* INTER-CLOUD RESOURCE MANAGEMENT

→ This section characterizes the various cloud service models & their extensions.

Cloud application (SaaS)	Concur, RightNOW, Teleo, kenexa, Webex, Blackbaud, Salesforce.com etc
Cloud Software environment (PaaS)	Force.com, App Engine, Facebook, MS Azure, NetSuite, IBM BlueCloud, SGI Cyclone, eBay.
Cloud Software Infrastructure Computational resource (IaaS) | Storage (DaaS) | Communication (CaaS)	Amazon AWS, OpSourceCloud, IBM Ensembles, Rackspace cloud, Window Azure, HP, Banknorth.
Collocation cloud service (LaaS)	Savvis, Internap, NTTCommunications, Digital realty trust, 365 Main
Network cloud services (NaaS)	Qwest, AT&T, AboveNet
Hardware / Virtualization cloud services (HaaS)	VMware, Intel, IBM, XenEnterprise

A stack of Six layers of cloud services & their providers.

→ It consist of six layers of cloud services, ranging from hardware, network, & collocation to Infrastructure, platform, & Soft. applications. We already introduced the top three service layers as SaaS, PaaS & IaaS respectively. The cloud platform provides PaaS, which sits on top of the IaaS inf..

(c) The top layer offers SaaS. These must be implemented on the cloud platforms provided. Although the three basic models are dissimilar in usage. They are built one on top of another. The implication is that one cannot launch SaaS applications with a cloud platform. The cloud platform cannot be built if compute & storage infrastructures are not there.

→ The bottom three layers are more related to physical requirements. The bottommost 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 collection service to house, power, & secure all the physical hardware and network resources. Some authors say this layer provides Security as a service ("SaaS"). The cloud infrastructure layer can be further subdivided as Data as a service (DaaS) & communication as a service (CaaS) in addition to compute & storage in IaaS.

→ Cloud players are divided into 3 classes:
(1) Cloud service providers & IT administrators
(2) Software developers or vendors, and

(3) end users or business users. These cloud players vary in their roles under the IaaS, PaaS & SaaS models. The table entries distinguish the three cloud models as viewed by diff. players. From the S/W vendor's perspective, application performance on a given cloud platform is most important. From the providers' perspective, cloud infrastructure performance is the primary concern. From the end users' perspective, the quality of services, including security, is the most important.

* Resource Provisioning:

→ In this section, we will discuss techniques to provision computer resources or VMs.

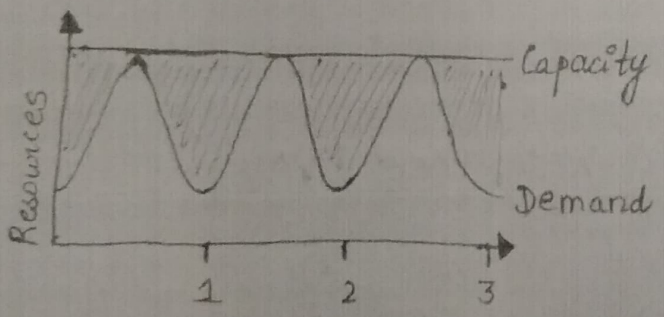
(i) Provisioning of compute resources (VMs).

→ Providers supply cloud services by signing SLAs with end users. The SLAs must commit sufficient resources such as CPU, memory & bandwidth that the user can use for a preset period. Underprovisioning 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.

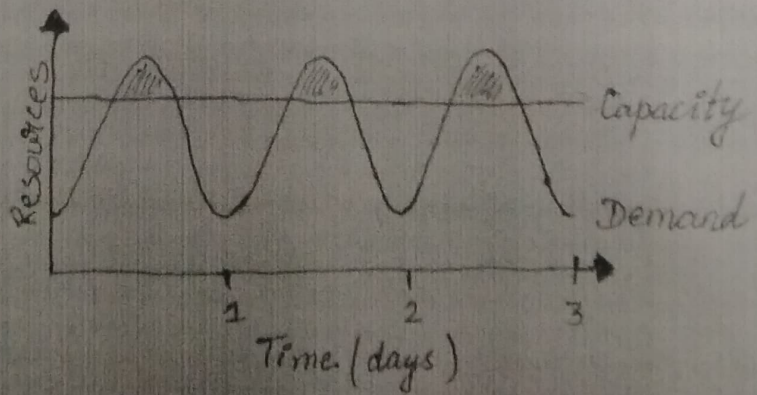
→ The difficulty comes from the unpredictability of consumer demand, soft. & hardware failures, heterogeneity of services, power manag.; & conflicts in signed SLAs b/w consumers & service providers.

→ Efficient VM provisioning depends on the cloud architecture and manag. of cloud infrastructures. Resource provisioning schemes also demand fast discovery of services and data in cloud computing infrastructures. In a virtualized cluster of servers, this demands efficient installation of VMs, live VM migration & fast recovery from failures.

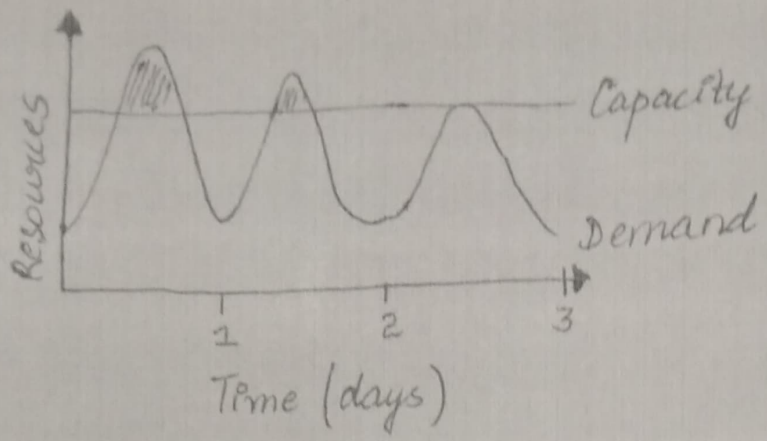
(ii) Resource Provisioning Methods :



(a) Provisioning for peak load



(b) Underprovisioning-1



(c) Underprovisioning 2

→ It shows three cases of static cloud resource provisioning policies. In case (a), overprovisioning with the peak load causes heavy resource waste (shaded area). In case (b), under-provisioning (along the capacity line) of resources results in losses by both user & provider in that paid demand by the users (the shaded area above the capacity). In case (c), the constant provisioning of resources with fixed capacity to a declining user demand could result in even worse resource waste. The user may give up the sevice by canceling the demand, resulting in reduced revenue for the provider. Both the user & provider may be losers in resource provisioning without elasticity.

→ Three resource-provisioning methods are presented in the following sections. The

demand-driven method provides static resources & has been used in grid computing for many years. The event-driven method is based on predicted workload by time. The popularity-driven method is based on Internet traffic monitored.

(a) Demand-driven Resource provisioning:

→ This method add or removes computing instances based on the current utilization level of the allocated resources. The demand driven method automatically allocates two Xeon processors for the user app., when the user was using one Xeon processor more than 60% of the time for an extended period. In general, when a resource has surpassed a threshold for a certain amt. of time, the scheme inc. that resource based on demand. When a resource is below a threshold for a certain amt. of time, that resource would be decreased accordingly.

(b) Event-driven resource provisioning:

→ This scheme add or removes mach. instances based on a specific time event. This scheme anticipates peak traffic before it happens. The method results in a minimal

loss of QoS, if the event is predicted correctly. otherwise, wasted resources are even greater due to events that do not follow a fixed pattern.

(c) → POPULARITY - DRIVEN Resource Provisioning:
In this method, the Internet searches for popularity of certain applications & creates the instances by popularity demand. The scheme anticipates inc. traffic with popularity. Again, the scheme has a minimal loss of QoS, if the predicted popularity is correct. Resources may be wasted if traffic does not occur as expected.

* Global Exchange of Cloud resources:

→ In order to support a large no. of application service consumers from around the world, cloud infrastructure providers (i.e., IaaS providers) have established data centres in multiple geographical location to provide redundancy & ensure reliability in case of site failures.

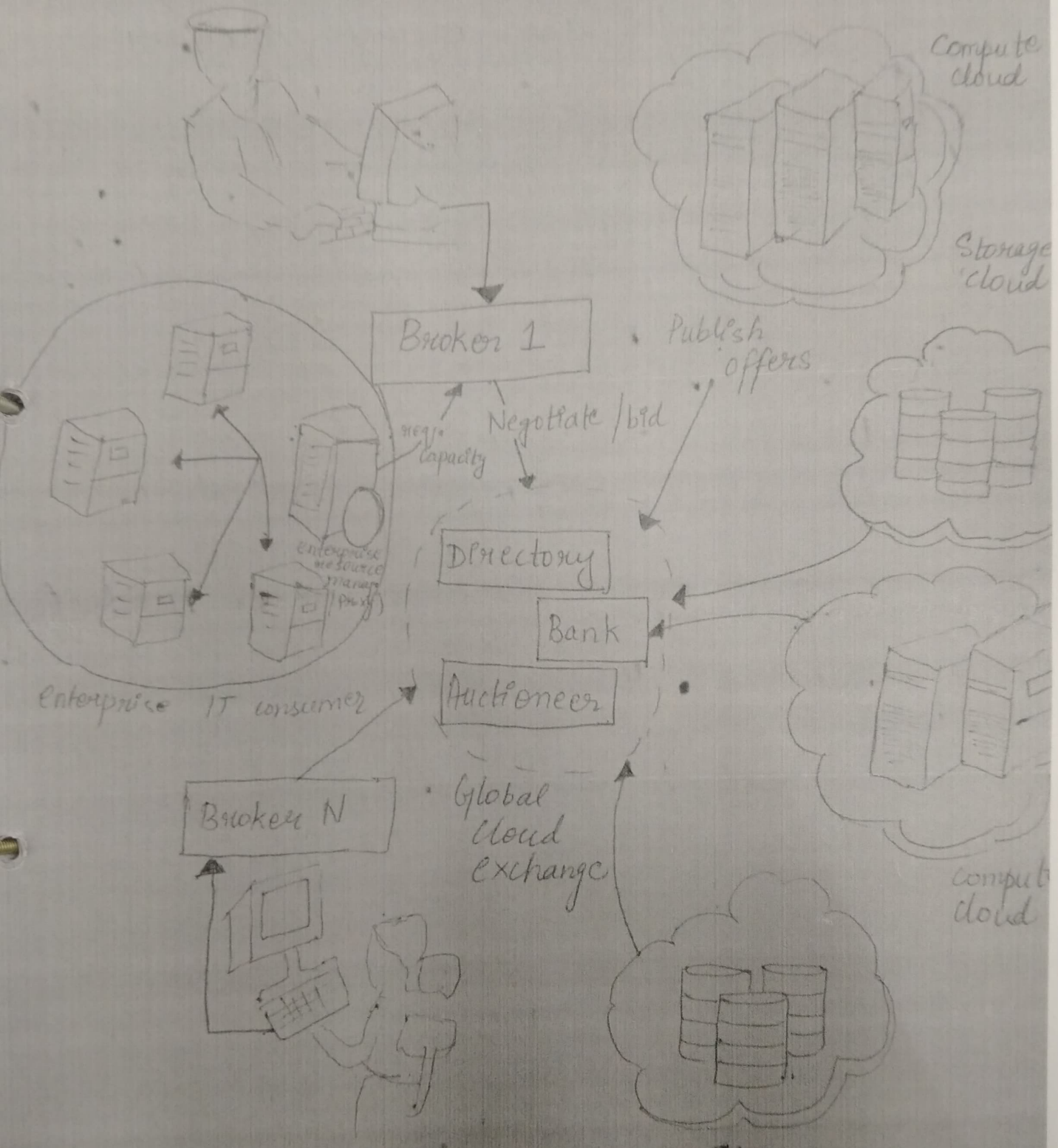
→ This approach has many shortcomings. First, it is difficult for cloud customers to determine in advance the best location for hosting their services as they may not know the origin of consumers of their services. Second, SaaS providers may not

The top layer offers SaaS. These must be implemented on the cloud platforms provided. Although the three basic models are dissimilar in usage. They are built one on top of another. The implication is that one cannot launch SaaS applications with a cloud platform. The cloud platform cannot be built if compute & storage infrastructures are not there.

→ The bottom three layers are more related to physical requirements. The bottommost 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 collection service to house, power, & secure all the physical hardware and network resources. Some authors say this layer provides Security as a Service ("SaaS"). The cloud infrastructure layer can be further subdivided as Data as a service (DaaS) & communication as a service (CaaS) in addition to compute & Storage in IaaS.

→ Cloud players are divided into 3 classes:
(1) Cloud service providers & IT administrators
(2) Software developers or vendors, and

architecture supporting brokering and exch. of cloud resources for scaling app. across multiple clouds service provider;



Inter-cloud exchange of cloud resources through brokering.

operate as a part of a market-driven resource leasing federation where application service providers such as Salesforce.com host their services based on negotiated SLA contracts driven by competitive market prices and delivery-on-demand, reliable, cost-effective, & QoS-aware services based on virtualization technologies while ensuring & high QoS standards & minimizing service cost. They need to be able to utilize market-based utility models as the basis for provisioning of virtualized soft. services & federated hardware inf. among users with heterogeneous applications.

→ The cloud exch. (CEx) acts as a market maker for bringing together service producers & consumers. It aggregates the inf. demands from app. brokers & evalutes them against the available supply currently published by the cloud coordinators. It supports trading of cloud services based on competitive economic models such as commodity markets & auctions. CEx allows participants to locate providers & consumers with matching offers.