

Cloud Infrastructure and Cloud Resource Management

Unit -2

ARCHITECTURAL DESIGN OF COMPUTE AND STORAGE CLOUDS

- An Internet cloud (CC) is envisaged as a public cluster of servers allocated on demand to perform collective web services or distributed apps using the resources of a data center.
- **Cloud Platform Design Goals:**
- The major goals of a cloud computing platform are:-
 - **scalability**
 - **efficiency**
 - **Virtualization**
 - **reliability**

- A cloud platform manager receives the user requests, finds the resources, and calls the provisioning services to allocate the appropriate amount of resources for the job.
- Note that a manager supports both physical and virtual machines.
- Security in shared resources and shared access of data centers also pose another design challenge.
- The platform also needs to establish an infrastructure that can obtain HPC.
- Scalability can be obtained by adding more data centers or servers, which leads to more efficient data distribution and, usage of less power and bandwidth.

- **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 system utilization via **multiplexing, virtualization, and dynamic resource provisioning**.
- Clouds are enabled by the progress in hardware, software, and networking technologies summarized

- In the hardware area, the rapid progress in multicore CPUs, memory chips, and disk arrays has made it possible to build faster data centers with huge amounts of storage space.
- Resource virtualization enables rapid cloud deployment and disaster recovery. Service-oriented architecture (SOA) also plays a vital role.

Table 4.3 Cloud-Enabling Technologies in Hardware, Software, and Networking

Technology	Requirements and Benefits
Fast platform deployment	Fast, efficient, and flexible deployment of cloud resources to provide dynamic computing environment to users
Virtual clusters on demand	Virtualized cluster of VMs provisioned to satisfy user demand and virtual cluster reconfigured as workload changes
Multitenant techniques	SaaS for distributing software to a large number of users for their simultaneous use and resource sharing if so desired
Massive data processing	Internet search and web services which often require massive data processing, especially to support personalized services
Web-scale communication	Support for e-commerce, distance education, telemedicine, social networking, digital government, and digital entertainment applications
Distributed storage	Large-scale storage of personal records and public archive information which demands distributed storage over the clouds
Licensing and billing services	License management and billing services which greatly benefit all types of cloud services in utility computing

A Generic Cloud Architecture

- The Internet cloud is envisioned as a massive cluster of servers.
- These servers are provisioned on demand to perform collective web services or distributed applications using data-center resources.
- The cloud platform is formed dynamically by provisioning or deprovisioning servers, software, and database resources.
- Servers in the cloud can be physical machines or VMs.
- User interfaces are applied to request services.
- The provisioning tool carves out the cloud system to deliver the requested service.
- In addition to building the server cluster, the cloud platform demands distributed storage and accompanying services

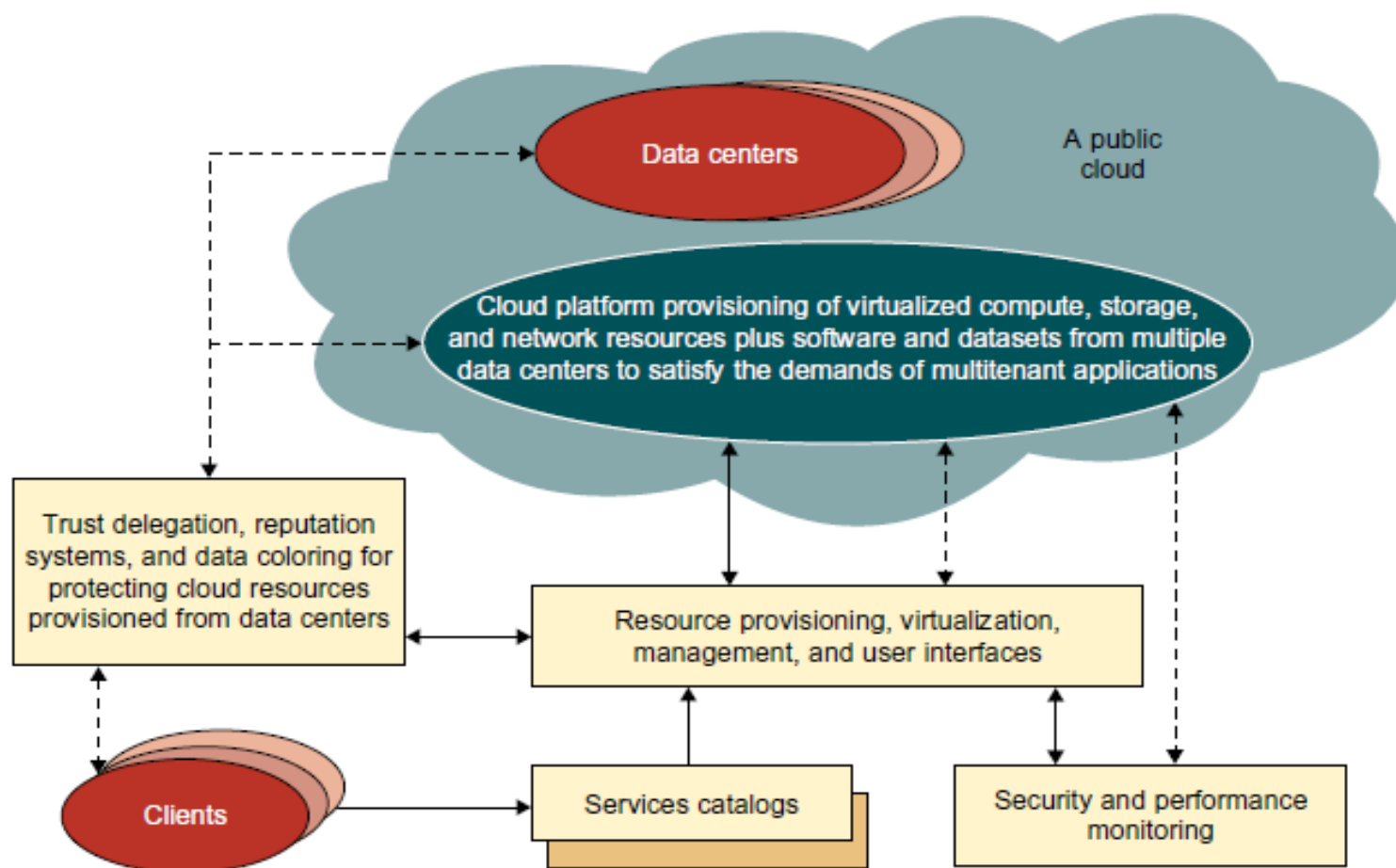


FIGURE 4.14

A security-aware cloud platform built with a virtual cluster of VMs, storage, and networking resources over the data-center servers operated by providers.

Layered Cloud Architectural Development

- 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 networking support over the Internet and intranets involved.
- It is clear that the **infrastructure layer is deployed first** to support IaaS services.
- This **infrastructure layer serves as the foundation for building the platform layer** of the cloud for supporting PaaS services.
- In turn, the platform layer is a foundation for implementing the application layer for SaaS applications.
- Different types of cloud services demand application of these resources separately.

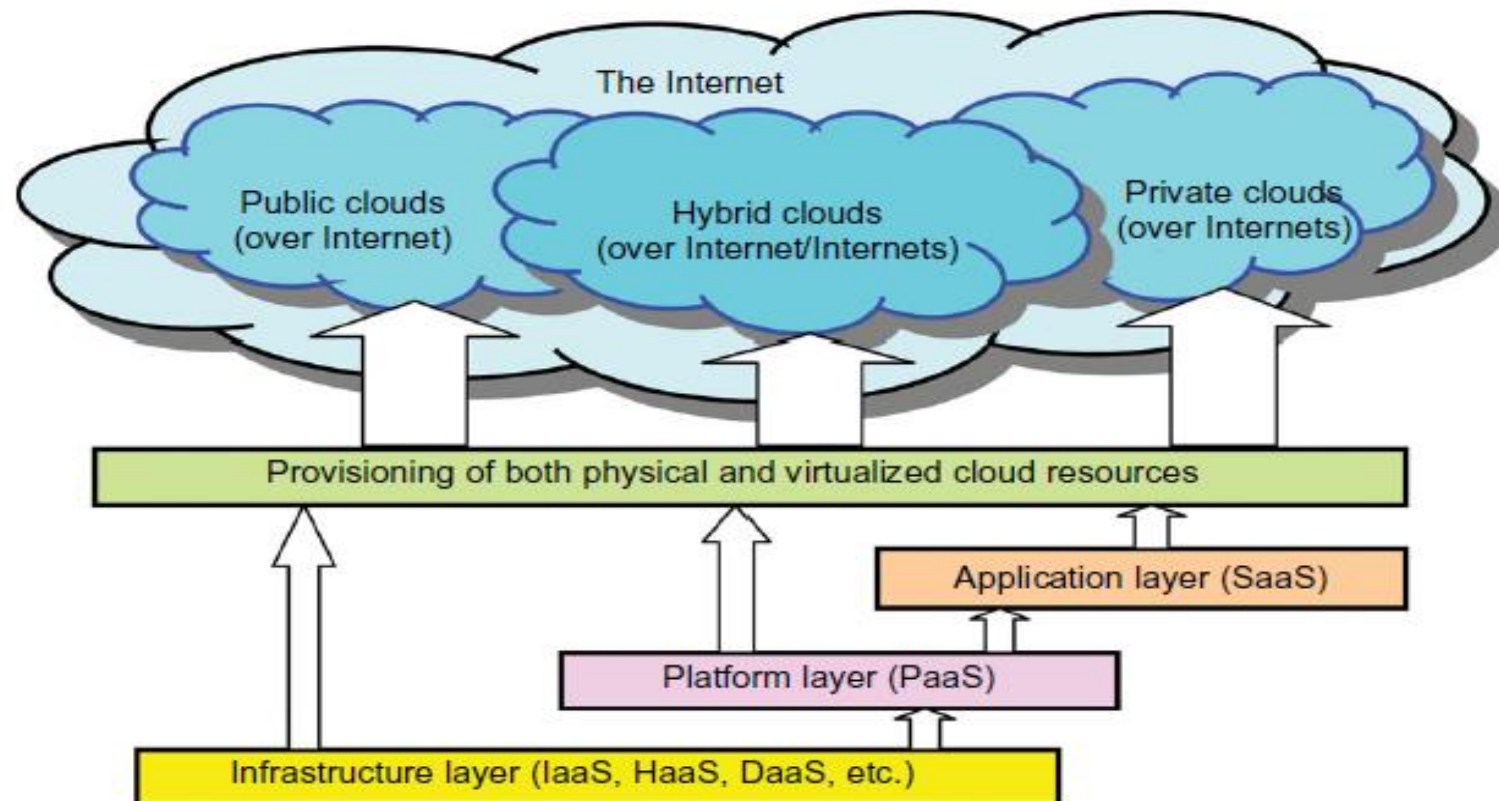


FIGURE 4.15

Layered architectural development of the cloud platform for IaaS, PaaS, and SaaS applications over the Internet.

- **The infrastructure layer** is built with virtualized compute, storage, and network resources.
- The abstraction of these hardware resources is meant to provide the flexibility demanded by users.
- Internally, virtualization realizes automated provisioning of resources and optimizes the infrastructure management process.

- **The platform layer** is for general-purpose and repeated usage of the collection of software resources.
- This layer provides users with an environment to develop their applications, to test operation flows, and to monitor execution results and performance.
- The platform should be able to assure users that they have scalability, dependability, and security protection.
- In a way, the virtualized cloud platform serves as a “system middleware” between the infrastructure and application layers of the cloud

- **The application layer** is formed with a collection of all needed software modules for SaaS applications.
- Service applications in this layer include daily office management work, such as information retrieval, document processing, and calendar and authentication services.
- The application layer is also heavily used by enterprises in business marketing and sales, consumer relationship management (CRM), financial transactions, and supply chain management.
- It should be noted that not all cloud services are restricted to a single layer.
- Many applications may apply resources at mixed layers.
- After all, the three layers are built from the bottom up with a dependence relationship.

- From the provider's perspective, the services at various layers demand different amounts of functionality support and resource management by providers.
- In general, SaaS demands the most work from the provider, PaaS is in the middle, and IaaS demands the least.
- For example, Amazon EC2 provides not only virtualized CPU resources to users, but also management of these provisioned resources.
- Services at the application layer demand more work from providers.
- The best example of this is the Salesforce.com CRM service, in which the provider supplies not only the hardware at the bottom layer and the software at the top layer, but also the platform and software tools for user application development and monitoring.

Architectural Design Challenges

- There are six open challenges of cloud architectural development
- **Challenge 1—Service Availability and Data Lock-in Problem**
- The management of a cloud service by a single company is often the source of single points of failure.
- To achieve HA, one can consider using multiple cloud providers.
- Even if a company has multiple data centers located in different geographic regions, it may have common software infrastructure and accounting systems.
- Therefore, using multiple cloud providers may provide more protection from failures.
- Another availability obstacle is distributed denial of service (DDoS) attacks.

- Criminals threaten to cut off the incomes of SaaS providers by making their services unavailable. Some utility computing services offer SaaS providers the opportunity to defend against DDoS attacks by using quick scale-ups.
- Software stacks have improved interoperability among different cloud platforms, but the APIs itself are still proprietary.
- Thus, customers cannot easily extract their data and programs from one site to run on another.
- The obvious solution is to standardize the APIs so that a SaaS developer can deploy services and data across multiple cloud providers.
- This will rescue the loss of all data due to the failure of a single company.
- In addition to mitigating data lock-in concerns, standardization of APIs enables a new usage model in which the same software infrastructure can be used in both public and private clouds.
- Such an option could enable “surge computing,” in which the public cloud is used to capture the extra tasks that cannot be easily run in the data center of a private cloud.

- **Challenge 2—Data Privacy and Security Concerns**

- Current cloud offerings are essentially public (rather than private) networks, exposing the system to more attacks.
- Many obstacles can be overcome immediately with well-understood technologies such as encrypted storage, virtual LANs, and network middleboxes (e.g., firewalls, packet filters).
- For example, you could encrypt your data before placing it in a cloud.
- Many nations have laws requiring SaaS providers to keep customer data and copyrighted material within national boundaries.
- Traditional network attacks include buffer overflows, DoS attacks, spyware, malware, rootkits, Trojan horses, and worms.
- In a cloud environment, newer attacks may result from hypervisor malware guest hopping and hijacking, or VM rootkits

- Another type of attack is the man-in-the-middle attack for VM migrations.
- In general, passive attacks steal sensitive data or passwords.
- Active attacks may manipulate kernel data structures which will cause major damage to cloud servers

- **Challenge 3—Unpredictable Performance and Bottlenecks**

- Multiple VMs can share CPUs and main memory in cloud computing, but I/O sharing is problematic.
- For example, to run 75 EC2 instances with the STREAM benchmark requires a mean bandwidth of 1,355 MB/second. However, for each of the 75 EC2 instances to write 1 GB files to the local disk requires a mean disk write bandwidth of only 55 MB/second.
- This demonstrates the problem of I/O interference between VMs.
- One solution is to improve I/O architectures and operating systems to efficiently virtualize interrupts and I/O channels.

- Internet applications continue to become more data-intensive.
- If we assume applications to be “pulled apart” across the boundaries of clouds, this may complicate data placement and transport.
- Cloud users and providers have to think about the implications of placement and traffic at every level of the system, if they want to minimize costs.
- This kind of reasoning can be seen in Amazon’s development of its new CloudFront service.
- Therefore, data transfer bottlenecks must be removed, bottleneck links must be widened, and weak servers should be removed.

- **Challenge 4—Distributed Storage and Widespread Software Bugs**

- The database is always growing in cloud applications.
- The opportunity is to create a storage system that will not only meet this growth, but also combine it with the cloud advantage of scaling arbitrarily up and down on demand.
- This demands the design of efficient distributed SAN(storage area network).
- Data centers must meet programmers' expectations in terms of scalability, data durability, and HA.
- Data consistence checking in SAN-connected data centers is a major challenge in cloud computing.

- Large-scale distributed bugs cannot be reproduced, so the debugging must occur at a scale in the production data centers.
- No data center will provide such a convenience.
- One solution may be a reliance on using VMs in cloud computing.
- The level of virtualization may make it possible to capture valuable information in ways that are impossible without using VMs.
- Debugging over simulator is another approach to attacking the problem, if the simulator is well designed.

- **Challenge 5—Cloud Scalability, Interoperability, and Standardization**

- The pay-as-you-go model applies to storage and network bandwidth, both are counted in terms of the number of bytes used.
- Computation is different depending on virtualization level.
- GAE(Google App Engine) automatically scales in response to load increases and decreases; users are charged by the cycles used.
- AWS charges by the hour for the number of VM instances used, even if the machine is idle.
- The opportunity here is to scale quickly up and down in response to load variation, in order to save money, but without violating SLAs.

- Open Virtualization Format (OVF) describes an open, secure, portable, efficient, and extensible format for the packaging and distribution of VMs.
- It also defines a format for distributing software to be deployed in VMs.
- This VM format does not rely on the use of a specific host platform, virtualization platform, or guest operating system.
- The approach is to address virtual platform-agnostic packaging with certification and integrity of packaged software.
- The package supports virtual appliances to span more than one VM.

- **Challenge 6—Software Licensing and Reputation Sharing**

- Many cloud computing providers originally relied on open source software because the licensing model for commercial software is not ideal for utility computing.
- The primary opportunity is either for open source to remain popular or simply for commercial software companies to change their licensing structure to better fit cloud computing.
- One can consider using both pay-for-use and bulk-use licensing schemes to widen the business coverage.

- One customer's bad behaviour can affect the reputation of the entire cloud.
- For instance, blacklisting of EC2 IP addresses by spam-prevention services may limit smooth VM installation.
- An opportunity would be to create reputation-guarding services similar to the "trusted e-mail" services currently offered (for a fee) to services hosted on smaller ISPs.
- Another legal issue concerns the transfer of legal liability.
- Cloud providers want legal liability to remain with the customer, and vice versa.
- This problem must be solved at the SLA level. We will study reputation systems for protecting data centers in the next section.

INTER-CLOUD RESOURCE MANAGEMENT

- **Extended Cloud Computing Services:-**
- There are six layers of cloud services, ranging from hardware, network, and collocation to infrastructure, platform, and software applications.
- We already introduced the top three service layers as SaaS, PaaS, and IaaS, respectively.
- The cloud platform provides PaaS, which sits on top of the IaaS infrastructure.
- 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 and storage infrastructures are not there.

Cloud application (SaaS)			Concur, RightNOW, Teleo, Kenexa, Webex, Blackbaud, salesforce.com, Netsuite, Kenexa, etc.
Cloud software environment (PaaS)			Force.com, App Engine, Facebook, MS Azure, NetSuite, IBM BlueCloud, SGI Cyclone, eBay
Cloud software infrastructure			Amazon AWS, OpSource Cloud, IBM Ensembles, Rackspace cloud, Windows Azure, HP, Banknorth
Computational resources (IaaS)	Storage (DaaS)	Communications (Caas)	
Collocation cloud services (LaaS)			Savvis, Internap, NTTCommunications, Digital Realty Trust, 365 Main
Network cloud services (NaaS)			Owest, AT&T, AboveNet
Hardware/Virtualization cloud services (HaaS)			VMware, Intel, IBM, XenEnterprise

FIGURE 4.23

A stack of six layers of cloud services and their providers.

- 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 collocation service to house, power, and 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) and Communication as a Service (CaaS) in addition to compute and storage in IaaS.

- cloud players are divided into three classes:
- (1) cloud service providers and IT administrators
- (2) software developers or vendors
- (3) end users or business users
- These cloud players vary in their roles under the IaaS, PaaS, and SaaS models.
- From the software vendors' perspective, application performance on a given cloud platform is most important.

Table 4.7 Cloud Differences in Perspectives of Providers, Vendors, and Users			
Cloud Players	IaaS	PaaS	SaaS
IT administrators/cloud providers	Monitor SLAs	Monitor SLAs and enable service platforms	Monitor SLAs and deploy software
Software developers (vendors)	To deploy and store data	Enabling platforms via configurators and APIs	Develop and deploy software
End users or business users	To deploy and store data	To develop and test web software	Use business software

- **Cloud Service Tasks and Trends:-**

- Cloud services are introduced in five layers. The top layer is for SaaS applications, as further subdivided into the five application areas
- For example, CRM is heavily practiced in business promotion, direct sales, and marketing services. CRM offered the first SaaS on the cloud successfully. The approach is to widen market coverage by investigating customer behaviours and revealing opportunities by statistical analysis.
- SaaS tools also apply to distributed collaboration, and financial and human resources management.
- These cloud services have been growing rapidly in recent years.
- PaaS is provided by Google, [Salesforce.com](https://www.salesforce.com), and Facebook, among others.
- IaaS is provided by Amazon, Windows Azure, and RackRack, among others.
- Collocation services require multiple cloud providers to work together to support supply chains in manufacturing.
- Network cloud services provide communications such as those by AT&T, Qwest, and AboveNet.

- **Software Stack for Cloud Computing:-**

- Despite the various types of nodes in the cloud computing cluster, the overall software stacks are built from scratch to meet rigorous goals .
- Developers have to consider how to design the system to meet critical requirements such as high throughput, HA, and fault tolerance.
- Even the operating system might be modified to meet the special requirement of cloud data processing.
- Based on the observations of some typical cloud computing instances, such as Google Microsoft, and Yahoo!, the overall software stack structure of cloud computing software can be viewed as layers.
- Each layer has its own purpose and provides the interface for the upper layers just as the traditional software stack does.
- However, the lower layers are not completely transparent to the upper layers.

- **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.
- One of the most important facilities is the cluster job management system is the **scheduler queues** the tasks submitted to the whole cluster and assigns the tasks to the processing nodes according to node availability.
- The distributed scheduler for the cloud application has special characteristics that can support cloud applications, such as scheduling the programs written in MapReduce style.
- The runtime support system keeps the cloud cluster working properly with high efficiency.

- Runtime support is software needed in browser-initiated applications applied by thousands of cloud customers.
- The SaaS model provides the software applications as a service, rather than letting users purchase the software.
- As a result, on the customer side, there is no upfront investment in servers or software licensing.
- On the provider side, costs are rather low, compared with conventional hosting of user applications.
- The customer data is stored in the cloud that is either vendor proprietary or a publicly hosted cloud supporting PaaS and IaaS.

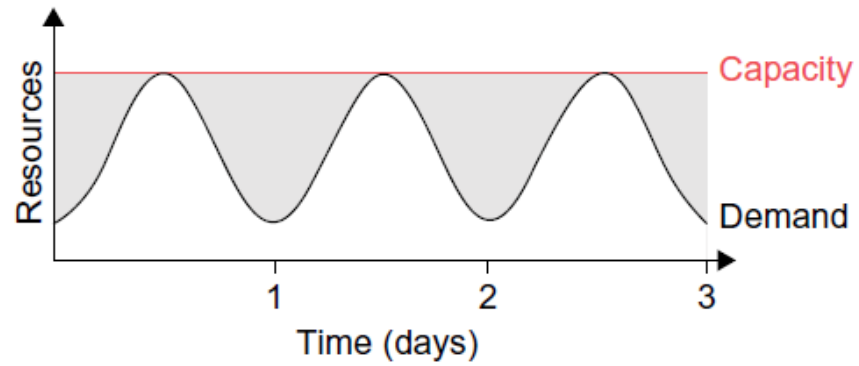
Resource Provisioning and Platform Deployment

- The emergence of computing clouds suggests fundamental changes in software and hardware architecture.
- Cloud architecture puts more emphasis on the number of processor cores or VM instances.
- **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, and bandwidth that the user can use for a present 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.

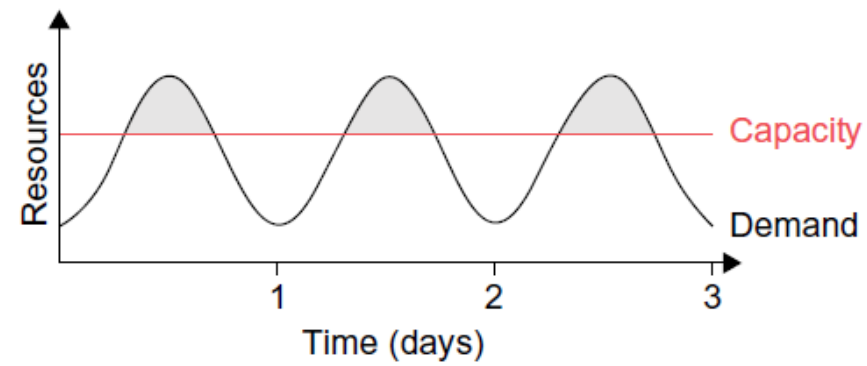
- Deploying an autonomous system to efficiently provision resources to users is a challenging problem.
- The difficulty comes from the **unpredictability of consumer demand, software and hardware failures, heterogeneity of services, power management, and conflicts in signed SLAs between consumers and service providers.**
- Efficient VM provisioning depends on the **cloud architecture and management 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, and fast recovery from failures.
- To deploy VMs, users treat them as physical hosts with customized operating systems for specific applications.

- **Resource Provisioning Methods:-**

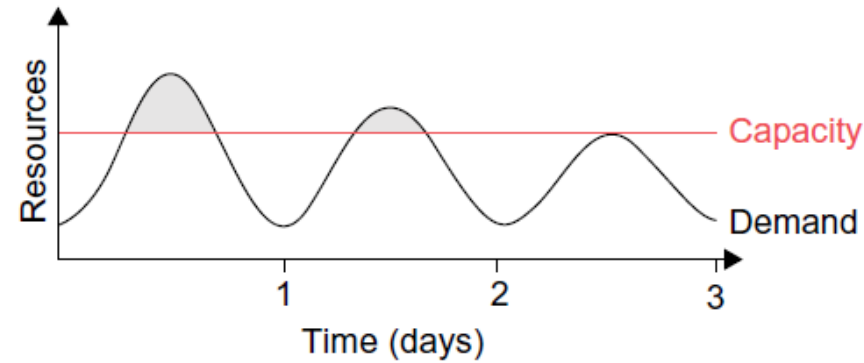
- three cases of static cloud resource provisioning policies.
- In **case (a), overprovisioning** with the peak load causes heavy resource waste (shaded area).
- In **case (b), under provisioning** (along the capacity line) of resources results in losses by both user and provider in that paid demand by the users (the shaded area above the capacity) is not served and wasted resources still exist for those demanded areas below the provisioned 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 service by cancelling the demand, resulting in reduced revenue for the provider.
- Both the user and provider may be losers in resource provisioning without elasticity



(a) Provisioning for peak load



(b) Underprovisioning 1



(c) Underprovisioning 2

FIGURE 4.24

Three cases of cloud resource provisioning without elasticity: (a) heavy waste due to overprovisioning, (b) underprovisioning and (c) under- and then overprovisioning.

- Three resource-provisioning methods are there.
- The **demand-driven method** provides static resources and 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

- **Demand-Driven Resource Provisioning:-**

- This method adds 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 application, when the user was using one Xeon processor more than 60 percent of the time for an extended period.
- In general, when a resource has surpassed a threshold for a certain amount of time, the scheme increases that resource based on demand.
- When a resource is below a threshold for a certain amount of time, that resource could be decreased accordingly.
- Amazon implements such an auto-scale feature in its EC2 platform.
- This method is easy to implement.
- The scheme does not work out right if the workload changes abruptly.

- **Event-Driven Resource Provisioning**

- This scheme adds or removes machine instances based on a specific time event.
- The scheme works better for seasonal or predicted events such as Christmastime in the West and the Lunar New Year in the East.
- During these events, the number of users grows before the event period and then decreases during the event period.
- 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

- **Popularity-Driven Resource Provisioning**

- In this method, the Internet searches for popularity of certain applications and creates the instances by popularity demand.
- The scheme anticipates increased 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.

- **Dynamic Resource Deployment:-**

- The cloud uses VMs as building blocks to create an execution environment across multiple resource sites.
- The **InterGrid-managed infrastructure** was developed by a Melbourne University group .
- Dynamic resource deployment can be implemented to achieve scalability in performance.
- The Inter- Grid is a **Java-implemented software system** that lets users create execution cloud environments on top of all participating grid resources.
- **Peering arrangements** established between gateways enable the allocation of resources from multiple grids to establish the execution environment.

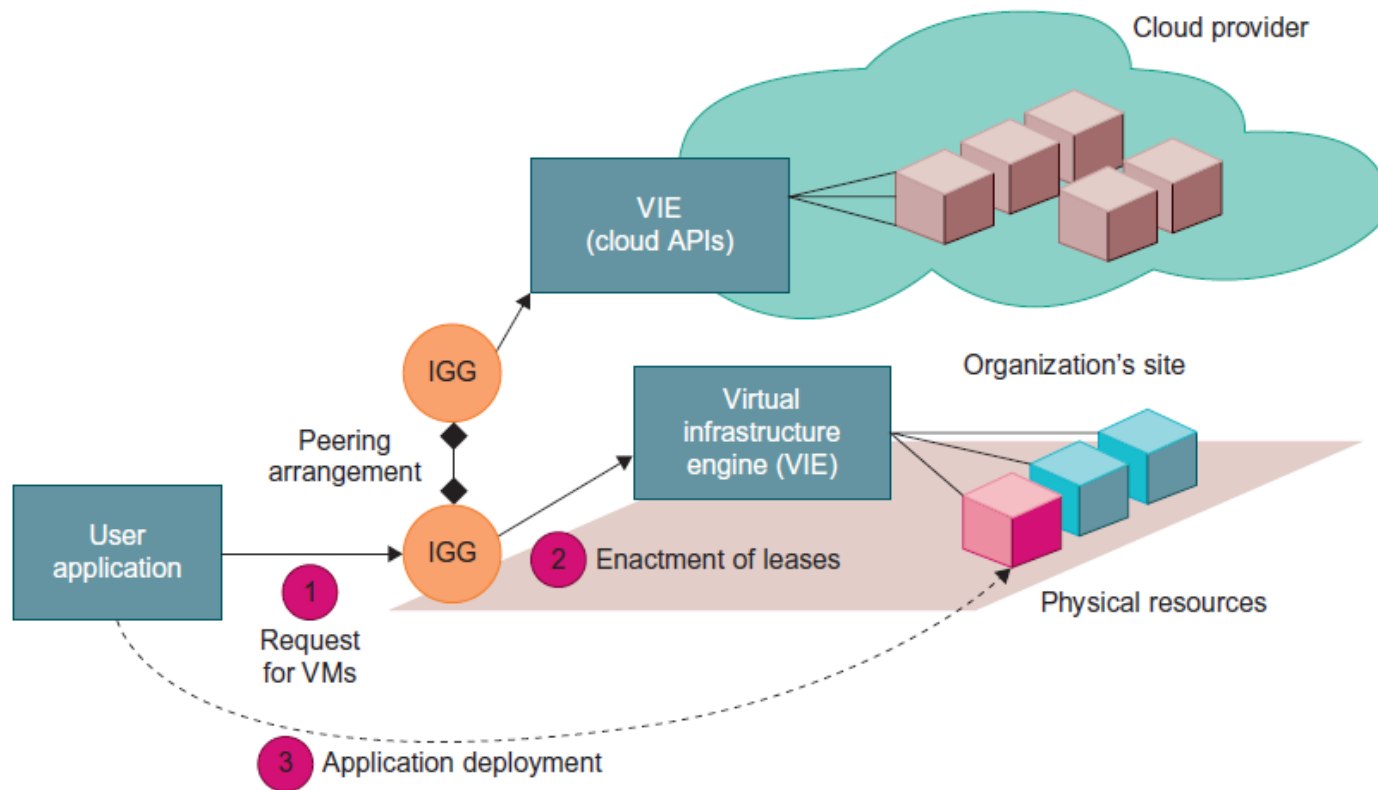


FIGURE 4.26

Cloud resource deployment using an IGG (intergrid gateway) to allocate the VMs from a Local cluster to interact with the IGG of a public cloud provider.

- In a scenario is illustrated by which an InterGrid gateway (IGG) allocates resources from a local cluster to deploy applications in three steps:
 - **(1) requesting the VMs**
 - **(2) enacting the leases**
 - **(3) deploying the VMs as requested.**
- Under peak demand, this IGG interacts with another IGG that can allocate resources from a cloud computing provider.

- A grid has **predefined peering arrangements with other grids**, which the IGG manages. Through multiple IGGs, the system coordinates the use of InterGrid resources.
- An IGG is aware of the peering terms with other grids, selects suitable grids that can provide **the required resources, and replies to requests from other IGGs**.
- Request redirection policies determine which peering grid InterGrid selects to process a request and a price for which that grid will perform the task.
- An IGG can also allocate resources from a cloud provider.
- The cloud system creates a virtual environment to help users deploy their applications. These applications use the distributed grid resources.

- The InterGrid allocates and provides **a distributed virtual environment (DVE)**.
- This is a virtual cluster of VMs that runs isolated from other virtual clusters.
- A component called **the DVE manager performs resource allocation and management on behalf of specific user applications**.
- The core component of the IGG is **a scheduler** for implementing provisioning policies and peering with other gateways.
- The communication component provides an asynchronous message-passing mechanism.
- Received messages are handled in parallel by a thread pool.

- **Provisioning of Storage Resources:-**

- The data storage layer is built on **top of the physical or virtual servers.**
- As the cloud computing applications often provide service to users, it is unavoidable that the data is stored in the clusters of the cloud provider.
- The service can be accessed anywhere in the world. One example is e-mail systems. Another example is a web searching application.
- In storage technologies, hard disk drives may be augmented with solid-state drives in the future.
- This will provide reliable and high-performance data storage.

- A **distributed file system** is very important for storing **large-scale data**.
- However, other forms of data storage also exist. Some data does not need the namespace of a tree structure file system, and instead, databases are built with stored data files.
- In cloud computing, another form of data storage is (Key, Value) pairs.
- Amazon S3 service uses SOAP(messaging protocol layer) to access the objects stored in the cloud.

Table 4.8 Storage Services in Three Cloud Computing Systems

Storage System	Features
GFS: Google File System	Very large sustainable reading and writing bandwidth, mostly continuous accessing instead of random accessing. The programming interface is similar to that of the POSIX file system accessing interface.
HDFS: Hadoop Distributed File System	The open source clone of GFS. Written in Java. The programming interfaces are similar to POSIX but not identical.
Amazon S3 and EBS	S3 is used for retrieving and storing data from/to remote servers. EBS is built on top of S3 for using virtual disks in running EC2 instances.

- Many cloud computing companies have developed large-scale data storage systems to keep huge amount of data collected every day.
- For example, **Google's GFS** stores web data and some other data, such as geographic data for Google Earth.
- A similar system from the open source community is the **Hadoop Distributed File System (HDFS)** for Apache. Hadoop is the open source implementation of Google's cloud computing infrastructure.
- Similar systems include Microsoft's **Cosmos file system** for the cloud.
- Despite the fact that the storage service or distributed file system can be accessed directly, similar to traditional databases, cloud computing does provide some forms of **structure or semi structure database processing capability**.
- For example, applications might want to process the information contained in a web page. Web pages are an example of semi structural data in HTML format.
- If some forms of database capability can be used, application developers will construct their application logic more easily.

- Another reason to build a database-like service in cloud computing is that **it will be quite convenient for traditional application developers to code for the cloud platform.**
- Databases are quite common as the underlying storage device for many applications.
- Thus, such developers can think in the same way they do for traditional software development.
- Hence, in cloud computing, **it is necessary to build databases like large-scale systems based on data storage or distributed file systems.**
- The scale of such a database might be quite large for processing huge amounts of data. The main purpose is to store the data in structural or semi-structural ways so that application developers can use it easily and build their applications rapidly.
- Typical cloud databases include **BigTable from Google, SimpleDB from Amazon,, and the SQL service from Microsoft Azure.**

Global Exchange of Cloud Resources

- In order to support a large number of application service consumers from around the world, cloud infrastructure providers (i.e., IaaS providers) have established data centers in multiple geographical locations to provide redundancy and ensure reliability in case of site failures.
- For example, Amazon has data centers in the United States (e.g., one on the East Coast and another on the West Coast) and Europe.
- However, currently Amazon expects its cloud customers (i.e., SaaS providers) to express a preference regarding where they want their application services to be hosted.
- Amazon does not provide seamless/automatic mechanisms for scaling its hosted services across multiple geographically distributed data centers.
- 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 be able to meet the QoS expectations of their service**
- consumers originating from multiple geographical locations. This necessitates building mechanisms for seamless federation of data centers of a cloud provider or providers supporting dynamic
- scaling of applications across multiple domains in order to meet QoS targets of cloud customers
- In addition, no single cloud infrastructure provider will be able to establish its data centers at all possible locations throughout the world.
- As a result, cloud application service (SaaS) providers will have difficulty in meeting QoS expectations for all their consumers.

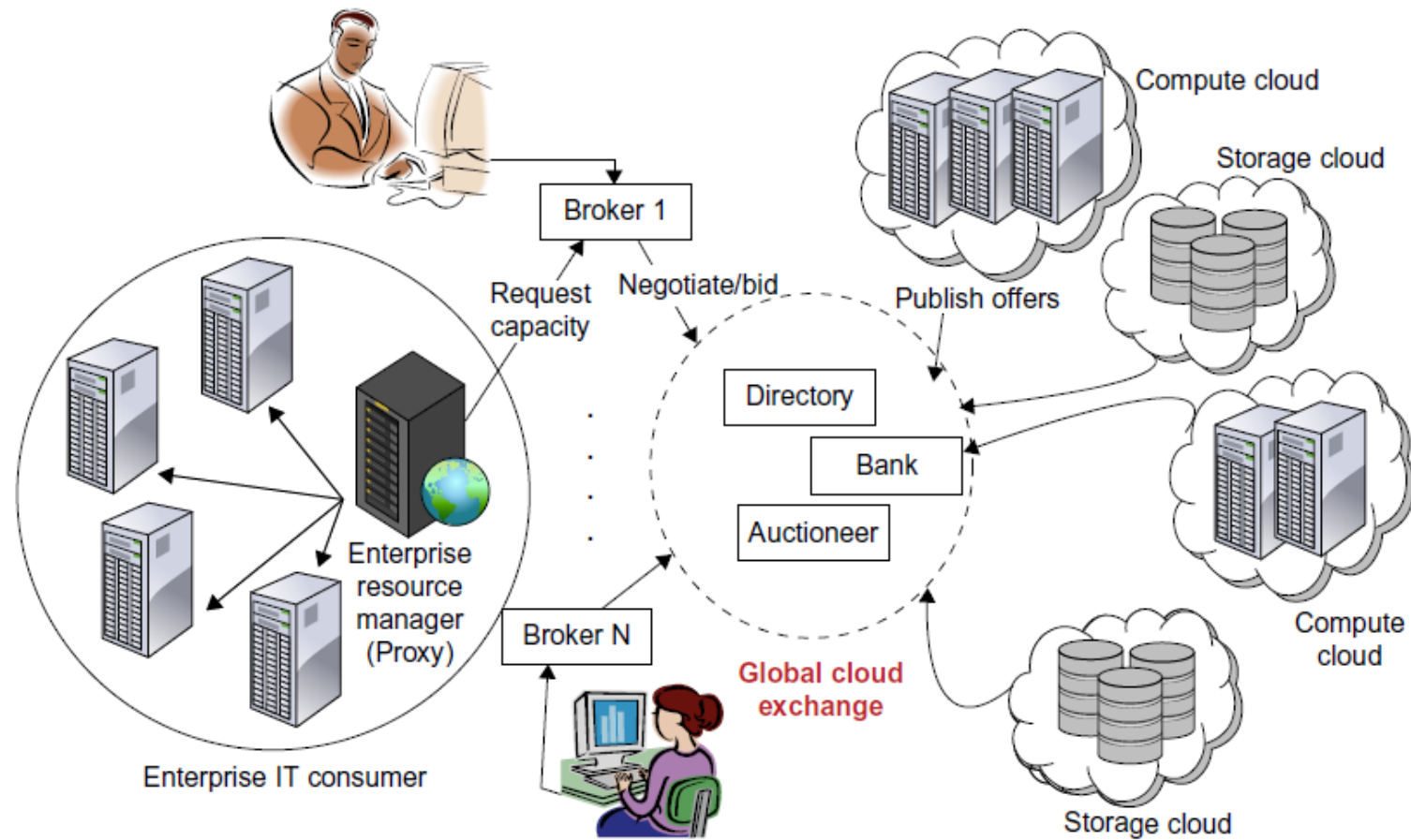


FIGURE 4.30

Inter-cloud exchange of cloud resources through brokering.

- Hence, they would like to make use of services of multiple cloud infrastructure service providers who can provide better support for their specific consumer needs.
- This kind of requirement often arises in enterprises with global operations and applications such as **Internet services, media hosting, and Web 2.0 applications.**
- This necessitates federation of cloud infrastructure service providers or seamless provisioning of services across different cloud providers.
- To realize this, **the Cloudbus Project at the University of Melbourne** has proposed **InterCloud architecture supporting brokering and exchange of cloud resources for scaling applications across multiple clouds.**

- By realizing InterCloud architectural principles in mechanisms in their offering, cloud providers will be able to **dynamically expand or resize their provisioning capability based on sudden spikes in workload demands by leasing available computational and storage capabilities from other cloud service providers;** operate as part of a market-driven resource leasing federation
- where application service providers such as [Salesforce.com](https://www.salesforce.com) host their services based on negotiated SLA contracts driven by competitive market prices; and deliver on-demand, reliable, cost-effective, and QoS-aware services based on virtualization technologies while ensuring high QoS standards and minimizing service costs.
- They need to be able to utilize market-based utility models as the basis for provisioning of virtualized software services and federated hardware infrastructure among users with heterogeneous applications.

- They consist of client brokering and coordinator services that support utility-driven federation of clouds: application scheduling, resource allocation, and migration of workloads.
- **The architecture cohesively couples the administratively and topologically distributed storage and compute capabilities of clouds as part of a single resource leasing abstraction.**
- The system will ease the cross domain capability integration for on-demand, flexible, energy-efficient, and reliable access to the infrastructure based on virtualization technology .
- The **Cloud Exchange (CEx)** 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.
- It supports trading of cloud services based on competitive economic models such as commodity markets and auctions.
- **CEx allows participants to locate providers and consumers with fitting offers.**

- Such markets enable services to be commoditized, and thus will pave the way for creation of **dynamic market infrastructure for trading based on SLAs**.
- An SLA specifies the details of the service to be provided in terms of metrics agreed upon by all parties, and incentives and penalties for meeting and violating the expectations, respectively.
- The availability of a banking system within the market ensures that financial transactions pertaining to SLAs between participants are carried out in a secure and dependable environment.

Administrating the Clouds

- The explosive growth in cloud computing services has led many vendors to rename their products and reposition them to get in on the gold rush in the clouds.
- What was once a network management product is now a cloud management product.
- Nevertheless, this is one area of technology that is very actively funded, comes replete with interesting start-ups, has been the focus of several recent strategic acquisitions, and has resulted in some interesting product alliances.

- These fundamental features are offered by traditional network management systems:
 - **Administration of resources**
 - **Configuring resources**
 - **Enforcing security**
 - **Monitoring operations**
 - **Optimizing performance**
 - **Policy management**
 - **Performing maintenance**
 - **Provisioning of resources**

- Network management systems are often described in terms of the acronym **FCAPS**, which stands for these features:
 - **F**ault
 - **C**onfiguration
 - **A**ccounting
 - **P**erformance
 - **S**ecurity
- Most network management packages have one or more of these characteristics; no single package provides all five elements of FCAPS.
- To get the complete set of all five of these management areas from a single vendor, you would need to adopt a network management framework.
- These large network management frameworks were industry leaders several years back: BMC PATROL, CA Unicenter, IBM Tivoli, HP OpenView, and Microsoft System Center.







































- **Management responsibilities**
- What separates a network management package from a cloud computing management package is the “cloudly” characteristics that cloud management service must have:
 - Billing is on a pay-as-you-go basis.
 - The management service is extremely scalable.
 - The management service is ubiquitous.
 - Communication between the cloud and other systems uses cloud networking standards.

- To monitor an entire cloud computing deployment stack, you monitor six different categories:
- **1. End-user services** such as HTTP, TCP, POP3/SMTP, and others
- **2. Browser performance** on the client
- **3. Application monitoring** in the cloud, such as Apache, MySQL, and so on
- **4. Cloud infrastructure monitoring** of services such as Amazon Web Services, GoGrid, Rackspace, and others
- **5. Machine instance monitoring** where the service measures processor utilization, memory usage, disk consumption, queue lengths, and other important parameters
- **6. Network monitoring and discovery using standard protocols** like the Simple Network Management Protocol (SNMP), Configuration Management Database (CMDB), Windows Management Instrumentation (WMI), and the like

- It's important to note that there are really two aspects to cloud management:
 - **Managing resources *in the cloud***
 - **Using the cloud to manage resources *on-premises***
- When you move to a cloud computing architecture from a traditional networked model like client/server or a three-tier architecture
- many of the old management tasks for processes going on in the cloud become irrelevant or nearly impossible to manage because the tools to effectively manage resources of various kinds fall outside of your own purview.
- In the cloud, the particular service model you are using directly affects the type of monitoring you are responsible for.

- The situation—as you move first to Platform as a Service (PaaS) like Windows Azure or Google App Engine and then onto Software as a Service (SaaS) for which Salesforce.com is a prime example— becomes even more restrictive.
- When you deploy an application on Google’s PaaS App Engine cloud service, the Administration Console provides you with the following monitoring capabilities:
 - Create a new application, and set it up in your domain.
 - Invite other people to be part of developing your application.
 - View data and error logs.
 - Analyse your network traffic.
 - Browse the application datastore, and manage its indexes.
 - View the application’s scheduled tasks.
 - Test the application, and swap out versions.

Management responsibilities by service model type

	Hosted	Managed services	Cloud (IaaS)	Cloud (PaaS)	SaaS
Example(s)	Hosted infrastructure	Network VoIP	Amazon AWS, Rackspace Cloud server	Google App Engine, Microsoft Azure	Salesforce.com
IT primary responsibilities	   		   	 	
Provider primary responsibilities		<i>Varies by business agreement</i>  		  	  
Shared responsibilities		    			 
 Business service/ user satisfaction	 Application	 Database	 Server	 Operating system	 Network

- The second aspect of cloud management is the role that cloud-based services can play in **managing on-premises resources**.
- From the standpoint of the client, a cloud service provider is no different than any other networked service.
- The full range of network management capabilities may be brought to bear to solve mobile, desktop, and local server issues, and the same sets of tools can be used for measurement.
- **Microsoft System Center** is an example of how management products are being adapted for the cloud.
- System Center provides tools for managing Windows servers and desktops.
- The management services include an Operations Manager, the Windows Service Update Service (WSUS), a Configuration Manager for asset management, a Data Protection Manager, and a Virtual Machine Manager, among other components.

- **Lifecycle management**
- Cloud services have a defined lifecycle, just like any other system deployment. A management program has to touch on each of the six different stages in that lifecycle:
- **1. The definition of the service as a template** for creating instances Tasks performed in Phase 1 include the creation, updating, and deletion of service templates.
- **2. Client interactions with the service**, usually through an SLA (Service Level Agreement) contract This phase manages client relationships and creates and manages service contracts.
- **3. The deployment of an instance to the cloud and the runtime management of instances** Tasks performed in Phase 3 include the creation, updating, and deletion of service offerings.

- **4. The definition of the attributes of the service** while in operation and performance of modifications of its properties The chief task during this management phase is to perform service optimization and customization.
- **5. Management of the operation of instances and routine maintenance** During Phase 5, you must monitor resources, track and respond to events, and perform reporting and billing functions.
- **6. Retirement of the service** End of life tasks include data protection and system migration, archiving, and service contract termination

Cloud Management Products

- Cloud management software and services is a very young industry, and as such, it has a very large number of companies, some with new products and others with older products competing in this area.
- When considering products in cloud management, you should be aware that—as in all new areas of technology—there is considerable churn as companies grow, get acquired, or fail along the way.

TABLE 11.1 (continued)

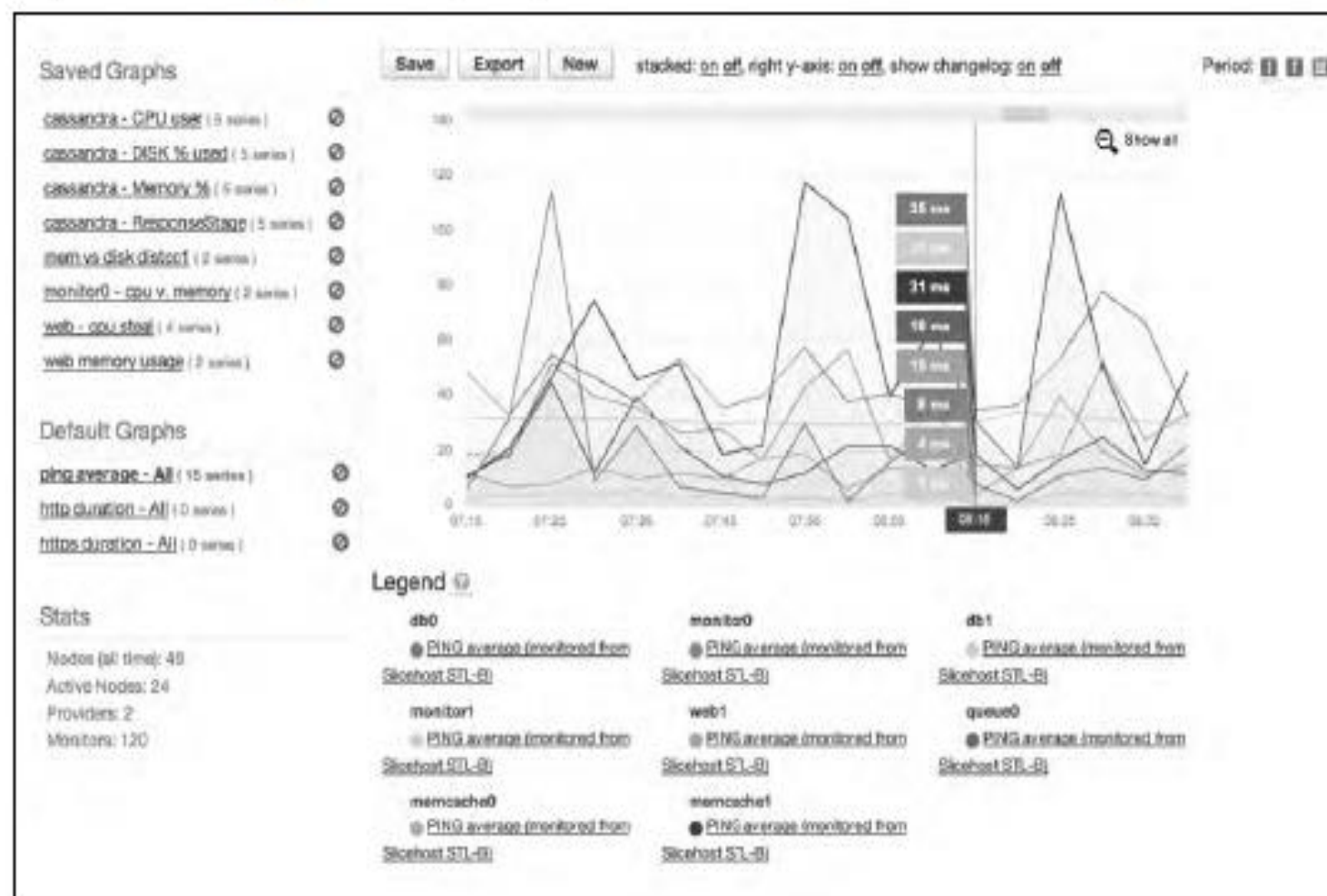
Product	URL	Description
Tapinsystems	http://www.tapinsystems.com/home	Provisioning and management service
Univa UD	http://univaud.com/index.php	Application and infrastructure management software for hybrid multi-clouds
VMware Hyperic	http://www.springsource.com/	Performance management for VMware deployed Java applications
Webmetrics	http://www.webmetrics.com/	Web performance management, load testing, and application monitor for cloud services
WebSitePulse	http://www.websitepulse.com/	Server, Web site, and application monitoring service
Whatsup Gold	http://www.whatsupgold.com/	Network monitoring and management software
Zenoss	http://www.zenoss.com/	IT operations monitoring
Zeus	http://www.zeus.com/	Web-based application traffic manager

- The **core management features** offered by most cloud management service products include the following:
 - Support of different cloud types
 - Creation and provisioning of different types of cloud resources, such as **machine instances, storage, or staged applications**
 - Performance reporting including **availability and uptime, response time, resource quota usage, and other characteristics**
 - The creation of dashboards that can be customized for a particular client's needs

- Automated deployment on IaaS systems represents one class of cloud management services.
- One of the more interesting and successful vendors in this area is
- 1. **RightScale** (<http://www.rightscale.com/>) whose software allows clients **to stage and manage applications on AWS (Amazon Web Service), Eucalyptus, Rackspace, and the Chef Multicloud framework or a combination of these cloud types.**
- RightScale creates **cloud-ready server templates and provides the automation and orchestration necessary to deploy them.**
- Eucalyptus and Rackspace both use Amazon EC2 and S3 services, although Eucalyptus is open source and portable.
- RightScale server templates and the Rightscript technology are highly configurable and can be run under batch control.
- The RightScale user interface also provides real-time measurements of individual server instances.

- **2. Cloudkick** is another infrastructure monitoring solution that is well regarded.
- Its service is noted for being agnostic and working with multiple vendor cloud platforms.
- The Cloudkick user interface is designed for **rapid deployment assessment, and its at-a-glance-monitoring Insight module is particularly easy to use.**
- Cloudkick's real-time server visualization tool, which is one of the more interesting presentation tools we've seen.
- Users have commented on Cloudkick's instant launching being difficult, and both Cloudkick and RightScale are known to be easy to use with Linux virtual servers and less so with Windows instances

Cloudkick's Insight module (https://www.cloudkick.com/site_media/images/graphs2.png) is powerful and particularly easy to use.



- All of the service models support monitoring solutions, most often through interaction with the service API.
- Tapping into a service API allows management software to perform command actions that a user would normally perform.
- Some of these APIs are themselves scriptable, while in some cases, scripting is supported in the management software.
- **One key differentiator in monitoring and management software is whether the service needs to install an agent or it performs its service without an agent.**
- The monitoring function normally can be performed through direct interaction with a cloud service or client using processes such as an HTTP GET or a network command like PING.
- **For management functions, an agent is helpful in that it can provide needed hooks to manipulate a cloud resource. Agents also, as a general rule, are useful in helping to solve problems associated with firewall NAT traversal.**

- ManageIQ and Service-now.com offer an integrated cloud stack that combines the ManageIQ Enterprise Virtualization Management Suite with Service-Now. com's ITSM SaaS service.
- The system has offers management, discovery, CMDB synchronization, and automated provisioning services. You can integrate these services into your Web applications using an open API that these companies offer.
- Distributed network applications often benefit from the deployment of a management appliance. Because cloud services tend to distribute applications across multiple sites, physical appliances need to be deployed in different locations—something that only cloud service providers can do.
- However, there has been a tendency to create virtual appliances, and those can be deployed as server instances wherever an application is deployed.
- Pareto Networks (<http://www.pareto-networks.com/>) has a cloud computing service that can monitor and manage distributed network services using a physical or virtual appliance. The system can be used to control and provision network services. Pareto Networks plans to add an API to this service.

Emerging Cloud Management Standards

- different cloud service providers use different technologies for creating and managing cloud resources.
- As the area matures, cloud providers are going to be under considerable pressure from large cloud users like the federal government to conform to standards and make their systems interoperable with one another.
- No entity is likely to want to make a major investment in a service that is a silo or from which data is difficult to stage or to extract.
- To this end, a number of large industry players such as VMware, IBM, Microsoft, Citrix, and HP have gotten together to create standards that can be used to promote cloud interoperability.
- Another effort just getting underway has been started by CA (the company formerly known as Computer Associates) in association with Carnegie Mellon called the Cloud Commons.
- This effort is aimed at creating an industry community and working group, and promoting a set of monitoring standards that were part of CA's cloud technology portfolio but are now open sourced.

- **DMTF cloud management standards**
- The Distributed Management Task Force is an industry organization that develops industry system management standards for platform interoperability.
- Its membership is a “who’s who” in computing, and since its founding in 1992, the group has been responsible for several industry standards, most notably the **Common Information Model (CIM)**.
- The DMTF organizes itself into a set of working groups that are tasked with specifying standards for different areas of technology.
- A recent standard called the Virtualization Management Initiative (VMAN) was developed to extend CIM to virtual computer system management.
- VMAN has resulted in the creation of the Open Virtualization Format (OVF), which describes a standard method for creating, packaging, and provisioning virtual appliances.

- OVF is essentially a container and a file format that is open and both hypervisor- and processor-architecture-agnostic.
- Since OVF was announced in 2009, vendors such as VirtualBox, AbiCloud, IBM, Red Hat, and VMWare have announced or introduced products that use OVF.
- It was, therefore, a natural extension of the work that DMTF does in virtualization to solve management issues in cloud computing.
- DMTF has created a working group called the Open Cloud Standards Incubator (OCSI) to help develop interoperability standards for managing interactions between and in public, private, and hybrid cloud systems.
- The group is focused on describing resource management and security protocols, packaging methods, and network management technologies.

- DMTF's cloud management efforts are really in their initial stages, but the group has broad industry support.
- Part of the group's task is to provide industry education, so you can find a number of white papers and technology briefs published on this site.
- It's an effort that's worth checking back with over time.
- Although the OCSI's work has not yet been joined by Amazon or Salesforce.com, a set of open standards that extend the use of industry standard protocols—such as the Common Information Model (CIM), the Open Virtualization Format (OVF), and WBEM—to the cloud are going to be hard for vendors to resist.

DMTF (<http://dmf.org/standards/cloud>) has a large and important effort underway for developing cloud interoperability management standards.

 **DISTRIBUTED MANAGEMENT TASK FORCE, INC.**
DMTF enables more effective management of IT systems worldwide.

[Workspace](#) [Members Area](#)

[Home](#) [About DMTF](#) [Standards & Technology](#) [News & Events](#) [Learning Center](#) [Contact Us](#) [Join Us](#)

[Home](#) > [Standards & Technology](#) > [Cloud](#)

Cloud Management



Cloud Management Standards

Workgroup formed to address management interoperability for cloud systems

Technologies like cloud computing and virtualization are rapidly being adopted by enterprise IT managers to better deliver services to their customers, lower IT costs and improve operational efficiency.

Using the recommendations developed by DMTF's Open Cloud Standards Initiative, the Cloud management workgroup (CMWG) is focused on standardizing interactions between cloud environments by developing specifications that deliver architectural overviews and implementation details to achieve interoperable cloud management between various providers as well as consumers and developers.

- [Read the Joint A Cloud Interoperability Overview Document](#)
- [Read the Interoperable Cloud White Paper](#)
- [Read the Architecture for Managed Clouds](#)
- [Read the Use Cases and Interactions for Managed Clouds](#)

Additional Resources:

- [Learning Center](#): Visit the DMTF Learning Center for a complete lot of overview documents, tutorials and other information.
- [Virtualization Management \(VMM\) standards](#): unleash the power of virtualization by defining broadly supported interoperability and portability standards to virtual computing environments through the Open Virtualization Format (OVF).
- [Cloud Standards Wiki](#): The Cloud Standards Wiki is a resource documenting the activities of the various Standards

CLOUD

Recent CLOUD News

- [Cloud Systems Group Develops Standard API](#)
- [DMTF Publishes "The Interoperable Cloud Standards"](#)
- [The Clouds: Enabling Open Service APIs](#)
- [Cloud APIs Set Open Source Treatment](#)
- [CloudPro Project - Consistent Solutions Framework](#)

Tutorials and Education

- [What is OVF Overview Document](#)
- [What is Cloud Interoperability Overview](#)
- [OVF Overview Document](#)
- [What is Overview Document](#)

- **Cloud Commons and SMI**
- CA Technologies, the company once known as Computer Associates, has taken some of its technologies in measuring distributed network performance metrics and repositioned its products as the following:
 - CA Cloud Insight, a cloud metrics measurement service
 - CA Cloud Compose, a deployment service
 - CA Cloud Optimize, a cloud optimization service
 - CA Cloud Orchestrate, a workflow control and policy based automation service
- Taken together, these products form the basis for CA's Cloud Connected Management Suite

- CA has lots of experience in this area through its Unicenter management suite and the products that were spawned from it.
- The company also has invested in cloud vendors such as 3Tera, Oblicore, and Cassatt to create their cloud services.
- CA acquired Nimsoft in March 2010. Nimsoft has a monitoring and management package called Nimsoft United Monitoring that creates a monitoring portal with customizable dashboards.
- The system can gather information from up to 100 types of data points and can work with both Google and Rackspace cloud deployments.
- Among the data points that can be monitored are resource usage and UPS status.

- At the heart of CA Cloud Insight is a method for measuring different cloud metrics that creates what CA calls a **Service Measurement Index or SMI**.
- The SMI measures things like SLA compliance, cost, and other values and rolls them up into a score.
- To help allow SMI to gain traction in the industry, CA has donated the core technology to the Software Engineering Institute at Carnegie Mellon as part of what is called the SMI Consortium.
- This same group is responsible for the Capability Maturity Model Integration (CMMI) process optimization technology and other efforts.
- The second CA initiative is the funding of an industry online community called the Cloud Commons

