4.1 Inter Cloud Resource Management

Resource management is a process for the allocation of computing, storage, networking and subsequently energy resources to a set of applications, in a context that aims to collectively meet the performance goals of infrastructure providers, cloud users and applications. The cloud users prefer to concentrate on application performance while the conceptual framework offers a high-level view of the functional aspect of cloud resource management systems and all their interactions. Cloud resource management is a challenge due to the scale of modern data centers, the heterogeneity of resource types, the interdependence between such resources, the variability and unpredictability of loads, and the variety of objectives of the different players in the cloud ecosystem.

Whenever any service is deployed on cloud, it uses resources aggregated in a common resource pool which are collected from different federated physical servers. Sometimes, cloud service brokers may deploy cloud services on shared servers for their customers which lie on different cloud platforms. In that situation, the interconnection between different servers needs to be maintained. Sometimes, there may be a loss of control if any particular cloud server faces downtime which may generate huge business loss. Therefore, it's quite important to look at inter cloud resource management to address the limitations related to resource provisioning.

We have already seen the NIST architecture for cloud computing which has three layers namely infrastructure, platform and application.

These three layers are referred by three services like Infrastructure as a service, Platform as a service and Software as a service respectively. The Infrastructure as a service is the foundation layer which provides compute, storage and network services to other two layers like platform as a service and software as a service. Even as the three basic services are different in use, they are built on top of each other. In practical there are five layers required to run cloud applications. The functional layers of cloud computing services are shown in Fig. 4.1.1.

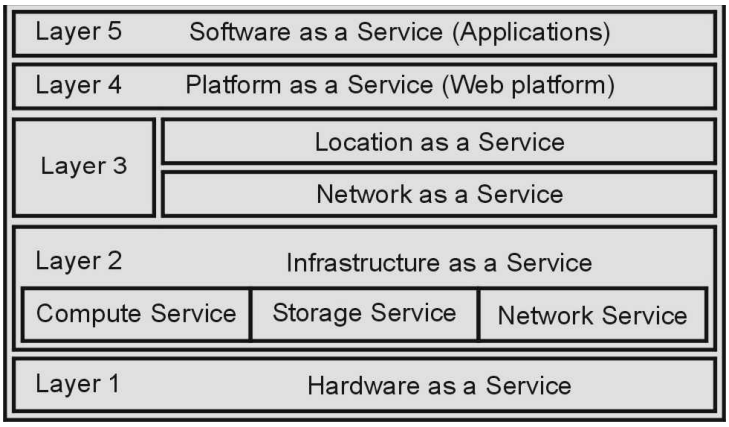

Fig. 4.1.1 Functional layers of Cloud computing

- The consequence is that one cannot directly launch SaaS applications on a cloud platform. The cloud platform for SaaS cannot be built unless there are compute, storage and network infrastructure are established.
- In above architecture, the lower three layers are more closely connected to physical specifications.
- The Hardware as a Service (HaaS) is the lowermost layer which provides various hardware resources to run cloud services.
- The next layer is Infrastructure as a Service that interconnects all hardware elements using computer, storage and network services.
- The next layer has two services namely Network as a Service (NaaS) to bind and provisioned cloud services over the network and Location as a Service (LaaS) to provide collocation service to control, and protect all physical hardware and network resources.
- The next layer is Platform as a Service for web application deployment and delivery while topmost layer is actually used for on demand application delivery.

In any cloud platform, the cloud infrastructure performance is the primary concern for every cloud service provider while quality of services, service delivery and security are the concerns for cloud users. Every SaaS application is subdivided into the different application areas for business applications like CRM is used for sales, promotion, and marketing services. CRM offered the first SaaS on the cloud successfully. The other tools may provide distributed collaboration, financial management or human resources management.

In inter cloud resource provisioning, developers have to consider how to design the system to meet critical requirements such as high throughput, HA, and fault tolerance. The infrastructure for operating cloud computing services may be either a physical server

or a virtual server. By using VMs, the platform can be flexible, i.e. running services are not associated with specific hardware platforms. This adds flexibility to cloud computing platforms. The software layer at the top of the platform is a layer for storing huge amounts of data.

Like in the cluster environment, there are some runtime support services accessible in the cloud computing environment. Cluster monitoring is used to obtain the running state of the cluster as a whole. The scheduler queues the tasks submitted to the entire cluster and assigns tasks to the processing nodes according to the availability of the node. The runtime support system helps to keep the cloud cluster working with high efficiency. Runtime support is the software needed for browser-initiated applications used by thousands of cloud customers. The SaaS model offers software solutions as a service, rather than requiring users to buy software. As a result, there is no initial investment in servers or software licenses on the customer side. On the provider side, the cost is rather low compared to the conventional hosting of user applications. Customer data is stored in a cloud that is either private or publicly hosted by PaaS and IaaS providers.

4.2 Resource Provisioning and Resource Provisioning Methods

The rise of cloud computing indicates major improvements in the design of software and hardware. Cloud architecture imposes further focus on the amount of VM instances or CPU cores. Parallelism is being used at the cluster node level. This section broadly focuses on the concept of resource provisioning and its methods.

4.2.1 Provisioning of Compute Resources

Cloud service providers offer cloud services by signing SLAs with end-users. The SLAs must commit appropriate resources, such as CPU, memory, and bandwidth that the user can use for a preset time. The lack of services and under provisioning of resources would contribute to violation of the SLAs and penalties. The over provisioning of resources can contribute to under-use of services and, as a consequence, to a decrease in revenue for the supplier. The design of an automated system to provision resources and services effectively is a difficult task. The difficulties arise from the unpredictability of consumer demand, heterogeneity of services, software and hardware failures, power management and disputes in SLAs signed between customers and service providers.

Cloud architecture and management of cloud infrastructure rely on effective VM provisioning. Resource provisioning schemes are also used for the rapid discovery of cloud computing services and data in cloud. The virtualized cluster of servers involve efficient VM deployment, live VM migration, and fast failure recovery. To deploy VMs,

users use virtual machines as a physical host with customized operating systems for different applications.

For example, Amazon's EC2 uses Xen as the Virtual Machine Monitor (VMM) which is also used in IBM's Blue Cloud. Some VM templates are also supplied on the EC2 platform. From templates, users can select different types of VMs. But no VM templates are provided by IBM 's Blue Cloud. Any form of VMs may generally be run on the top of Xen. In its Azure cloud platform, Microsoft also applied virtualization. A resource-economic services provider should deliver. The increase in energy waste by heat dissipation from data centers means that power-efficient caching, query processing and heat management schemes are necessary. Public or private clouds promise to streamline software, hardware and data as a service, provisioned in order to save on-demand IT deployment and achieving economies of scale in IT operations.

4.2.2 Provisioning of Storage Resources

As cloud storage systems also offer resources to customers, it is likely that data is stored in the clusters of the cloud provider. The data storage layer in layered architecture lies at the top of a physical or virtual server. The provisioning of storage resources in cloud is often associated with the terms like distributed file system, storage technologies and databases.

Several cloud computing providers have developed large scale data storage services to store a vast volume of data collected every day. A distributed file system is very essential for storing large data, as traditional file systems have failed to do that. For cloud computing, it is also important to construct databases such as large-scale systems based on data storage or distributed file systems. Some examples of distributed file system are Google's GFS that stores huge amount of data generated on web including images, text files, PDFs or spatial data for Google Earth. The Hadoop Distributed File System (HDFS) developed by Apache is another framework used for distributed data storage from the open source community. Hadoop is an open-source implementation of Google's cloud computing technology. The Windows Azure Cosmos File System also uses the distributed file system. Since the storage service or distributed file system can be accessed directly, similar to conventional databases, cloud computing does have a form of structure or semi-structured database processing capabilities. However, there are also other forms of data storage. In cloud computing, another type of data storage is (Key, Value) pair or object-based storage. Amazon DynamoDB uses (Key, Value) pair to store a data in a NOSQL database while Amazon S3 uses SOAP to navigate objects stored in the cloud.

In storage, numerous technologies are available like SCSI, SATA, SSDs, and Flash storages and so on. In future, hard disk drives with solid-state drives may be used as an enhancement in storage technologies. It would ensure reliable and high-performance data storage. The key obstacles to the adoption of flash memory in data centers have been price, capacity and, to some extent, lack of specialized query processing techniques. However, this is about to change as the I/O bandwidth of the solid-state drives is becoming too impressive to overlook.

Databases are very popular for many applications as they used as an underlying storage container. The size of such a database can be very high for the processing of huge quantities of data. The main aim is to store data in structured or semi-structured forms so that application developers can use it easily and construct their applications quickly. Traditional databases may meet the performance bottleneck while the system is being extended to a larger scale. However, some real applications do not need such a strong consistency. The size of these databases can be very growing. Typical cloud databases include Google's Big Table, Amazons Simple DB or DynamoDB and Azure SQL service from Microsoft Azure.

4.2.3 Provisioning in Dynamic Resource Deployment

The cloud computing utilizes virtual machines as basic building blocks to construct the execution environment across multiple resource sites. Resource provisioning in dynamic environment can be carried out to achieve scalability of performance. The Inter-Grid is a Java-implemented programming model that allows users to build cloud-based execution environments on top of all active grid resources. The peering structures established between gateways enable the resource allocation from multiple grids to establish the execution environment. The Intergrid Gateway (IGG) allocate resources from the local cluster to deploy applications in three stages, which include requesting virtual machines, authorizing leases and deploying virtual machines as demanded. At peak demand, this IGG interacts at another IGG that is capable of sharing resources from a cloud storage provider. The grid has pre-configured peering relationships with other grids that are controlled by the IGG. The system manages the use of Intergrid resources across several IGGs. The IGG is aware of peering parameters with other grids that selects appropriate grids that can provide the necessary resources, and responds to requests from other IGGs. The Request redirect policies decide which peering grid Intergrid wants to process the request and the rate at which that grid can perform the task. The IGG can even allocate resources from a cloud service provider. The cloud system provides a virtual environment that lets users to deploy their applications as like Intergrid, such

technologies use the tools of the distributed grid. The Intergrid assigns and manages a Distributed Virtual Environment (DVE). It is a cluster of available vms isolated from other virtual clusters. The DVE Manager component performs resource allocation and management on behalf of particular user applications. The central component of the IGG is the schedule for enforcing provisioning policies and peering with several other gateways. The communication system provides an asynchronous message-passing mechanism that is managed in parallel by a thread pool.

4.2.4 Methods of Resource Provisioning

There are three cases in the static cloud resource provisioning scheme, namely over-provisioning of resources at peak load, under provisioning of resources that results in losses for both the user and the providers because of wastage and shortage of resources below the allocated capacity and constant provisioning and Constant provision of resources with fixed capacity for declining user demand could result in to even worse waste of resources. In such cases, both the user and the provider may lose in the provisioning of resources with no elasticity.

- There are three resource-provisioning methods which are presented in the following sections.
- The demand-driven method offers static resources and has been used for many years in grid computing.
- The event-driven method is based on the expected time-dependent workload.
- The popularity-driven method is based on the monitoring of Internet traffic. We define these methods of resource provisioning as follows.

4.2.4.1 Demand-Driven Resource Provisioning

In demand driven resource provisioning, the resources are allocated as per demand by the users in dynamic environment. This method adds or eliminates computing instances depending on the current level of usage of allocated resources. The demand-driven method automatically allocates two the CPUs to the user application when the user uses one CPU more than 60 percent of the time for an extended period. In general, when a resource has met the threshold for a certain amount of time, the system increases the resource on the basis of demand. If the resource is utilized below the threshold for a certain amount of time, that resource could be reduced accordingly. This method is implemented by Amazon web services called as auto-scale feature that runs on its EC2

server. This method is very easy to implement. This approach does not work successfully if the workload changes abruptly.

4.2.4.2 Event-Driven Resource Provisioning

In event driven resource provisioning, the resources are allocated whenever an event generated by the users for at a specific time of interval in dynamic environment. This method adds or removes machine instances that are based on a specific time event. This approach works better for seasonal or predicted events when additional resources are required for shorter time of interval. During these events, the number of users increases before and decreases after the event period. Decreases over the course of the incident. This scheme estimates peak traffic before the event happens. This method results in a small loss of QoS if the occurrence is correctly predicted. Otherwise, its wasted resources are even larger due to events that do not follow a fixed pattern.

4.2.4.3 Popularity-Driven Resource Provisioning

In popularity driven resource provisioning, the resources are allocated based on popularity of certain applications and their demands. In this method, the internet checks for popularity of certain applications and produces instances by popularity demand. In this method, the Internet seeks the popularity and creates instances by popularity demand of certain applications. The scheme expects increased in traffic with popularity. Again, if the predicted popularity is correct, the scheme has a minimum loss of QoS. If traffic does not happen as expected, resources may get wasted.

4.3 Global Exchange of Cloud Resources

To serve a large number of users worldwide, the IaaS cloud providers have set up datacenters in various geographical locations to provide redundancy and ensure reliability in the event of site failure. However, Amazon is currently asking its cloud customers (i.e. SaaS providers) to give preference to where they want their application services to be hosted. Amazon does not have seamless/automatic frameworks for scaling hosted services across many geographically dispersed data centers. There are many weaknesses in this approach. First, cloud customers cannot find the best place for their services in advance because they do not know the origin of their services' consumers. Secondly, SaaS providers may not be able to meet QoS requirements from multiple geographical locations of their service consumers. It involves the development of structures that help complex applications across multiple domains to efficiently federate

cloud data centers to meet cloud customers' QoS targets. Moreover, not a single provider of cloud infrastructure will be able to set up its data centers, anywhere around the world. This will make it difficult to meet the QoS standards for all its customers by cloud applications service (SaaS) providers. They also want to take advantage of the resources of multiple providers that can best serve their unique needs in cloud infrastructure. In companies with global businesses and applications such as Internet services, media hosting and Web 2.0 applications, this form of requirement often arises. This includes the federation of providers of cloud infrastructure to offer services to multiple cloud providers. To accomplish it, Intercloud architecture has been proposed to enable brokerage and the sharing of cloud resources for applications across multiple clouds in order to scale applications. The generalized Intercloud architecture is shown in Fig. 4.3.1.

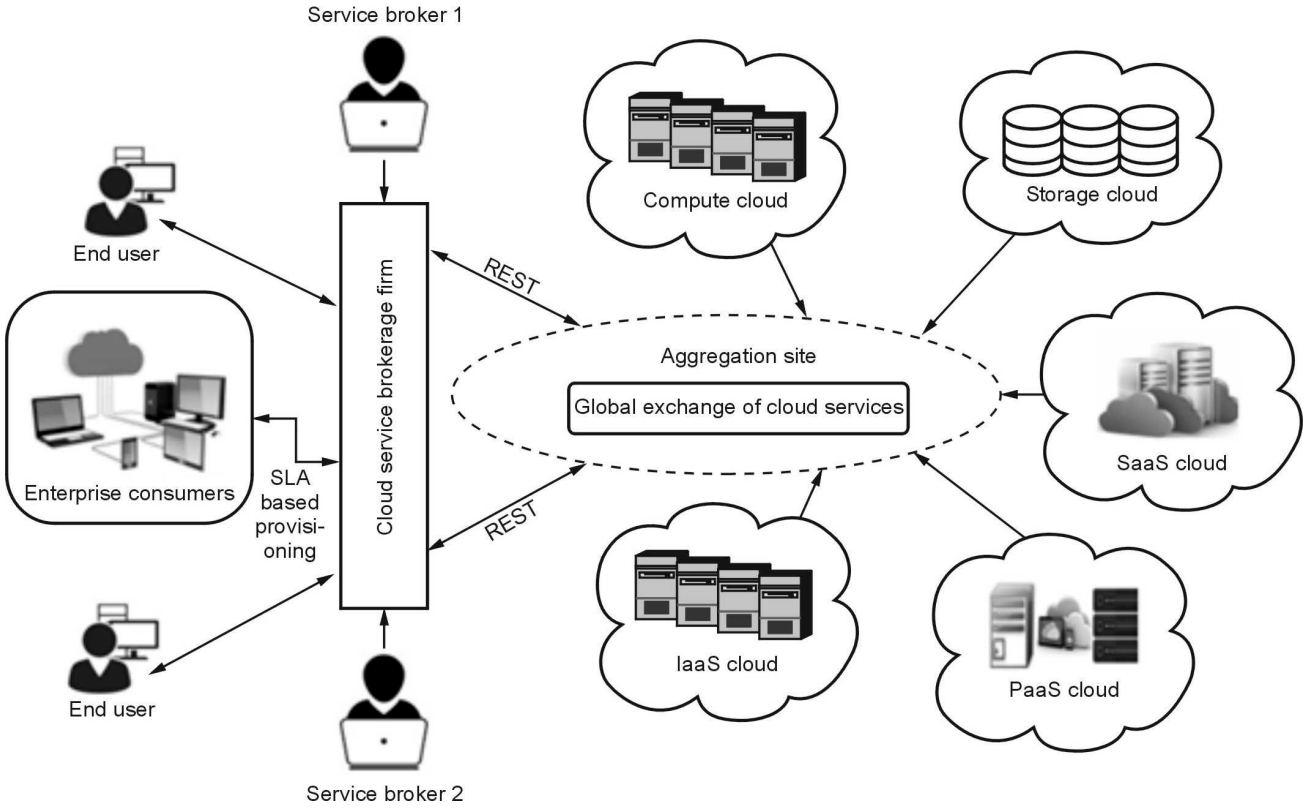

Fig. 4.3.1 Intercloud Architecture

The cloud providers can expand or redimension their provision capacity in a competitive and dynamic manner by leasing the computation and storage resources of other cloud service providers with the use of Intercloud architectural principles. It helps operators, such as Salesforce.com to host services based in an SLA contract that is agreed, to operate in a market-driven resource leased federation. It offers reliable, on demand, affordable and QoS-aware services using virtualization technology and ensures high QoS quality and reduces cost of operation. They must be able to employ market-based utility

models as the assumption to offer heterogeneous user applications to virtualize software services and federated hardware infrastructures.

The intercloud architecture consolidates the distributed storage and computing capabilities of clouds in a single resource-leasing abstraction. They comprise client brokerage and coordination services which support the utility based useful cloud federation: scheduling of applications, allocation of resources and workload migration. The system will facilitate the integration of cross-domain capability for on demand, adaptable, energized and reliable infrastructure access based on virtualization technology. The Cloud Exchange (CEX) is used to enhance and analyze the infrastructure demands of application brokers against the available supply. It acts as a marketing authority to bring service producers and consumers together to encourage cloud service trading on the basis of competitive economic models such as commodity prices and auctions. The SLA (Service Level Agreement) specifies the service details to be provided in accordance with agreed metrics, incentives and penalties for meeting and breaching expectations. The accessibility of a bank system within the market ensures that SLAs between participants are transacted in a safe and reliable environment.

4.4 Security Overview

The cloud computing is made as an on-demand service through the network to provision resources, applications and information. It includes a very high computational power and storage capacity. Nowadays most small and medium size companies (SMEs) move to the cloud because of their advantages such as lower infrastructure, no maintenance costs, model payoff, scalability, load balancing, independent venue, on-demand access, quicker deployment and flexibility, etc.

Although cloud computing has many benefits in most of the aspects, but security issues in cloud platforms led many companies to hesitate to migrate their essential resources to the cloud. In this new environment, companies and individuals often worry about how security, privacy, trust, confidentiality and integrity of compliance can be maintained. However, the companies that jump to the cloud computing can be even more worrying about the implications of placing critical applications and data in the cloud. The migration of critical applications and sensitive data to public and shared across multiple cloud environments is a major concern for companies that move beyond the network perimeter defense of their own data center. To resolve these concerns, a cloud software provider needs to ensure that customers continue to maintain the same security and privacy controls on their services and applications, provide customers with evidence that their company and consumers are secure, and can fulfill their service-level agreements,