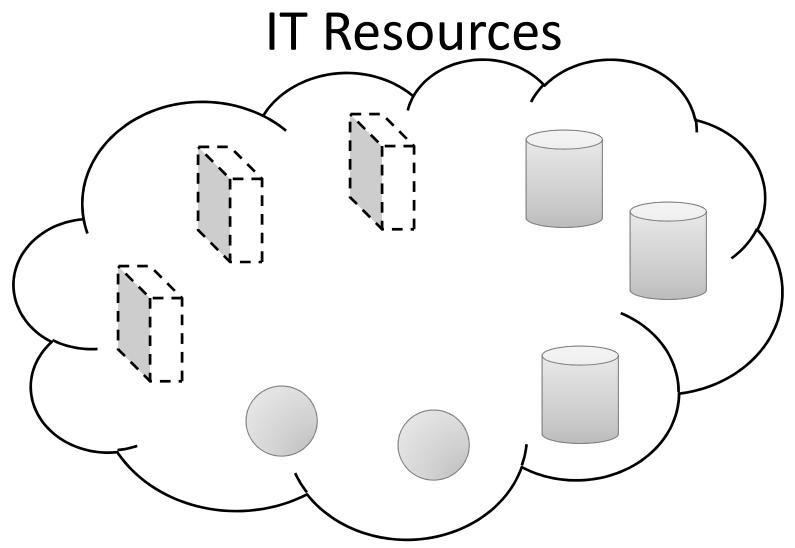
The symbol is used to denote the boundary of a cloud environment.

IT Resources

- An IT resource is a physical or virtual IT-related artifact that can be either
- Software-based
 - such as a virtual server or a custom software program.
- Hardware-based
 - such as physical server or a network device.

IT Resources





A cloud is hosting eight IT resources: three virtual servers, two cloud services, and three storage devices.

On-Premise

- An IT resource that is hosted in a conventional IT enterprise within an organizational boundary.
- On the premises of a controlled IT environment that is not cloud based.
- Key points:
 - A on-premise IT resource can access and interact with a cloud-based IT resource.
 - An on-premise IT resource can be moved to cloud, thereby changing it to a cloud-based service.
 - Redundant deployment of an IT resource can exist in both on-premise and cloud based environments.

Cloud Providers

 The party that provides cloud-based IT resources is the cloud provider.

Cloud Consumer

 The party that uses cloud-based IT resources is the cloud consumer.

Scaling

- Ability of the IT resource to handle increased or decreased usage demands.
- Two types of scaling
 - Horizontal scaling
 - Vertical scaling

Horizontal Scaling

pooled physical servers demand demand servers horizontal scaling

Vertical Scaling

- When an existing IT resource is upgraded or replaced by another with higher or lower capacity.
- Upgrading or replacing of an IT resource with another that has a higher capacity is referred to as *scaling up* and the downgrading or replacing an IT resource with another that has a lower capacity is considered *scaling down*.

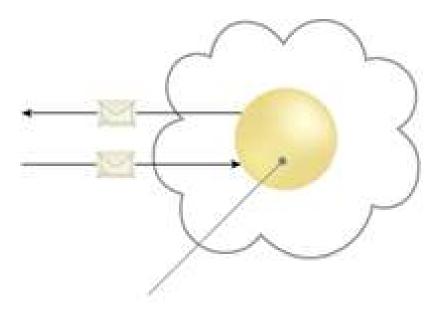
Vertical Scaling

 Vertical in cloud ne required environmo while replaced scaling

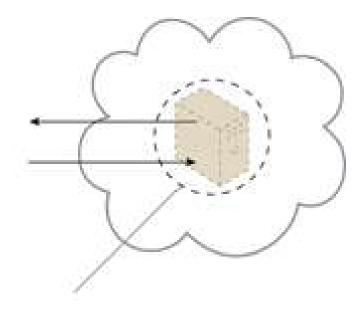
Horizontal Scaling	Vertical Scaling
less expensive (through commodity hardware components)	more expensive (specialized servers)
IT resources instantly available	IT resources normally instantly available
resource replication and automated scaling	additional setup is normally needed
additional IT resources needed	no additional IT resources needed
not limited by hardware capacity	limited by maximum hardware capacity

- Although a cloud is a remotely accessible environment, not all IT resources residing within a cloud can be made available for remote access.
- For example, a database or a physical server deployed within a cloud may only be accessible by other IT resources that are within the same cloud.
- A software program with a published API may be deployed specifically to enable access by remote clients.

- A *cloud service* is any IT resource that is made remotely accessible via a cloud.
- A cloud service can exist as a simple Webbased software program with a technical interface invoked via the use of a messaging protocol, or as a remote access point for administrative tools or larger environments and other IT resources.



Remotely accessed web service acting as a cloud service



Remotely accessed virtual server acting as a cloud service

- The driving motivation behind cloud computing is to provide IT resources as services that encapsulate other IT resources, while offering functions for clients to use and leverage remotely.
- A multitude of models for generic types of cloud services have emerged, most of which are labeled with the "as-a-service" suffix.

- Cloud service usage conditions are typically expressed in a service-level agreement (SLA)
- That is the human-readable part of a service contract between a cloud provider and cloud consumer
- It describes QoS features, behaviors, and limitations of a cloud-based service or other provisions.

- An SLA provides details of various measurable characteristics related to IT outcomes.
- Such as uptime, security characteristics, and other specific QoS features, including availability, reliability, and performance.
- Since the implementation of a service is hidden from the cloud consumer, an SLA becomes a critical specification.

Cloud Service Consumer

• The *cloud service consumer* is a temporary runtime role assumed by a software program when it accesses a cloud service.



 Common types of cloud service consumers can include software programs and services capable of remotely accessing cloud services with published service contracts, as well as workstations, laptops and mobile devices running software capable of remotely accessing other IT resources positioned as cloud services.

ANY QUESTION?

Goals and Benefits

Benefits

- Reduced Investments and Proportional Costs
- Increased Scalability
- Increased Availability and Reliability

- Business model on the mass-acquisition of IT resources. (Just like wholesaler)
- Attractively priced leasing packages.
- Access to powerful infrastructure.
- Elimination of up-front IT investments.
- Measured Usage characteristic allows measured operational expenditures.

- Elimination or minimization of up-front financial commitments leads to
 - Increase IT resource allocation as and when required.
 - Allow capital to be redirected to the core business investment.
- No need to set up large scale data center as it is replaced by cloud providers.
 - No need to invest in real estate, IT professionals, and network bandwidth.

- The same rationale applies to operating systems, middleware or platform software, and application software.
- Pooled IT resources are made available to and shared by multiple cloud consumers, resulting in increased or even maximum possible utilization.
- Operational costs and inefficiencies can be further reduced by applying proven practices and patterns for optimizing cloud architectures, their management and governance.

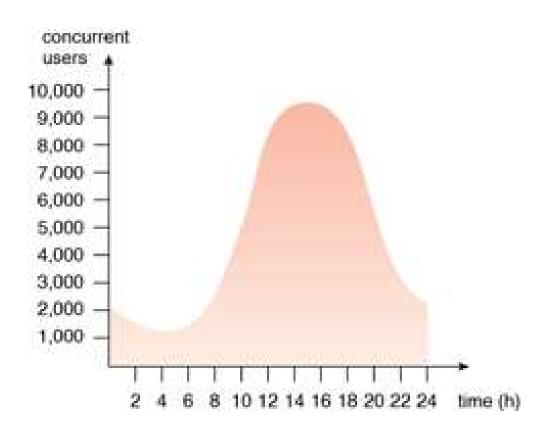
- Common measurable benefits to cloud consumers include:
 - On-demand access to pay-as-you-go computing resources on a short-term basis (such as processors by the hour), and the ability to release these computing resources when they are no longer needed.
 - i.e. Using 100 servers for one hour costs the same as using one server for 100 hours.
 - This "elasticity" of IT resources, achieved without requiring steep initial investments to create a large-scale computing infrastructure, can be extremely compelling.

- The perception of having unlimited computing resources that are available on demand, thereby reducing the need to prepare for provisioning.
- The ability to add or remove IT resources at a finegrained level, such as modifying available storage disk space by single gigabyte increments.
- Abstraction of the infrastructure so applications are not locked into devices or locations and can be easily moved if needed.

- Another area of cost savings offered by clouds is the "as-a-service" usage model, whereby technical and operational implementation details of IT resource provisioning are abstracted from cloud consumers and packaged into "ready-to-use" or "off-the-shelf" solutions.
- These services-based products can simplify and expedite the development, deployment, and administration of IT resources when compared to performing equivalent tasks with on-premise solutions.
- The resulting savings in time and required IT expertise can be significant and can contribute to the justification of adopting cloud computing.

 By providing pools of IT resources, along with tools and technologies designed to leverage them collectively, clouds can instantly and dynamically allocate IT resources to cloud consumers, on-demand or via the cloud consumer's direct configuration.

- This empowers cloud consumers to scale their cloud-based IT resources to accommodate processing fluctuations and peaks automatically or manually.
- Similarly, cloud-based IT resources can be released (automatically or manually) as processing demands decrease.
- It is directly related to proposional cost benefits.



An example of an organization's changing demand for an IT resource over the course of a day.

 The ability of IT resources to always meet and fulfill unpredictable usage demands avoids potential loss of business that can occur when usage thresholds are met.

Increased Availability and Reliability

- The availability and reliability of IT resources are directly associated with tangible business benefits.
- Outages limit the time an IT resource can be "open for business" for its customers, thereby limiting its usage and revenue generating potential.

Increased Availability and Reliability

- Runtime failures that are not immediately corrected can have a more significant impact during high-volume usage periods.
- Not only is the IT resource unable to respond to customer requests, its unexpected failure can decrease overall customer confidence.

Increased Availability and Reliability

 A hallmark of the typical cloud environment is its intrinsic ability to provide extensive support for increasing the availability of a cloud-based IT resource to minimize or even eliminate outages, and for increasing its reliability so as to minimize the impact of runtime failure conditions.

Increased Availability and Reliability

- An IT resource with increased availability is accessible for longer periods of time (for example, 22 hours out of a 24 hour day).
- Cloud providers generally offer "resilient" IT resources for which they are able to guarantee high levels of availability.

Increased Availability and Reliability

- An IT resource with increased reliability is able to better avoid and recover from exception conditions.
- The modular architecture of cloud environments provides extensive failover support that increases reliability.

Summary

- Cloud environments are comprised of highly extensive infrastructure that offers pools of IT resources that can be leased using a pay-foruse model whereby only the actual usage of the IT resources is billable.
- When compared to equivalent on-premise environments, clouds provide the potential for reduced initial investments and operational costs proportional to measured usage.

Summary

- The inherent ability of a cloud to scale IT resources enables organizations to accommodate unpredictable usage fluctuations without being limited by predefined thresholds that may turn away usage requests from customers.
- Conversely, the ability of a cloud to decrease required scaling is a feature that relates directly to the proportional cost benefit.

Summary

- By leveraging cloud environments to make IT resources highly available and reliable, organizations are able to increase quality-ofservice (QoS) guarantees to customers.
- And further reduce or avoid potential loss of business resulting from unanticipated runtime failures.

ANY QUESTION?

Risks and Challenges

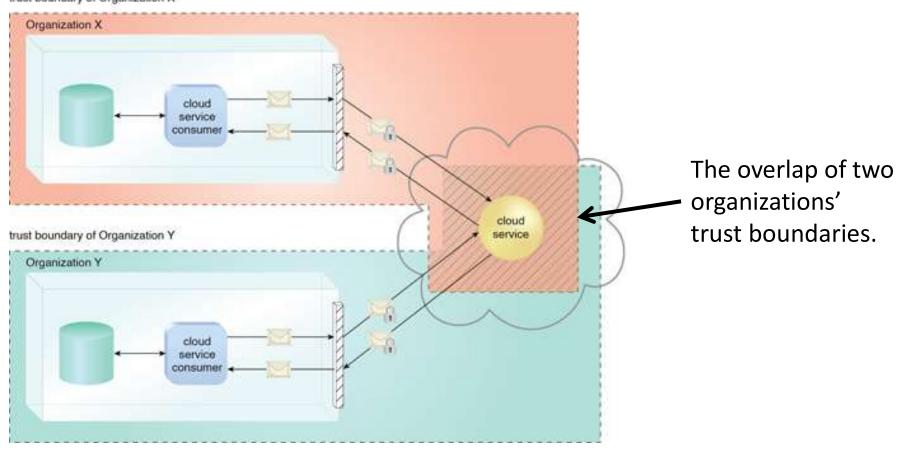
- The moving of business data to the cloud means that the responsibility over data security becomes shared with the cloud provider.
- The remote usage of IT resources requires an expansion of trust boundaries by the cloud consumer to include the external cloud.

 It can be difficult to establish a security architecture that spans such a trust boundary without introducing vulnerabilities, unless cloud consumers and cloud providers happen to support the same or compatible security frameworks—which is unlikely with public clouds.

- Another consequence of overlapping trust boundaries relates to the cloud provider's privileged access to cloud consumer data.
- The extent to which the data is secure is now limited to the security controls and policies applied by both the cloud consumer and cloud provider.
- Furthermore, there can be overlapping trust boundaries from different cloud consumers due to the fact that cloud-based IT resources are commonly shared.

 The overlapping of trust boundaries and the increased exposure of data can provide malicious cloud consumers (human and automated) with greater opportunities to attack IT resources and steal or damage business data.

trust boundary of Organization X



- It can be challenging for the cloud provider to offer security mechanisms that accommodate the security requirements of both cloud service consumers.
- Overlapping trust boundaries is a security threat.

- Cloud consumers are usually allotted a level of governance control that is lower than that over on-premise IT resources.
- This reduced level of governance control can introduce risks associated with how the cloud provider operates its cloud.
- As well as the external connections that are required for communicate between the cloud and the cloud consumer.

- Consider the following examples:
 - An unreliable cloud provider may not maintain the guarantees it makes in the SLAs that were published for its cloud services.
 - This can jeopardize the quality of the cloud consumer solutions that rely on these cloud services.
 - Longer geographic distances require additional network hops that introduce fluctuating latency and potential bandwidth constraints.

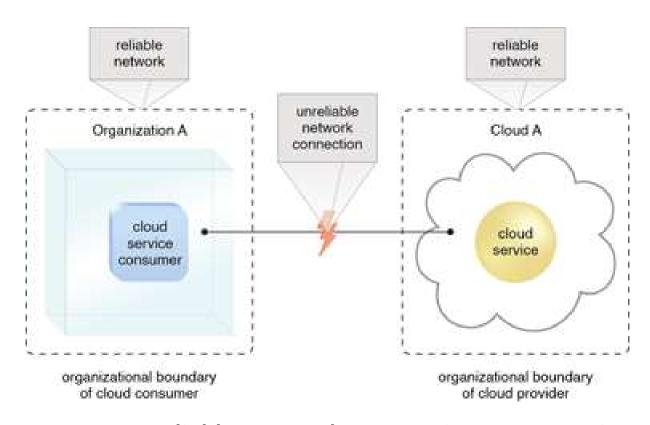


Figure 1 – An unreliable network connection compromises the quality of communication between cloud consumer and cloud provider environments.

- Legal contracts, when combined with SLAs, technology inspections, and monitoring, can mitigate governance risks and issues.
- A cloud governance system is established through SLAs, given the "as-a-service" nature of cloud computing.
- A cloud consumer must keep track of the actual service level being offered and the other warranties that are made by the cloud provider.

Limited Portability Between Cloud Providers

- Due to a lack of established industry standards within the cloud computing industry, public clouds are commonly proprietary to various extents.
- For cloud consumers that have custom-built solutions with dependencies on these proprietary environments, it can be challenging to move from one cloud provider to another.

Limited Portability Between Cloud Providers

Cloud A (Cloud Provider X) Portal e the supports message impac urces encryption and digital signatures and d requires encryption and cloud digital signing consumer of messages Cloud B (Cloud Provider Y) supports message encryption only

- Third-party cloud providers will frequently establish data centers in affordable or convenient geographical locations.
- Cloud consumers will often not be aware of the physical location of their IT resources and data when hosted by public clouds.

- For some organizations, this can pose serious legal concerns pertaining to industry or government regulations that specify data privacy and storage policies.
- For example, some UK laws require personal data belonging to UK citizens to be kept within the United Kingdom.

- Another potential legal issue pertains to the accessibility and disclosure of data.
- Countries have laws that require some types of data to be disclosed to certain government agencies or to the subject of the data.
- For example, a European cloud consumer's data that is located in the U.S. can be more easily accessed by government agencies (due to the U.S. Patriot Act) when compared to data located in many European Union countries.

 Most regulatory frameworks recognize that cloud consumer organizations are ultimately responsible for the security, integrity, and storage of their own data, even when it is held by an external cloud provider.

END

Introduction to Cloud Computing

Dr. Jitendra Nasriwala

Computing

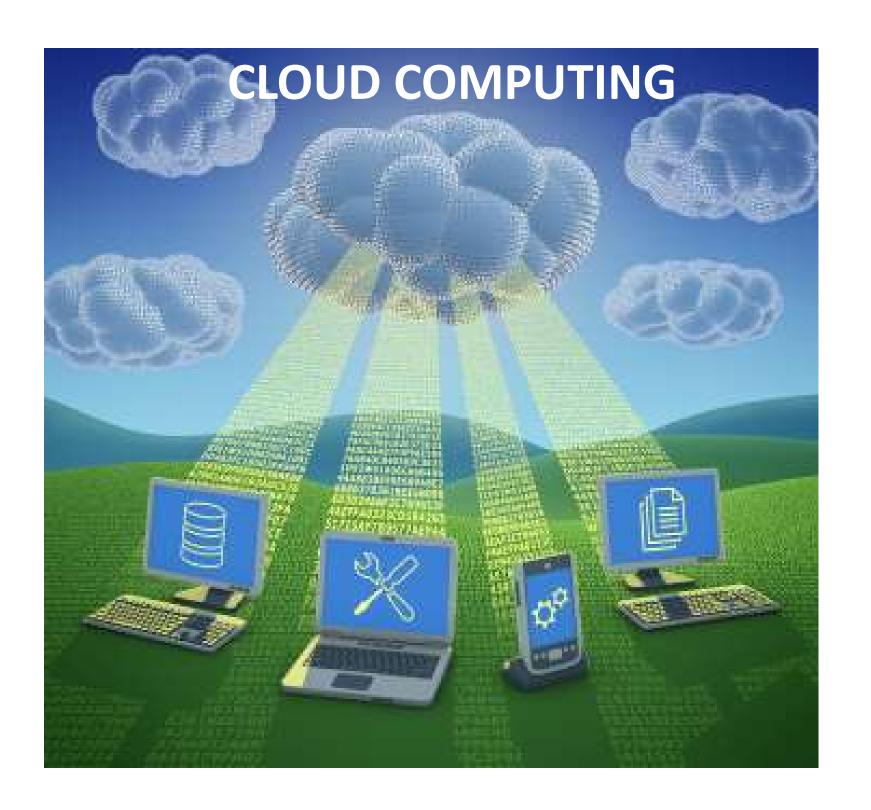
- Computing is any activity that uses computers to manage, process, and communicate information.
- It includes development of both hardware and software.
- Computing is a critical, integral component of modern industrial technology.

Computing

- Computing is any goal-oriented activity requiring, benefiting from, or creating computers.
- For example, computing includes
 - designing, developing and building hardware and software systems;
 - processing, structuring, and managing various kinds of information;
 - doing scientific research on and with computers;
 - making computer systems behave intelligently;
 - creating and using communications and entertainment media etc.

The Cloud

- A cloud refers to a distinct IT environment that is designed for the purpose of remotely provisioning scalable and measured IT resources.
- Internet , a network of networks providing remote access to a set of decentralized IT resources.



- Utility computing a concept that computer scientist John McCarthy publicly proposed in 1961:
 - "If computers of the kind I have advocated become the computers of the future, then computing may someday be organized as a public utility just as the telephone system is a public utility. ... The computer utility could become the basis of a new and important industry."

- (1969) Leonard Kleinrock, Chief Scientist of ARPANET (Advanced Research Projects Agency Network)
 - As of now, computer networks are still in their infancy, but as they grow up and become sophisticated, we will probably see the spread of 'computer utilities' which, like present electric and telephone utilities, will service individual homes and offices across the country."

- Mid 1990s incarnation of search engines
 - Dec 1990 archie
 - Jan 1994 Yahoo
 - 1998 to 2000 Google Search Engine goto.com to google.com
- Open publishing platform Youtube,
 MySpace, facebook
- Email services hotmail, yahoo, gmail

- Late 1990s, Salesforce.com pioneered the notion of bringing remotely provisioned services into enterprise.
- In 2002, Amazon.com launched the Amazon Web Services platform, a suite of enterpriseoriented services that provide remotely provisioned storage, computing resources, and business functionality.

- The term "Network Cloud" or "Cloud" was introduced in the early 1990s throughout the networking industry.
- It is an abstract layer to represent delivery methods of data across heterogeneous public and semi-public networks.
- The networking method support transmission of data from one end-point(local network) to the "Cloud" (wide are network) and then further decomposed to another intended end-point.
- It is used under the utility computing.

- Aug 2006, Amazon launched Elastic Compute Cloud (EC2) services that enabled organizations to "lease" computing capacity and processing power to run their enterprise applications.
- Google Apps also began providing browser based enterprise applications in the same year.

Definitions

- Gartner is an information technology (IT) research and consultancy company, formerly known as Gartner Group.
- Gartner definition
 - Cloud computing is a style of computing in which scalable and elastic IT-enabled capabilities are delivered as a service using Internet technologies.

Definitions

- Forrester Research:
 - "... A standardized IT capability (services, software, or infrastructure) delivered via Internet technologies in a pay-per-use, self-service way."

Definition

- Cloud is a model where users have a convenient, on-demand access to a shared pool of resources, such as servers, storage, and applications, over the Internet.
- Users don't have a control of underlying hardware infrastructure that is owned and managed by the provider.
- They access the services or allocated resources by using a Web browser.

Definition

- The National Institute of Standards and Technology (NIST), USA
 - "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."

- In NIST definition, the five salient features of cloud are:
- 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
- Resources can be dynamically assigned and reassigned according to customer demand
- Customer generally may not care where the resources are physically located but should be aware of risks if they are located offshore

Rapid Elasticity

- Capabilities can be expanded or released automatically (i.e., more CPU power, or ability to handle additional users)
- To the customer this appears seamless, limitless, and responsive to their changing requirements

Measured Service

- Customers are charged for the services they use and the amounts
- There is a metering concept where customer resource usage can be monitored, controlled, and reported, providing transparency for both the provider and consumer of the utilized service

Definition

- "Cloud computing is a specialized form of distributed computing that introduces utilization models for remotely provisioning scalable and measured resources."
 - Definition by Thomas Erl book

Cloud computing can be summarized as:

- I don't care where my servers are, who manages them, where my documents are stored, or where my applications are hosted.
- I just want them always available and access them from any device connected through Internet. And
- I am willing to pay for this service for as a long as I need it.

Motivation towards development of Cloud Computing

Business Drivers
Cloud enabling/similar technologies

- Several of the primary business drivers that fostered modern cloud-based technology.
 - Capacity planning
 - Process of determining and fulfilling future demands of an organization's IT resources, products and services.
 - Capacity represents the maximum amount of work that an IT resource is capable of delivering in a given period of time.
 - Discrepancy between the capacity of an IT resource and its demand can result in a system becoming either inefficient or unable to fulfill user needs.

- Capacity planning is focused on minimizing this discrepancy to achieve predictable efficiency and performance.
- Different capacity planning strategies exist:
- Lead strategy
 - Adding capacity to an IT resource in anticipation of demand
- Lag strategy
 - Adding capacity when the IT resource reaches its full capacity
- Match strategy
 - Adding IT resource capacity in small increments, as demand increases

- Planning capacity can be challenging because it requires estimating usage load fluctuations.
 - Outfitting IT infrastructure → unreasonable financial investments
 - Moderating investment → under-provisioning resulting leading to transaction losses and other usage limitations from lowered usage thresholds.

Cost Reduction

- Difficult to establish relationship between IT cost and business performance.
- Automation solution is always be limited by the processing power of its underlying infrastructure.
- The IT cost includes: the cost of acquiring new infrastructure and the cost of its ongoing ownership.
- The total cost of ownership (TCO) is the purchase price of an asset plus the costs of operation.
- Operational overhead represents a considerable share of IT budgets, often exceeding up-front investment costs.

- Common forms of infrastructure-related operating overhead include the following:
 - Technical personnel required to keep the environment operational
 - Upgrades and patches that introduce additional testing and deployment cycles
 - Utility bills and capital expense investments for power and cooling
 - Security and access control measures that need to be maintained and enforced to protect infrastructure resources
 - Administrative and accounts staff that may be required to keep track of licenses and support arrangements

- The on-going ownership of internal technology infrastructure can encompass burdensome responsibilities that impose compound impacts on corporate budgets.
- An IT department consequently become a significantand at times overwhelming-drain on the business, potentially inhibiting its responsiveness, profitability and overall evolution.

- Organizational Agility
 - Organizational agility is a company's ability to rapidly change or adapt in response to change caused by both internal or external factors.
 - An IT enterprise often needs to respond to business change by scaling its IT resources beyond the scope of what was previously predicted or planned for.

ASP

- Application Service Provider Term coined in 1996.
- Defined as an organization that hosts and manages one or more applications ant its underlying infrastructure.
- Customer could use these application over the internet and billed for the amount of utilization.

Autonomic Computing

- It is a set of self managing characteristics of distributed computing resources that operate on the basis of a set of pre defined policies.
- These systems are capable of self healing, selfconfiguration of their components, selfoptimization of their resources, and selfprotection from malware and attacks.

Cluster

- It is a group of independent IT resources that are interconnected and work as single system.
- System failure rates are reduced.
- All the nodes are actively working or some nodes are in standby mode, waiting to take over after the failure of an active node.

- Due to identical hardware and operating systems,
 failover is possible with similar performance.
- Component devices that form a cluster are kept in synchronization through dedicated, high-speed communication link.

Distributed Computing

- Different roles or tasks are distributed among separate nodes in the network.
- Grid computing, peer-to-peer architecture, and client-server architecture are some forms of distributed computing.

- High Performance Computing
 - This technique divides a task into pieces, and uses parallel processing algorithms to execute each piece on different processors on the same node or multiple nodes in the network.

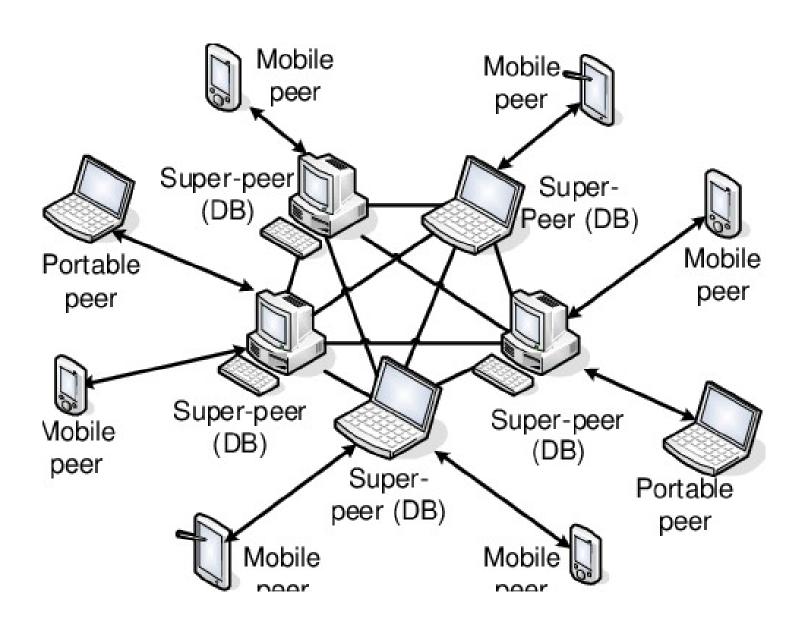
Utility Computing

- it started in the early days of mainframe in the 1960s.
- Mainframe were very expensive, even for large, profitable companies.
- Hence, mainframe manufacturer provided form of utility computing called time-sharing, where they offered database storage and compute power to banks and other large organization for a fee.

- Peer to Peer
- Client-server
- Grid Computing

Cloud Computing Vs Peer to Peer Architecture

Peer to Peer Architecture



Peer to Peer Architecture

- Resource sharing, processing and communication control are completely decentralized.
- Each host act as a server or provider of certain services.
- Host relies on other nodes within network for other services.
- All clients are equal in terms of providing and using resources and users are authenticated by each individual workstation.

Cloud Computing Vs P2P Architecture

Benefits over P2P

- easily scalable
- Enabled access to any type of hosted applications
- Does not burden user-end devices
- Configured with the highest level of security

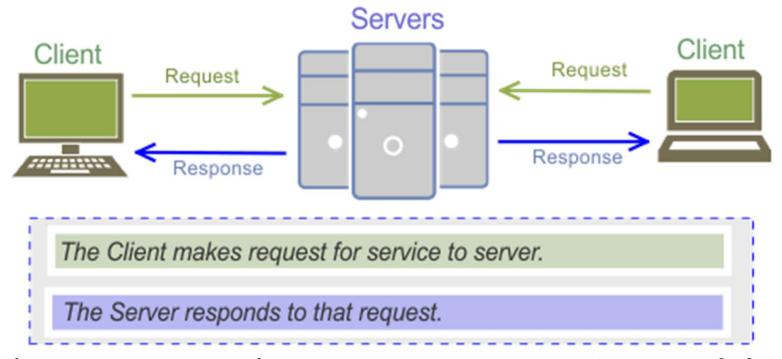
Benefits of P2P

- Deployment are relatively inexpensive
- Simple to set up and manage

Cloud Computing Vs P2P Architecture

- Shortcomings with cloud computing
 - High initial capital investment
 - Need good technology expertise to establish and manage
- P2P downside
 - Limited in extensibility
 - Tend to overburden user workstation by making them work as server for other user
 - Lax security
 - Unable to provide system-wide service

Cloud Computing Vs Client-Server Architecture



Client Server Architecture is a computing model in which the server hosts, delivers and manages most of the resources and services to be consumed by the client

Cloud Computing Vs Client-Server Architecture

- Client-Server Architecture is a form of distributed computing.
- Clients depends on a number or servers for various services or resources such as database, applications, security, printing and backups.
- Central server for authentication services.
- These servers provides access to shared files, printers, hardware storage, and applications.

Cloud Computing Vs Client-Server Architecture

- In cloud, running application is a part of clientserver architecture.
- Cloud computing can provide increased performance, flexibility and significant cost savings.

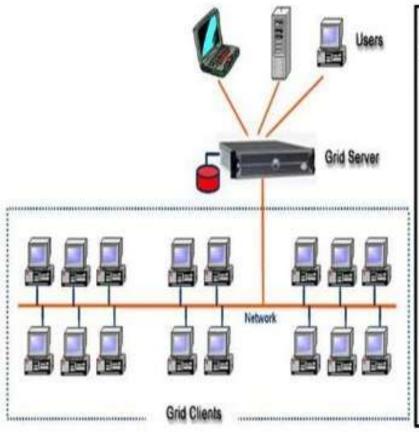
- In early 1990s, Carl Kesselman and Ian Foster formulated the concept of grid computing.
- It is a cluster of computer system that were geographically distributed.
- It provides a platform in which computing resources are organized into one or more logical pools.
- Work together to perform a common task.

- Grid computing sometimes referred as "super virtual computer".
- In a grid, cluster of loosely coupled computers work together to solve a single problem.
- Problem involve massive amount of numerical calculation and compute cycles.

- Grid controlling software that divides the work into smaller pieces and assigns each piece to a pool of thousands of computers.
- The controlling unit later assembles the results to build the output.
- Grids are usually used to harness idle computer power.

- Grid computing is based on a middleware layer that is deployed on computing resources.
- These IT resources participate in a grid pool that implements a series of workload distribution and coordination functions.
- This middle tier contain load balancing logic, failover controls, and automatic configuration management.

How Grid computing works?



In general, a grid computing system requires:

- At least one computer, usually a server, which handles all the administrative duties for the System
- A network of computers running special grid computing network software.
- A collection of computer software called middleware

- Cloud computing harnesses idle computer power over a network connection.
- It is metered utility service.
- In both, user need not to do any capital expenses or upfront implementation.

Virtualization

- It is a technology platform used for creation of virtual instances of IT resources.
- It allows physical IT resources to provide multiple virtual images of themselves.
- So that their underlying processing capabilities can be shared by multiple user.

Cloud Enabling Technologies

- Broadband Networks and Internet Architecture
- Data Center Technology
- Virtualization Technology
- Web Technology
- Multitenant Technology
- Service Technology