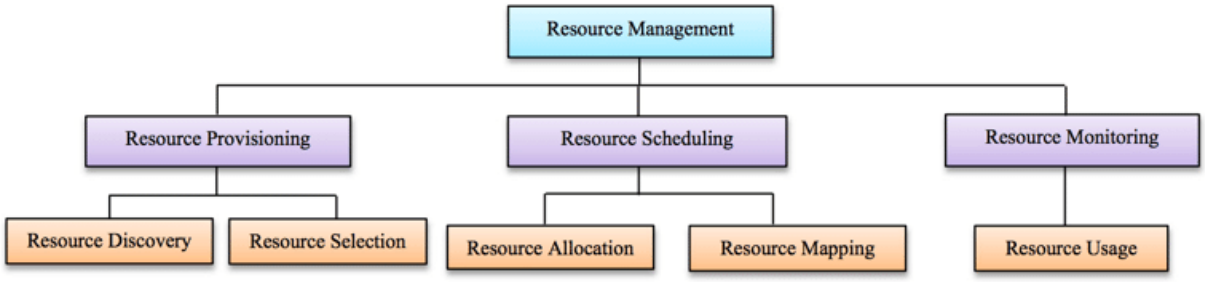
INTER-CLOUD RESOURCE MANAGEMENT

Resource management is an umbrella activity comprising different stages of resources : Resource Provisioning, Scheduling and Monitoring.

Resource provisioning is defined to be the stage to identify adequate resources for a given workload based on QoS requirements described by cloud consumers whereas resource scheduling is mapping and execution of cloud consumer workloads based on selected resources through resource provisioning. Resource monitoring is a key tool for controlling and managing hardware and software infrastructures. It provides information and Key Performance Indicators for both platforms and applications in cloud to be used for data collection to assist in the decision method of allocating the resources.



Firstly, cloud consumers submit requests for workload execution in the form of work-load details. Based on these details the broker (resource provisioner) finds the suitable resource(s) for a given workload and determines the feasibility of provision-ing of resources based on QoS requirements. Broker sends requests to the resource scheduler for scheduling after successful provisioning of resources. Other responsibilities of the broker include: release of extra resources to the resource pool, contains information of provisioned resources and monitor performance to add or remove resources. After resource provisioning, resource scheduling is done in second stage. All the provisioned resources are kept in the resource queue while other remaining resources are in the resource pool. Submitted workloads are processed in the workload queue. In this stage, the scheduling agent maps the provisioned resources to a given workload(s), executes the workload(s) and releases the resources back to the resource pool after successful completion of workload(s).

Important Activities:

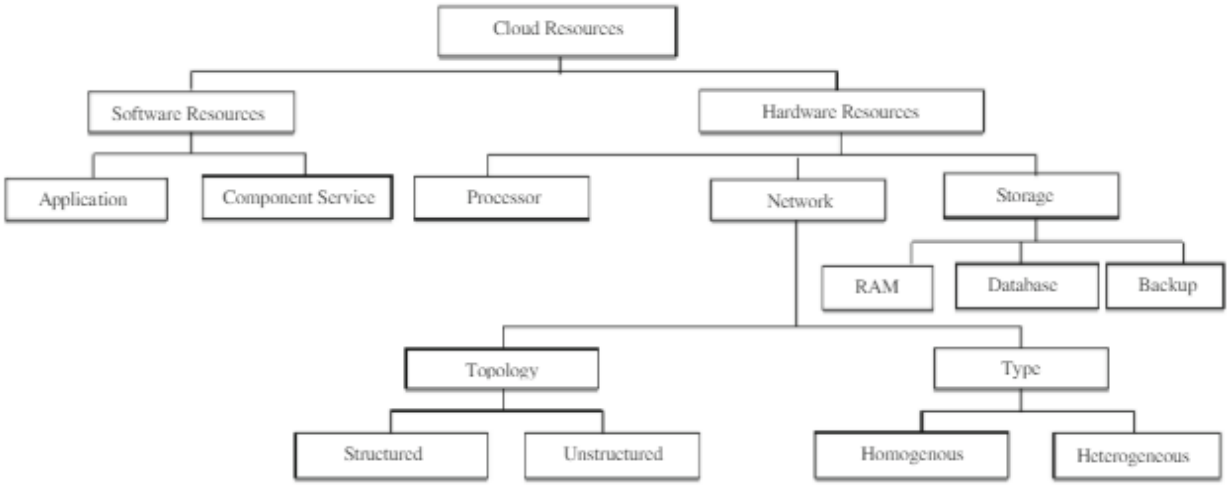
1. Resource discovery
2. Resource matching
3. Selection
4. Composition
5. Negotiation
6. Scheduling
7. Monitoring

Resource Provisioning and Platform Deployment

Resource Provisioning means the selection, deployment, and run-time management of software (e.g., database server management systems, load balancers) and hardware resources (e.g., CPU, storage, and network) for ensuring guaranteed performance for applications.

There are many resource provisioning techniques, each one having its own advantages and also some challenges. These resource provisioning techniques used must meet Quality of Service (QoS) parameters like availability, throughput, response time, security, reliability etc., and thereby avoiding Service Level Agreement (SLA) violations.

Service Level Agreement (SLA) : This is an initial agreement between the cloud users and cloud service providers which ensures Quality of Service (QoS) parameters like performance, availability, reliability, response time etc. The SLAs must commit sufficient resources such as CPU, memory, and bandwidth that the user can use for a preset period.
The fundamental element of resource management is the discovery process. It involves searching for the appropriate resource types available that match the application requirements. The process is managed by the cloud service provider. This process is being taken by the resource broker or user broker to discover available resources. Discovery consists of detailed descriptions of resources available.
The following are the various classifications of cloud resources:



Provisioning of Compute Resources (VMs)

Providers supply cloud services by signing SLAs(Service-level Agreement) with end users. 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.

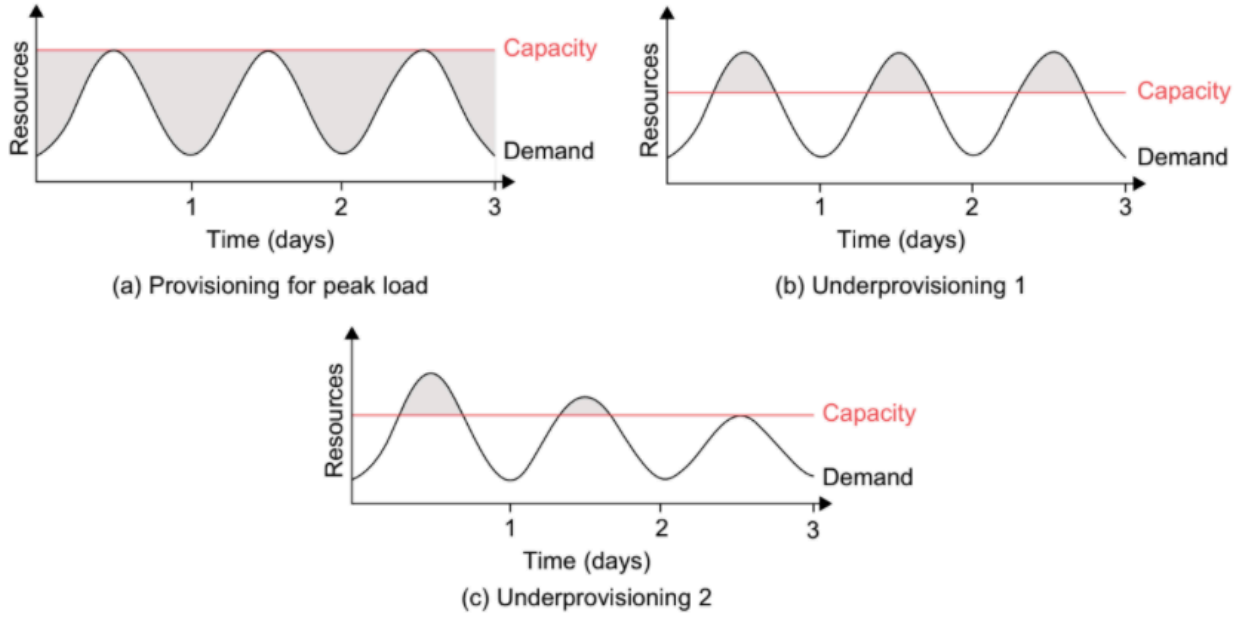
Efficient VM provisioning depends on the cloud architecture and management of cloud infrastructure. Resource provisioning schemes also demand fast discovery of services and data in cloud computing infrastructures. In a virtualized cluster of servers, this demands efficient installation of VMs, live VM migration, and fast recovery from failures. To deploy VMs, users treat them as physical hosts with customized operating systems for specific applications.

The provider should offer resource-economic services. Power-efficient schemes for caching, query processing, and thermal management are mandatory due to increasing energy waste by heat dissipation from data centers. Public or private clouds promise to streamline the

on-demand provisioning of software, hardware, and data as a service, achieving economies of scale in IT deployment and operation.

Resource Provisioning Methods

Figure below 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), underprovisioning (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 canceling the demand, resulting in reduced revenue for the provider.

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 processors for the user application, when the user was using one 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

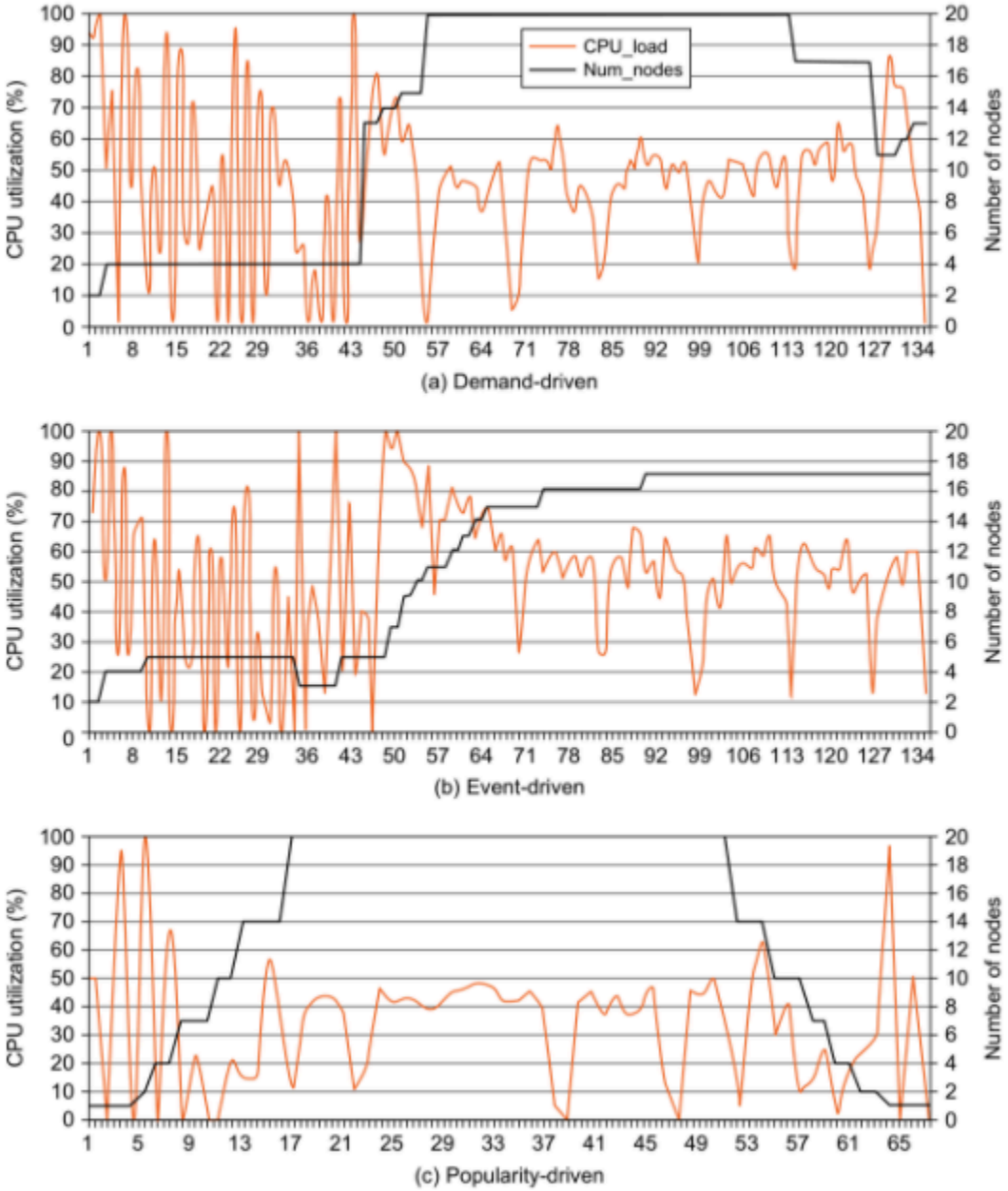
This scheme adds or removes machine instances based on a specific time event. The scheme works better for seasonal or predicted events such as Christmastime in the West and the Lunar

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.

Comparison of the three methods



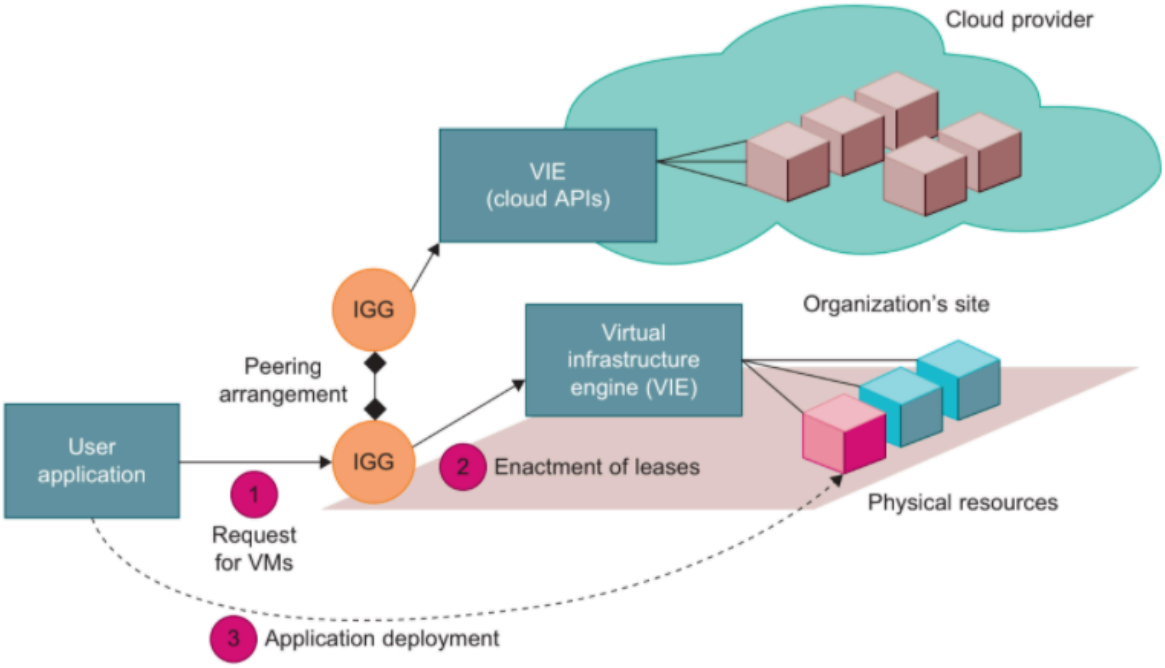
The x-axis in the figure is the time scale in milliseconds. In the beginning, heavy fluctuations of CPU load are encountered. All three methods have demanded a few VM instances initially. Gradually, the utilization rate becomes more stabilized with a maximum of 20 VMs (100 percent utilization) provided for demand-driven provisioning in Figure (a). However, the event-driven method reaches a stable peak of 17 VMs toward the end of the event and drops quickly in Figure (b). The popularity provisioning shown in Figure (c) leads to a similar fluctuation with peak VM utilization in the middle of the plot.

Dynamic Resource Deployment

The cloud uses VMs as building blocks to create an execution environment across multiple resource sites. Dynamic resource deployment can be implemented to achieve scalability in performance.

The InterGrid 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.

In Figure, 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, and (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.

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.

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. The biggest barriers to adopting flash memory in data centers have been price, capacity, and, to some extent, a lack of sophisticated query-processing techniques. However, this is about to change as the I/O bandwidth of solid-state drives becomes too impressive to ignore.

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 to access the objects stored in 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 semistructure database processing capability. For example, applications might want to process the information contained in a web page. Web pages are an example of semistructural 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. Traditional databases will meet the performance bottleneck while the system is expanded to a larger scale. However, some real applications do not need such strong consistency. The scale of such databases can be quite large. Typical cloud databases include BigTable from Google, SimpleDB from Amazon, and the SQL service from Microsoft Azure.

Resource Scheduling

Virtual Machine Creation and Management

By using independent service providers, the cloud applications can run different services at the same time. In addition to gateway applications, the cloud computing platform provides the extra capabilities of accessing backend services or underlying data.

Virtual Machine Manager : The VM manager is the link between the gateway and resources. The gateway doesn’t share physical resources directly, but relies on virtualization technology for

abstracting them. Hence, the actual resources it uses are VMs. The manager manage VMs deployed on a set of physical resources. The VM manager implementation is generic so that it can connect with different VIEs. Typically, VIEs can create and stop VMs on a physical cluster. Users submit VMs on physical machines using different kinds of hypervisors, which enables the running of several operating systems on the same host concurrently. The VMM also manages VM deployment on grids and IaaS providers. To deploy a VM, the manager needs to use its template.

Virtual Machine Templates : A VM template is analogous to a computer's configuration and contains a description for a VM with the following static information:

- The number of cores or processors to be assigned to the VM
- The amount of memory the VM requires
- The kernel used to boot the VM's operating system
- The disk image containing the VM's file system
- The price per hour of using a VM

Before starting an instance, the scheduler gives the network configuration and the host's address, it then allocates MAC and IP addresses for that instance. The template specifies the disk image field. To deploy several instances of the same VM template in parallel, each instance uses a temporary copy of the disk image. The IGG works with a repository of VM templates, called the VM template directory.

Distributed VM Management

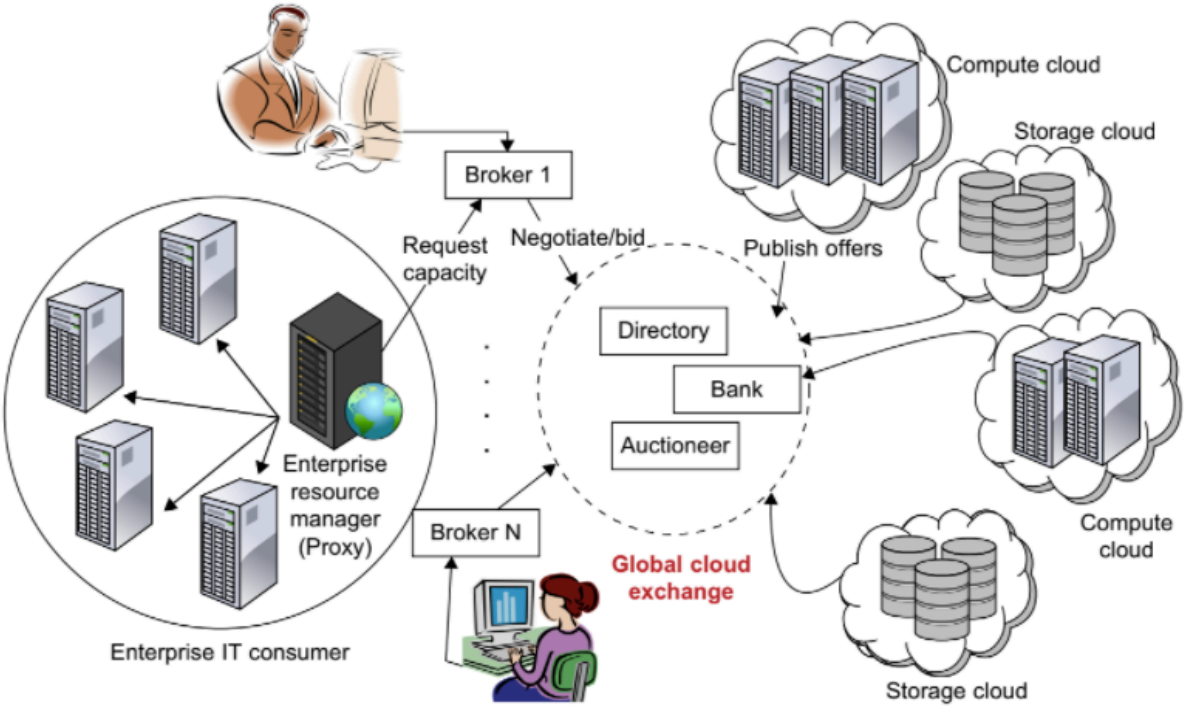
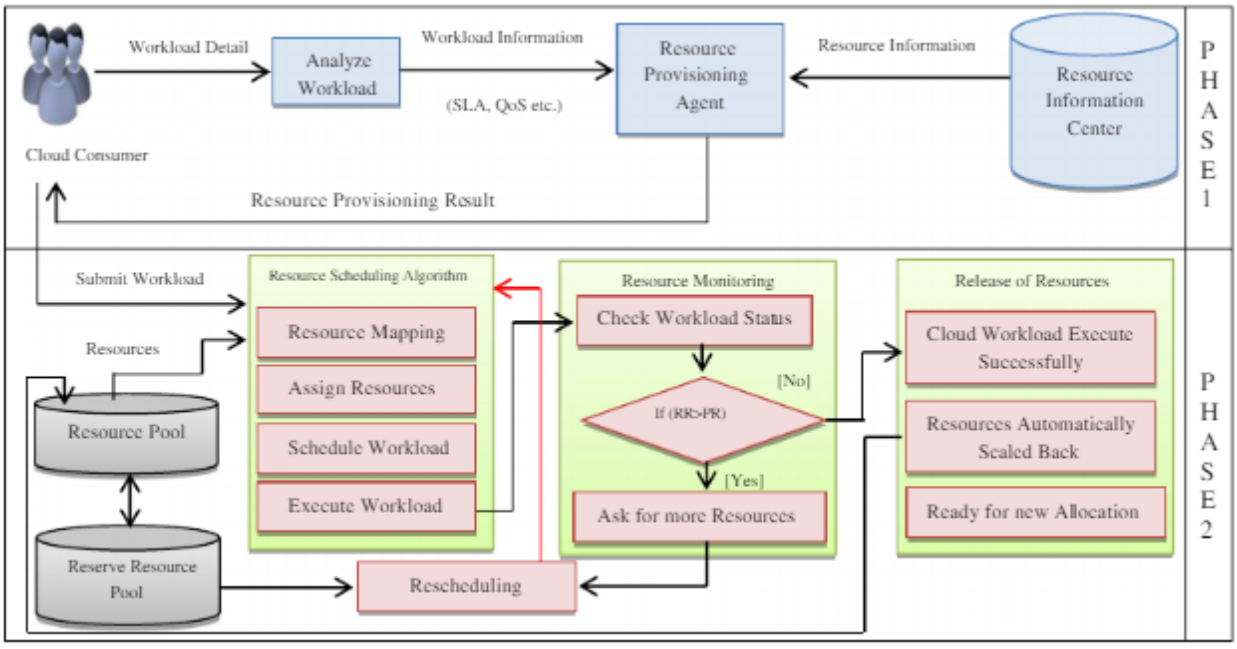


Figure illustrates the interactions between InterGrid's components. A distributed VM manager makes requests for VMs and queries their status. This manager requests VMs from the gateway on behalf of the user application. The manager obtains the list of requested VMs from the gateway. This list contains a tuple of public IP/private IP addresses for each VM with Secure Shell (SSH) tunnels. Users must specify which VM template they want to use and the number of VM instances needed, the deadline, the wall time, and the address for an alternative gateway. The local gateway tries to obtain resources from the underlying VIEs. When this is impossible, the local gateway starts a negotiation with any remote gateways to fulfill the request. When a gateway schedules the VMs, it sends the VM access information to the requester gateway. Finally, the manager configures the VM, sets up SSH tunnels, and executes the tasks on the VM. Under the peering policy, each gateway's scheduler uses conservative backfilling to schedule requests. When the scheduler can't start a request immediately using local resources, a redirection algorithm will be initiated.

Figure below gives a summarization of cloud resource management:

Note : Phase 2 comprises of both resource scheduling and resource monitoring



Global Exchange of Cloud Resources

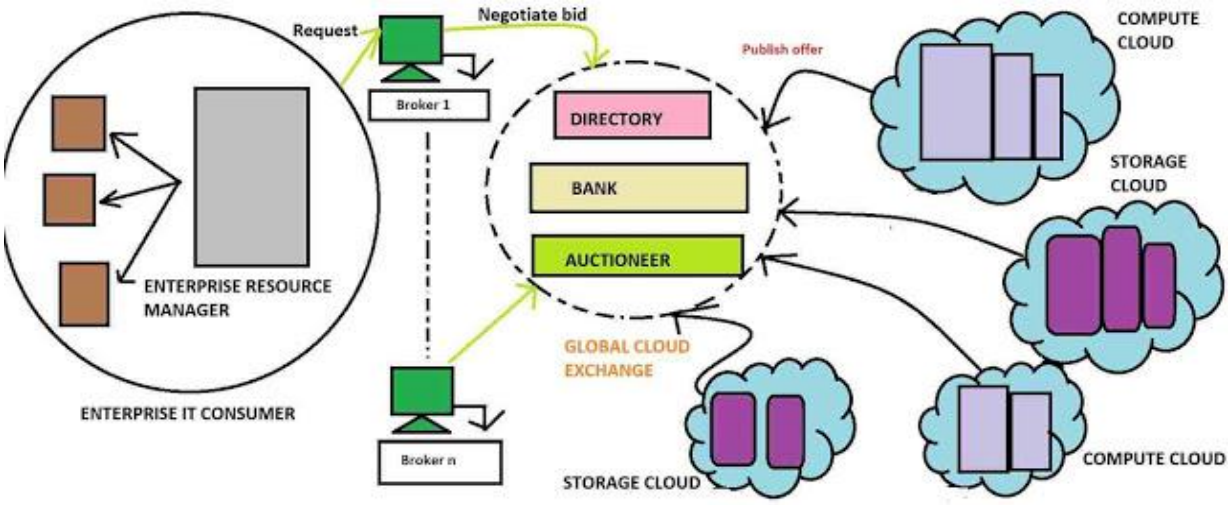
In cloud computing, large numbers of customers use cloud services from all over the world. To ensure reliability in the cloud server, the service provider established various data centers in different locations worldwide.

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.

If there are no automatic mechanisms, then different cloud service providers can't exchange their resources, and we were unable to solve various shortcomings in the **global exchange of cloud resources**, which are given below.

1- In cloud computing technology, we know that customers are distributed across different locations. It will be difficult for the customers in the cloud to determine the best locations in advance that they would like to have for hosting their services.

2- We know that by the quality of services (i.e., a response time of system, task performing capability, reliability, etc.), customers judge the different provider's services in the cloud. Because customers are from different geographical locations due to which the quality of services expectations will not meet.



In this architecture an enterprise resources manager was present which was an organization or any enterprise in which a central server system was located, and it was responsible for managing all the resources. The various supporting structures are connected to it and all these are mentioned by enterprise IT consumers; they are basically small organizations connected to the central system.

Then a different number of brokers will be present, and the numbers of brokers worked as an interface or medium between the consumer. Different service providers in the cloud have different kinds of storage cloud, different kinds of compute cloud, and these brokers are there to negotiate with different service providers to get the necessary resources requested by the enterprise resource managers.

And a central directory or global cloud exchange is mentioned in which a directory block, a bank, and an auctioneer were present. All the resources or services which users want to use will be mentioned on the directory block, and whatever the payment user pays for resources will be managed by the bank block. There will be an auctioneer whose primary role is to check all the organization's services and check the charges of resources or what basic service is provided to the user by the service provider.

Both the brokers and central directory system were linked together and provided services to the users.

Different cloud service providers (compute cloud, infrastructure cloud), attached with the cloud exchange system, as mentioned in the above image storage cloud (particular area for storing resources) were there. After that, there will be a requirement of a compute cloud in which data processing mechanism has happened.

And the publish offers mentioned in the above image with red color means that all the services/resources that the services provider will create are stored in the particular directory block because when the services were available only, the user will be able to access them. So different cloud service providers publish its resources into this cloud exchange.

How will the process start?

Suppose a user wants to access any particular resources, then with the broker help user can search his services from the directory, then different resources available, such as storage, computing can easily be used.

Different entities that were involved in the global exchange of the cloud resources

1- Market directory

The market directory makes the record of different services in which multiple services providers store their services.

2- Banking system

This banking system managed all the different transactions which were done by the users in cloud computing.

3- Brokers

They present the services of different providers to the users.

4- Price settings

In the price-setting mechanism, different users can determine the actual price of the available services present in the cloud.

5- Admission control mechanism

In admission control mechanisms, a priority was set for users' demands in the cloud so that different users can get the resources.

6- Resources management system

The resource management systems managed the resources available on the cloud; for example, in resource management, there will be scheduling, provisioning, resource usage, etc.

7- consumer utility function

We know that cloud services are based on the pay per use model, so the user needs to pay only for that utility or resources which they purchase from the cloud service provider.

8- Enterprise IT consumer

They are all small organizations that purchase the main organization's services and provide those services to their clients.