4.1 Architectural Design of Compute and Storage Clouds

- Major design goals of a cloud computing platform is scalability, virtualization, efficiency, and reliability. Clouds support Web 2.0 applications.
- The cloud management receives the user request and then finds the correct resources, and then calls the provisioning services which invoke resources in the cloud. The cloud management software need to support both physical and virtual machines.
- The platform needs to establish a very large-scale HPC infrastructure. The hardware and software systems are combined together to make it easy and efficient to operate. The system scalability can benefit from cluster architecture.
- A cloud platform should be built to serve many users simultaneously. Therefore, multitasking is a necessity to assess distributed system performance.
- Five basic performance metrics are shown in Fig. 4.1.1. Refined performance models could be extended form basic attributes to include program behavior, environmental demand, QoS and cost-effectiveness.

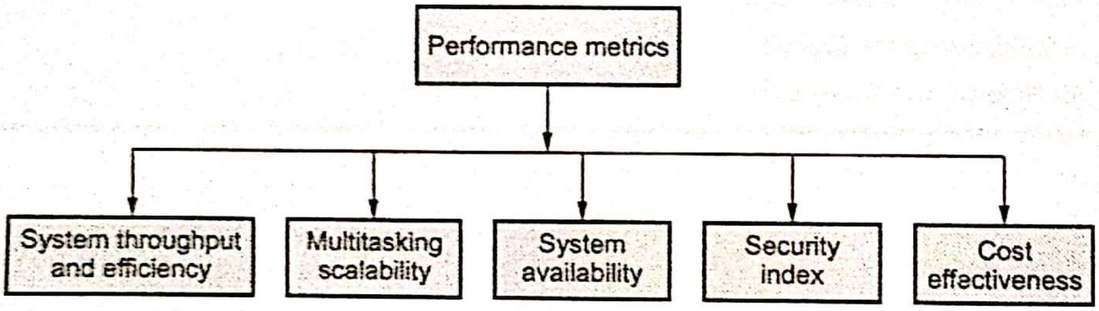

Fig. 4.1.1 Performance Metrics

- Enabling technologies for clouds : The key driving forces behind cloud computing are the ubiquity of broadband and wireless networking, falling storage costs, and progressive improvements in Internet computing software.
- Cloud users are able to demand more capacity at peak demand, reduce costs, experiment with new services, and remove unneeded capacity, whereas service providers can increase the system utilization via multiplexing, virtualization and dynamic resource provisioning.
- Resource virtualization enables rapid cloud deployment faster and fast disaster recovery. Service-oriented architecture (SOA) also plays a vital role. The progress in providing Software as a Service, Wed.2.0 standards and Internet performance have all contributed to the emergence of cloud services.
- The cloud computing resources are built in data centers, which are typically owned and operated by a third-party provider. Consumers do not need to know the underlying technologies.

- Web service providers offer special APIs that enable developers to exploit Internet clouds. Monitoring and metering units are used to track the usage and performance of resources provisioned. The software infrastructure of a cloud platform must handle all resource management and do most of the maintenance, automatically

4.1.1 Layered Cloud Architecture Development

- Fig. 4.1.2 shows layered architectural development of the cloud platform for IaaS, PaaS, and SaaS applications over the Internet and intranet.

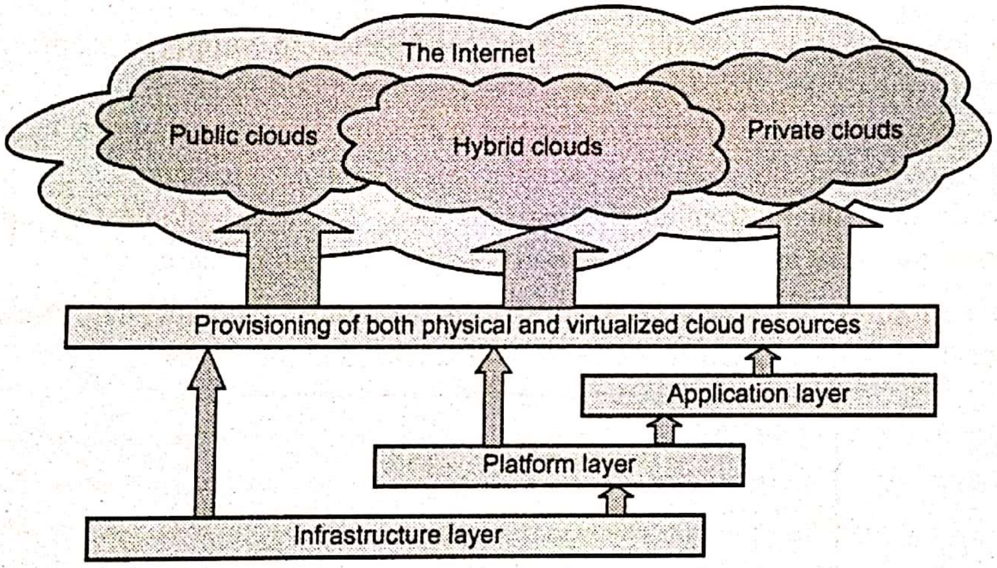

Fig. 4.1.2 Layered architectural development of the cloud platform

- The architecture of a cloud is developed at three layers : Infrastructure, platform, and application. These three development layers are implemented with virtualization and standardization of hardware and software resources provisioned in the cloud.
- The services to public, private, and hybrid clouds are conveyed to users through the networking support over the Internet and intranets involved. It is clear that the infrastructure layer is deployed first to support IaaS type of services.
- This infrastructure layer serves as the foundation to build the platform layer of the cloud for supporting PaaS services. The infrastructure layer is built with virtualized compute, storage and network resource.
- The platform layer is for general-purpose and repeated usage of the collection of software resources. The application layer is formed with a collection of all needed software modules for SaaS application.

4.1.2 Design Challenges

1. Service availability and data lock-in problem
2. Data privacy and security concerns
3. Unpredictable performance and bottlenecks
4. Distributed storage and wide-spread software bug
5. Cloud scalability, interoperability and standardization
6. Software licensing and reputation sharing

4.2 Inter Cloud Resource Management

- The inter cloud is a cloud of clouds constructed to support resource sharing between the clouds. The resources under the inter cloud environment are managed in distributed model without any central authority. The inter cloud communication and resource identification is a complex task. The software agents are small piece of code that can be used to perform any task. The agent models are applied to execute the tasks as small fragments for a specified requirement.
- Fig. 4.2.1 shows six layers of cloud services, ranging from hardware, network and collocation to infrastructure, platform and software applications

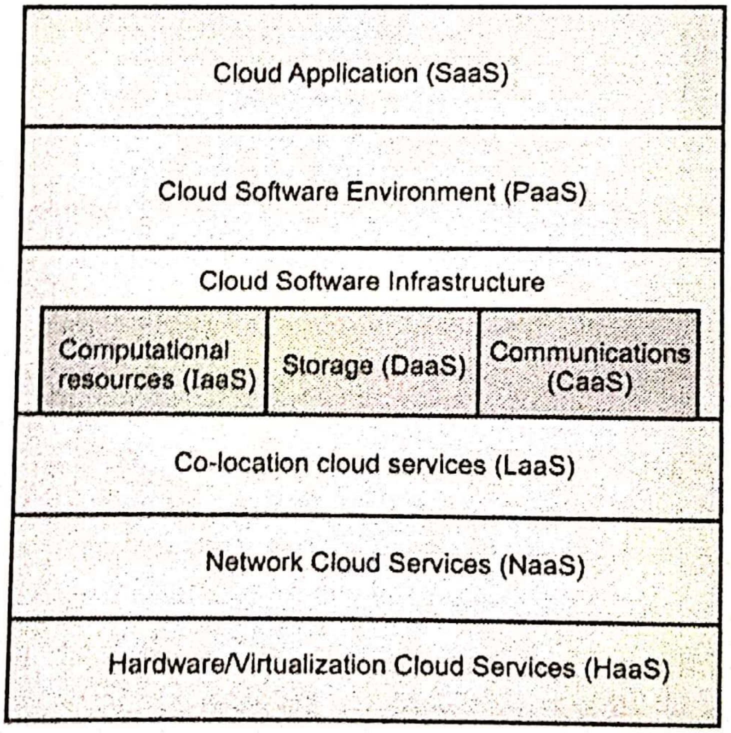
Cloud Application (SaaS)
Cloud Software Environment (PaaS)
Cloud Software Infrastructure
Computational resources (IaaS)
Storage (DaaS)
Communications (CaaS)
Co-location cloud services (LaaS)
Network Cloud Services (NaaS)
Hardware/Virtualization Cloud Services (HaaS)

Fig. 4.2.1 Six layer stack

- The bottom most layer provides Hardware as a Service (HaaS). The next layer is for interconnecting all the hardware components, and is simply called Network as a Service (NaaS). Virtual LANs fall within the scope of NaaS.

- The next layer up offers Location as a Service (LaaS), which provides a collocation service to house, power, and secure all the physical hardware and network resources. The cloud infrastructure layer can be further subdivided as data as a service and communication as a service in addition to compute and storage in IaaS.
- The top layer is for SaaS applications. For example, CRM is heavily practiced in business promotion, direct sales, and marketing services. CRM offered the first SaaS on the cloud successfully.
- PaaS is provided by Google, Salesforce.com and Facebook, among others. IaaS is provided by Amazon, Windows Azure, and RackRack, among others.
- Runtime support services : As in a cluster environment, there are also some runtime supporting services in the cloud computing environment. Cluster monitoring is used to collect the runtime status of the entire cluster. Runtime support is software needed in browser-initiated applications applied by thousands of cloud customers.

4.2.1 Resource Provisioning and Platform Deployment

- Cloud architecture puts more emphasis on the number of processor cores.
- Provisioning of compute resources : Providers supply cloud services by signing SLAs with end users. The SLAs must commit sufficient resources such as CPU, memory and bandwidth that the user can use for a pre-set period.
- Under-provisioning of resources will lead to broken SLAs and penalties. Overprovisioning of resources will lead to resource underutilization and consequently, a decrease in revenue for the provider. Efficient VM provisioning depends on the cloud architecture and management of cloud infrastructures.
- Resource provisioning methods :

  a) The demand-driven method provides static resources and has been used in grid computing for many years. When a resource has surpassed a threshold for a certain amount of time, the scheme increases that resource based on demand.

  b) The event driven method is based on predicted workload by time. This scheme adds or removes machine instances based on a specific time event.

  c) The popularity-driven method is based on Internet traffic monitored. the Internet searches for popularity of certain applications and creates the instances by popularity demand.

4.2.2 Global Exchange of Cloud Resources

- In cloud computing, large numbers of customers use cloud services from all over the world. To ensure reliability in the cloud server, the service provider established various data centers in different locations worldwide.
- For example, the famous e-commerce website AMAZON has data centers in different geographical areas across the world. Even though the site has different data centers, it has specific limitations; for example, they don't have an automatic mechanism by which data centers at different locations can cooperate better and scale their different hosting services.
- Fig. 4.2.2 shows Inter-cloud exchange of cloud resources through brokering.

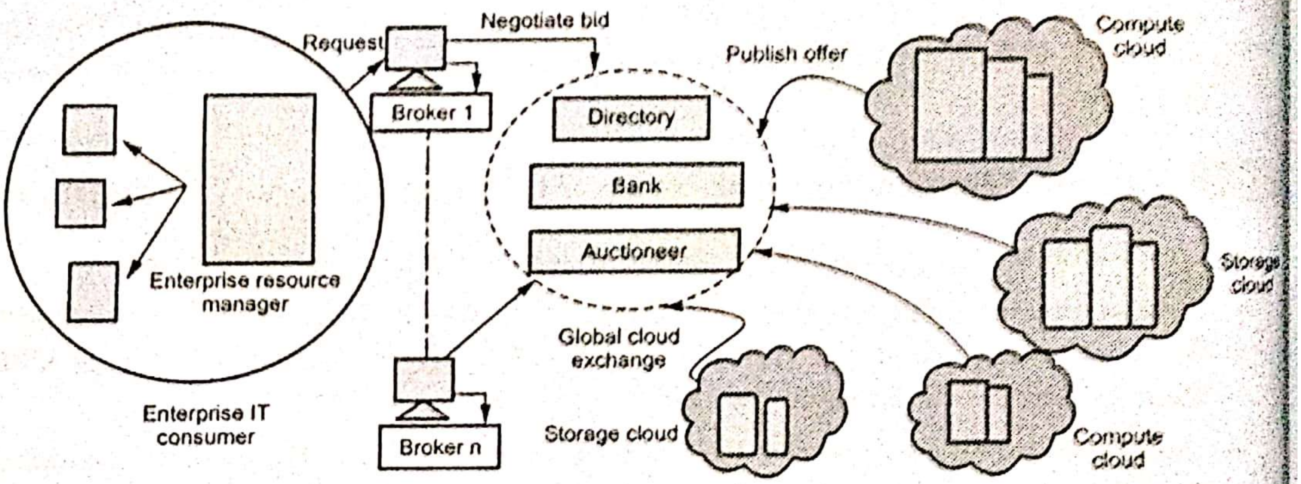

Fig. 4.2.2 Inter-cloud exchange of cloud resources through brokering

- The cloud exchange acts as a market maker for bringing together service producers and consumers. It aggregates the infrastructure demands from application brokers and evaluates them against the available supply currently published by the cloud coordinators.

4.3 Administrating the Clouds

- Administering cloud computing services is an important process when you have hosted your business data on the cloud. The business owners need to know whether the performance is at the right level and whether the deleted data is permanently gone.
- Cloud service provider can definitely build and provide a stable service that are cost effective and efficient. However, there can be a serious gap between the actual service and the promised services.

- You would need evaluate the solution providers when you are choosing a cloud application. Some of the questions that you can ask your vendors are :
  a) Are the vendors available to solve any software issues ?
  b) How will they manage if there is an outage ?
  c) How much experience do they hold in managing customer issues ?
  d) Will they provide training to the customers ?

4.3.1 Cloud Management Products

- Cloud management is the organized oversight, control, administration and maintenance of public cloud, private cloud or more commonly, hybrid multi-cloud computing infrastructure, services and resources.
- Cloud management services combine different technologies and products to deliver a cohesive, consistent strategy and process. Administrators can orchestrate delivery and management of cloud infrastructure, applications, data, services and access control. They can access resources, automate processes, make changes as needed and monitor utilization and cost.
- Cloud management platforms help IT teams secure and optimize cloud infrastructure, including the applications and data residing on it. Administrators can manage compliance, set up real-time monitoring, and pre-empt cyberattacks and data breaches.
- Typically, a cloud management system will be installed on the target cloud. It captures information on activity and performance then sends analysis to a web-based dashboard where administrators can see and act accordingly. Where there is an issue, administrators can issue commands back to the cloud through the cloud management platform, that servers as a consolidated point of control.

4.3.1.1 Dynamo

- Dynamo is propriety key value structured storage system. It can act as database and also distributed hash table.
- Dynamo dynamically partitions a set of keys over a set of storage nodes
- It is most powerful relational database available in WWW. Relational databases have been used a lot in retail sites, to make visitors browse and search for product easily.
- Dynamo does not support replication.
- Dynamo is used to manage the state of services that have very high reliability requirements and need tight control over the tradeoffs between availability, consistency, cost-effectiveness and performance.