Chapter 2

Background

2.1 Cloud Computing Paradigm

The term *Cloud* first appeared in the early 1990s, referring mainly to large ATM networks. Cloud Computing is not a completely new concept; it has connection to the established grid Computing, utility computing, cluster computing, and distributed systems [8]. In this computing technology dynamically scalable and often virtualized resources are provided as a service over the internet. It comprises of hardware and software resources available on the internet and handled by third-party services [9, 10]. These services generally provide access to upgraded software applications with high-end networks of server computers.

Buyya et al. [3] defined cloud computing as a parallel and distributed systems comprising a collection of inter-connected and virtualized servers that are provisioned dynamically and delivered as one or more unified computing resources established between the cloud provider and consumers based on service-level agreements (SLA) .

As per Vaquero et al. [11] clouds are a large pool of virtualized resources consisting of hardware, software and development platforms including services which can be easily accessible and usable by the service consumer. These resources can dynamically be configured to a variable load allowing for an optimal resource utilization. The resources are typically used by a pay-as-per-use pattern in which the service Provider assures the performance of the system which are defined in the SLAs.

McKinsey et al. [12] specified clouds are the hardware based services providing computer, network, and storage capability where the infrastructure management is abstracted from the cloud users, users receive hardware costs as variable operational expenses, and the hardware capacity is extremely elastic.

National Institute of Standards and Technology (NIST) [13] defines cloud computing as a pay-as-per-use model for providing networks, servers, storage, applications and services from a shared pool of configurable computing resources [14, 15] that can be rapidly provisioned and released with minimal management effort.

Slabeva and Wozniak [16] described cloud computing as follows :

- The hardware, storage, system software and applications are provided as X-as-a-Service mode. Cloud Computing is based on pay as per usage business models and these services are provided by an service provider.
- Virtualization and dynamic scalability on demand.
- Utility computing and Software as a Service(SaaS) may be provided in an integrated manner, although utility computing might be needed separately .
- The services offered by the providers are utilized either through web browser or through a defined API.

The common characteristics among the notable ones are

- pay-per-use (no ongoing commitment, utility prices)
- elastic capacity and the illusion of infinite resources
- self-service interface and

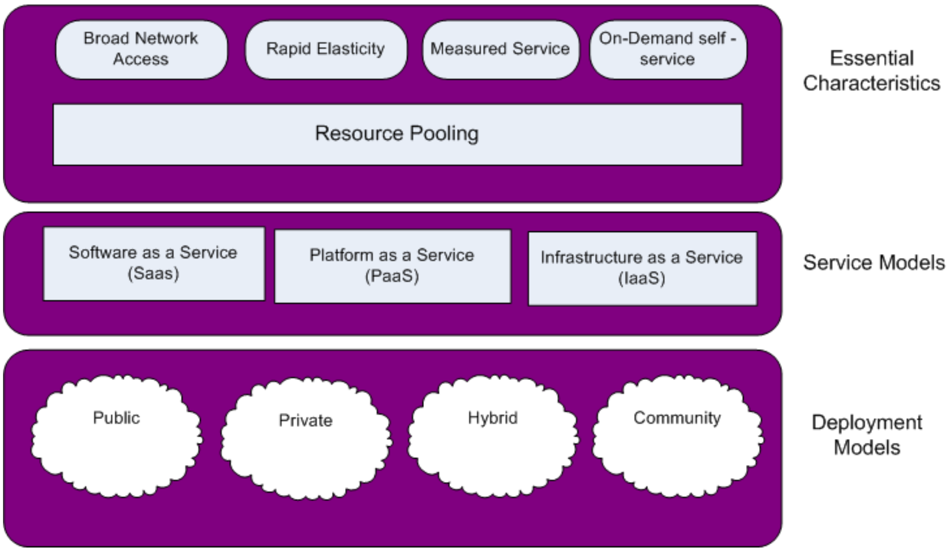


Figure 2.1: System model of NIST Working Definition of Cloud Computing.

- resources that are abstracted or virtualized

Three things are required to work with cloud computing. They are : thin clients (or clients with a thick-thin switch), grid computing which links heterogeneous computers to make a large infrastructure, and utility computing (paying for what users consume)[17, 18].

2.2 Cloud Computing Architecture

Figure 2.1 shows the architecture of cloud computing as defined by National Institute of Standards and Technology(NIST) [13]. This cloud model is composed of five essential characteristics, three basic components, three layers, three cloud service models and four cloud deployment models.

2.2.1 Essential Characteristics

- *On-demand-self-service.* The consumer can provision the services provided by the service providers without the human interaction.

- *Broad Network Access.* The capabilities are available over the network and accessed through standard mechanism which is used by heterogeneous thin or thick client platforms such as mobile phones, laptops and PDAs.

- *Resource Pooling.* Depending on the consumer demand various physical and virtual resources are dynamically assigned to the consumers from the providerś computing resource pool. The provider serves multiple consumers using a multi-tenant model.

- *Rapid Elasticity.* The services are quickly and elastically provided.

- *Measured Service.* The resource such as storage, processing, bandwidth and active user accounts are automatically controlled and optimized by providing a metering capability.

The ways in which the above mentioned characteristics are demonstrated in an enterprise context vary according to the deployment model employed.

2.2.2 Components of Cloud Computing

Cloud computing involves the following three basic components:

- Clients: The internet data or other services are accessed by the clients through desktop computers, laptops, tablet computers or other mobile devices.

- Datacenter: A set of servers hosting a specific collection of applications constitute a datacenter. Through the technology of virtualization a large number of virtual servers can run on one physical server. Depending on the size, speed of the physical server and nature of the applications running in the virtual server the number of virtual servers are initiated.

- Distributed servers: The cloud providers host their physical servers in different physical locations which do not affect the interaction of cloud end-users. Due to any reason if datacenter is inaccessible, the service will still be available through another distributed server. In accession, if the cloud needs more hardware capacity to support its peak workload, it is not necessary to attach more servers onto the primary datacenter but another group of distributed servers can be automatically embedded to the cloud.

2.2.3 Layers of Cloud Model

The technical, business, and policy issues of cloud computing play out across three layers:

- The Infrastructure layer : It includes the hardware, networks and operating systems responsible for handling the underlying resources such as data storage, computation and network bandwidth. The ability of cloud infrastructure layer is to virtualize the resources and to establish a connection between physical resources and the services that consume them. There may be several virtual machines that can be instantiated on a physical server, or there may be more than one physical server functioning one virtual machine. Service providers can dynamically add, remove or modify hardware resources without reconfiguring the services that depend on them.

- The Platform layer : This layer serves two purposes. It renders a set of usual services that are dealt by applications, such as messaging, business rules engines and databases. It also isolates application creators via a set of higher level Application Programming Interfaces (APIs) from the complexity of the inherent infrastructure.

- The Application layer : It provides the applications that dispel the cloud and are usable to web browsers or endeavors on a pay-as-you-go basis and responsible in the management of the databases and software, updates and removal, including installation. The business logic for the application is a function in the application layer of the cloud datacenter.

2.2.4 Cloud Service Models

Cloud computing services can be broadly divided into three service models : Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a service (IaaS) [13]. These service models describe the degree of service and control the offers of cloud service provider(CSP) and the degree of freedom of customers. Figure 2.2 gives a graphical representation of the different service models, and their components while Figure 2.3 depicts the offerings at various levels of the cloud model. Figure 2.4 shows the offered services at each levels of cloud models with the providers.

- Traditional IT: The organization has to manage and locate all the computing infrastructure by itself. The organization buys its own servers and manages the complete infrastructure from networking to application levels.

- Infrastructure as a Service (IaaS): The IaaS customers essentially hires data center space, servers, software, network equipment, etc. as a fully outsourced service. Generally, the service is billed on a monthly basis like a utility bills to customers. The customer is charged only for the resources they consume. Developers may create a specific operating system instance with native applications running.

- Platform as a Service (PaaS): PaaS customers are concerned solely on web based development and mostly do not care whatever operating system is used. PaaS services permit customers to focus only on innovations instead of thinking for the complex infrastructure. Customers can redirect a substantial percentage of their budgets to create applications that provide real business value instead of worrying about all the infrastructure issues.

- Software as a Service (SaaS): In the Software as a Service (SaaS) model, the CSP offers all infrastructure as a service, including the application. The applications are accessible from various client devices through a

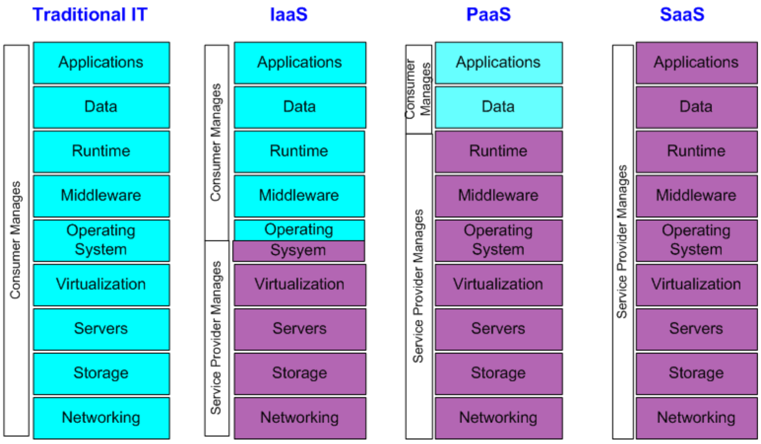


Figure 2.2: Behavior of Cloud Computing Service Models

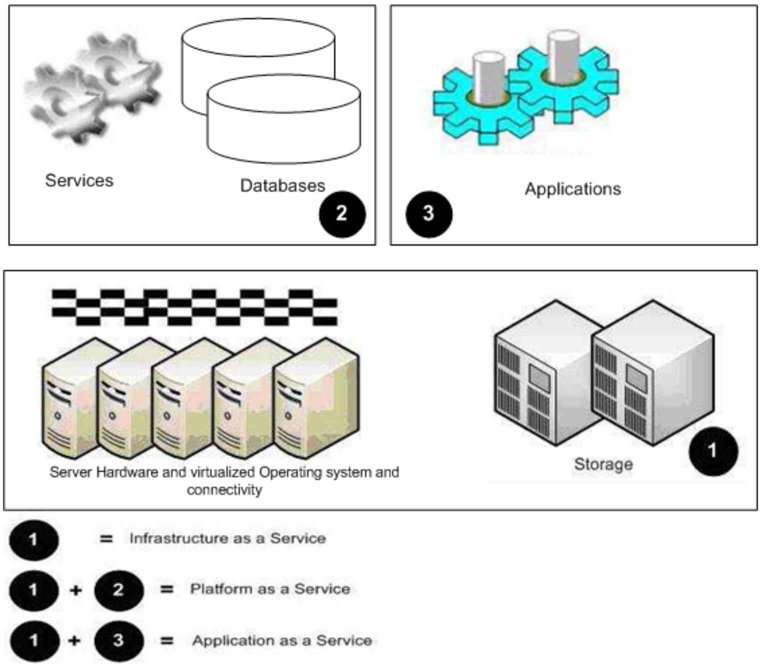
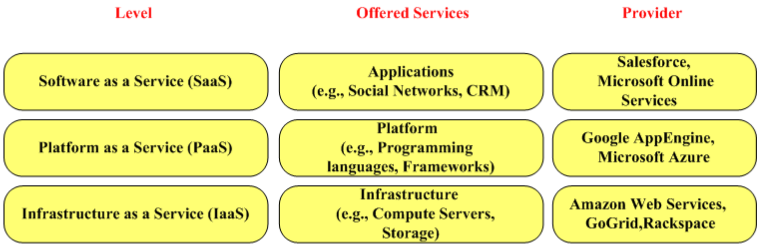

Services
Databases
2
3
Applications
Server Hardware and virtualized Operating system and connectivity
Storage
1
1 = Infrastructure as a Service
1 + 2 = Platform as a Service
1 + 3 = Application as a Service

Figure 2.3: Cloud Computing Service Models



Level	Offered Services	Provider
Software as a Service (SaaS)	Applications (e.g., Social Networks, CRM)	Salesforce, Microsoft Online Services
Platform as a Service (PaaS)	Platform (e.g., Programming languages, Frameworks)	Google AppEngine, Microsoft Azure
Infrastructure as a Service (IaaS)	Infrastructure (e.g., Compute Servers, Storage)	Amazon Web Services, GoGrid,Rackspace

Figure 2.4: The Cloud Computing Stack

thin client interface such as a web browser. The consumer does not manage or control the underlying cloud infrastructure, but may be able to set limited user-specific application configuration settings.

2.2.5 Deployment Models

- Private Cloud: In Private Cloud model, an internal datacenter is assembled within an organizations and infrastructure is devoted to a particular organization and not dealt with other organizations. Private clouds are of two types: on-premise private clouds [19] and externally hosted private clouds. When a product or service of a company demands to be maintained beneath rigid check and a usual demand to fiddle with the infrastructure, the idealistic condition is to use private cloud. A private cloud existing among a shared or public cloud is known as a virtual private cloud (VPC). Amazon web services established Amazon Virtual Private Cloud, which permits the Amazon Elastic Compute Cloud service to be linked to bequest infrastructure throughout an Internet Protocol Security virtual private network connection [20]. Figure 2.5(a) shows that the private cloud is only used by one consumer, resources are not shared with other consumers.

- Public Cloud: Cloud infrastructure is hosted at vendors premises. Public cloud permits the user to access the cloud facilities employing web browser as the interim layer. The computing infrastructure and the services of cloud supplier are usable through internet and are dealt be-

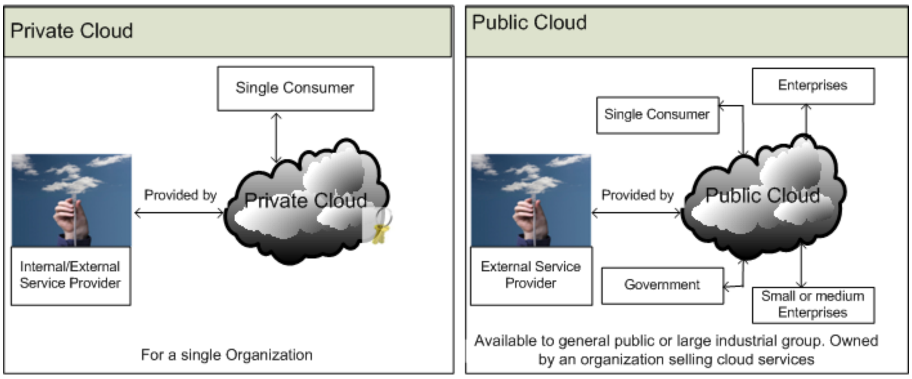

Private Cloud
Single Consumer
Provided by
Private Cloud
Internal/External Service Provider
For a single Organization
Public Cloud
Enterprises
Single Consumer
Provided by
Public Cloud
External Service Provider
Government
Small or medium Enterprises
Available to general public or large industrial group. Owned by an organization selling cloud services

Figure 2.5: (a) Private and (b) public cloud deployment models.

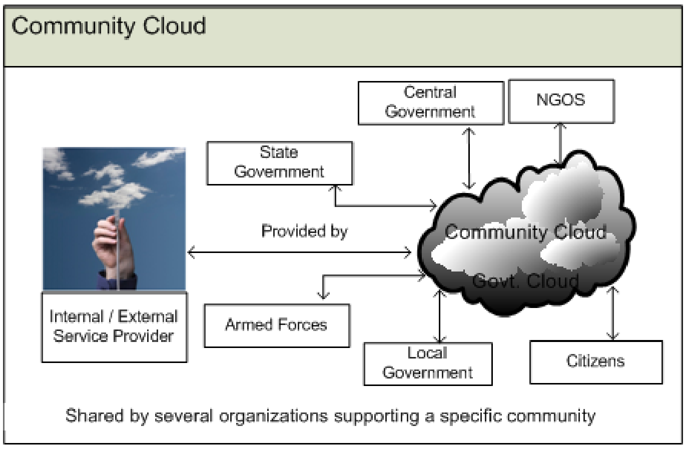


Figure 2.6: Community cloud deployment models.

tween any organizations. It cuts the capital disbursements as the cost is allotted and dealt throughout a large group of businesses and individuals [21]. Figure 2.5(b) shows that in the public cloud, resources are shared with multiple consumers, which may operate globally, and have different security demands.

- Community Cloud: The cloud infrastructure is shared by several organizations and supports a specific community that has shared concerns (e.g., mission, security requirements, policy, and compliance considerations). It may be managed by the organizations or a third party and may exist on premise or off premise. Figure 2.6 shows an example of a

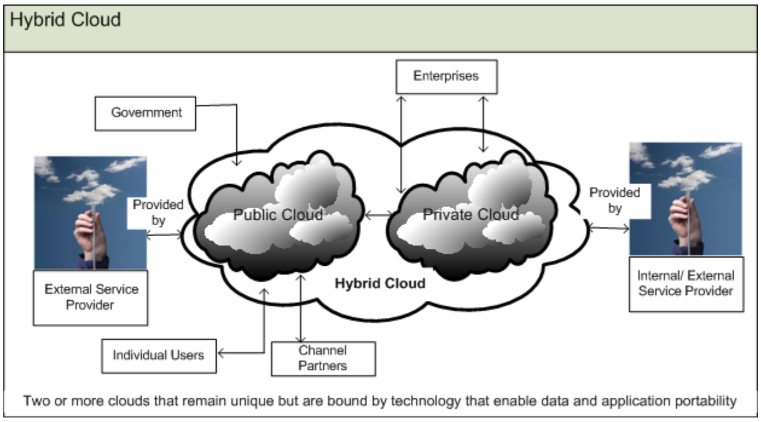
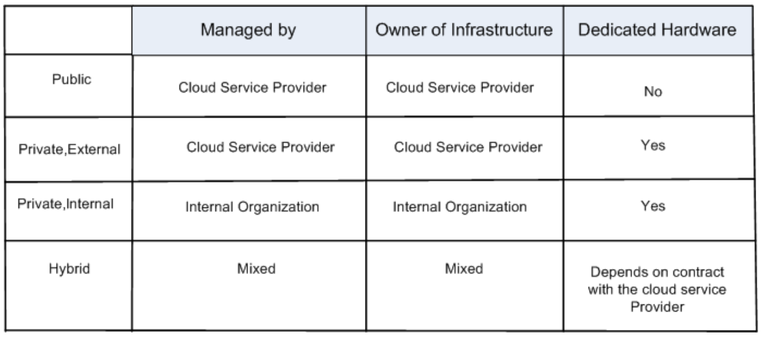

Hybrid Cloud
Government
Enterprises
Provided by
Public Cloud
Private Cloud
Provided by
External Service Provider
Hybrid Cloud
Internal/ External Service Provider
Individual Users
Channel Partners
Two or more clouds that remain unique but are bound by technology that enable data and application portability

Figure 2.7: Hybrid cloud deployment models.



	Managed by	Owner of Infrastructure	Dedicated Hardware
Public	Cloud Service Provider	Cloud Service Provider	No
Private,External	Cloud Service Provider	Cloud Service Provider	Yes
Private,Internal	Internal Organization	Internal Organization	Yes
Hybrid	Mixed	Mixed	Depends on contract with the cloud service Provider

Figure 2.8: Cloud Deployment Models

community cloud, which is used for a government community. Google offers a government cloud named Google Gov Cloud.

- Hybrid Cloud: A hybrid cloud is a private cloud linked to one or more external cloud services, centrally managed, provisioned as a single unit, and circumscribed by a secure network. It renders virtual IT solutions through a usage of both public and private clouds collectively [22]. It permits several parties to get information over the internet and provides more secure control of the applications and data. It also has an open architecture that allows interfaces with other management systems. The approach of temporarily renting capacity to handle spikes in load is known as cloud-bursting [23]. Figure 2.7 gives a graphical representation of a hybrid cloud, consisting of a public cloud and private cloud. The private cloud is only used by the consumer, while the public cloud is shared with other consumers. The private cloud and public cloud may be offered by different service providers. Figure 2.8 describes the characteristics of different deployment models.

2.3 Cloud Resource Virtualization

Earlier days in a data center all the individual systems have installed with standard operating systems and they rely on conventional operating system techniques to ensure resource sharing, performance isolation and application protection. System administration, accounting, security, and resource management were very challenging for the service providers and in the same way application development and performance optimization were equally challenging for the users.

Virtualization is a basic principle of cloud computing is the alternative to the above problem. It simplifies some of the resource management tasks. A VM running under a virtual machine monitor (VMM) can de saved and migrated to another node to balance the load. At the same time, virtualization allows users to operate in environments they are familiar with rather than forcing them to work in single environment.

2.3.1 Virtualization

Through virtualization multiple independent operating system runs on a single physical server. It includes virtual networks, such as virtual private network (VPN) and VMs, such as VMware and Xen. Virtual network affirm customers with a custom-made network environment to access cloud resources and VMs render virtualized IT-infrastructures on-demand [24]. Virtualization hides the technical detail through encapsulation [25].

The virtualization technique is depicted in Figure 2.9. The Virtual Ma-

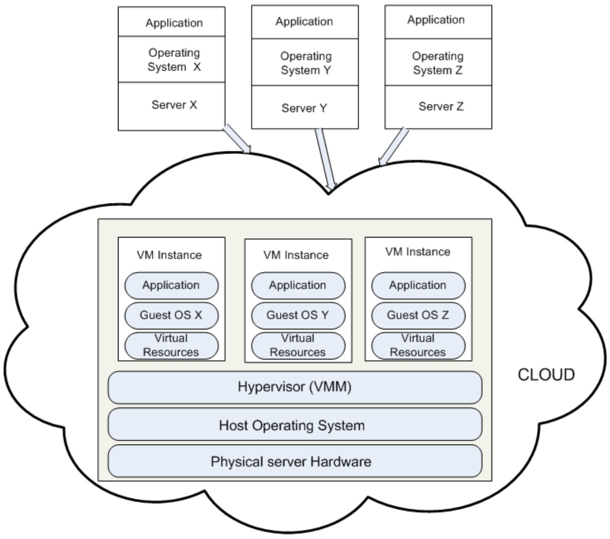


Figure 2.9: Virtualization technology

chine Monitor(VMM), also known as the hypervisor is a virtualization platform that allows multiple operating systems to run on a host computer at the same time. It establishes the abstraction layer that encapsulates and isolates each VM. The hypervisor runs on the physical machine and maps physical resources such as processing power, storage, memory and network to VMs. The Operating System (OS) running inside a VM can therefore use the virtual resources mapped to the confining VM. Therefore, the physical resources of a machine can be shared among the multiple VMs. Two types of hypervisors are available:

- *Type 1 hypervisors* is software that are installed directly on the given hardware platform or onto bare metal. A guest operating system runs above the type 1 hypervisor. Type 1 hypervisors actually offer better performance than the type 2 systems. Recent examples are Xen [26], VMware's ESX Server, and Sun's Hypervisor (released in 2005).

- *Type 2 hypervisors* are software applications that run on top of a host operating system.

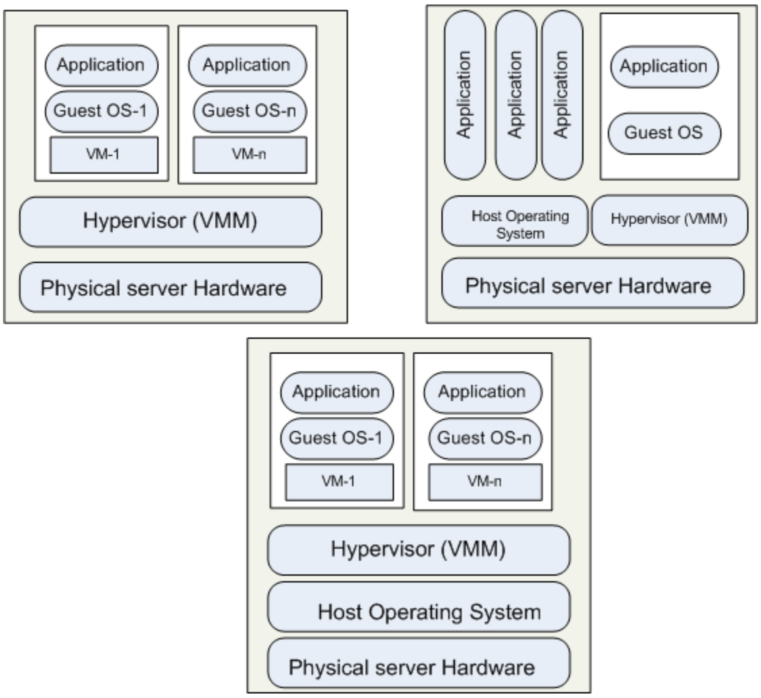


Figure 2.10: Three different types of system virtual machine

2.3.2 Virtual Machine

A VM which appears to be a whole computer works in an isolated environment. It has the access capability to a portion of the computer resources. Each VM appears to be running on the bare hardware are supported by a single physical system. Figure 2.10 shows three different types of system VMs such as traditional, hybrid and hosted virtual machines.

- Traditional VMs; the VMM supports multiple VMs and runs directly on the hardware.

- Hybrid VM; the VMM shares the hardware with a host operating system and supports multiple VMs.

- Hosted VM; the VMM runs under a host operating system.

2.3.3 Virtual Machine Monitors

A Virtual Machine Monitor (VMM) also called hypervisor is a software that securely partition the resources of a physical server into more than one VMs. A guest operating system runs under the control of a VMM rather than directly on the hardware. The VMM runs in kernel mode while a guest OS runs in user mode. VMM allow several operating systems to run concurrently on a single hardware platform. It controls how the guest operating system uses the hardware resources such that the events occurring in one VM should not affect any other VM running under the same VMM. At the same time, the VMM allows:

- Multiple services to share the same platform.
- The movement of a server from one platform to another, the so-called live migration.

2.3.4 Cloud Provisioning Approach

Cloud provisioning is the process of deployment and management of application services on cloud infrastructures. This process is complex as it computes the best software and hardware configuration to assure the QoS objectives of application services with the maximization of system efficacy and usage. These are the following unpredictable situations which obstructs the smooth provisioning and delivery of application services :

- *Estimation Error*: IT managers or SaaS owners faces the difficulties to understand the demands because of the complexness of cloud based IT resources and applications. It is enormously hard for IT managers to discover the decent combination of IT resources that can accomodate the current and predict the future workload.
- *Highly dynamic workload*: A popular application may be accessed by huge end-users resulting a extremely variable load spikes. The feature of workload [27] spikes could vary over application types . This induces

serious problems while estimating the workload behavior and related resource requirements.

- *Uncertain Behaviour*: The availability, load, and throughput of Cloud-based IT resources and network links can vary in an unpredictable way in cloud data centers. The provisioning problem is shown to be computationally intractable (i.e., NP-hard) [28].

The provisioning technique consists of three key steps [27]:

- *VM provisioning*: The process instantiates one or more VMs that fits the specific hardware and software requirements of an application. Several cloud providers provides a class of general use VMs having generic resource configurations and software. For instance Amazon EC2 provides 11 categories of VMs, having different alternatives of processors, memory, and I/O performance.

- *Resource Provisioning*: This process maps and schedules VMs onto physical cloud servers . It is completely hidden from the application providers.

- *Application Provisioning*: It is the deployment of particular applications such as ERP, BLAST experiments, and web servers within VMs and mapping of users requests to application instances.

2.3.5 Virtual Machines Provisioning & Manageability

Figure 2.11 shows the system architecture consisting of a dispatcher, global manager and local manager [7]. The local managers as a part of the virtual machine monitor (VMM) reside inside the physical servers. Besides monitoring the thermal state of each node the local manager checks the utilization of the current node's resources. The following are the situations in which the local manager chooses VMs to migrate to some other node to maintain the QoS of the system and better efficiency of the system:

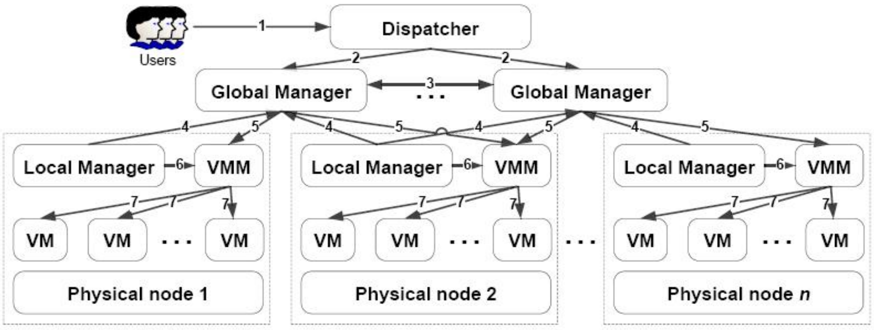


Figure 2.11: VM provisioning opeation

- If the resource utilization of a node is very low, instead of unnecessarily utilizing the resources of the node, the VMs should be reallocated to some other node and the idle node should be turned off.
- If the utilization of some resources approaches towards 100% and the system may not able to provide the requisite performance as per the SLA, some VMs need to be migrated to other nodes.
- If two VMs are in different nodes and there is a frequent need of network communication, migrate and put both VMs in a single node.
- If the temperature of a node exceeds a threshold value, some VMs are migrated to some other node to cool the node naturally.

As the VMs have been selected by local managers for migration, the information is communicated to global manager about the resource utilization. The local manager issues commands for VM resizing and turning on/off idle physical nodes as shown in Figure 2.12 [29]. Each of the global manager is tied up with several physical nodes and processes data obtained from their local managers.

The VM provisioning operation is consists of the following steps:

- *Fresh request for VM provisioning.* Cloud users sends requests for VM provisioning.

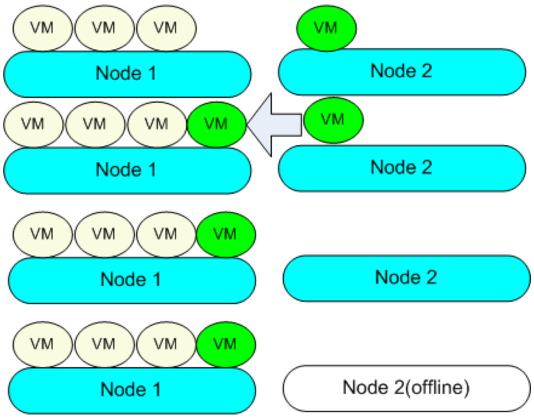


Figure 2.12: VM management dynamic shutdown technique

- *Dispatching requests for VM provisioning.* Once the requests are received from the users the dispatcher distributes the requests among the global managers.

- *Intercommunication between global managers.* The exchange of information takes place among the global managers about the resource utilization and allocation of VMs.

- *VMs chosen to migrate and Utilization of resources.* The local manager pass the information to the global manager about the utilization of resources and the VMs chosen to migrate.

- *Migration commands.* The command for the VM migration are issued by the global manager for optimizing the recent allocation.

- *VM Resizing commands and power states adjustment.* The host nodes are monitored by the local managers residing on the VMM to issue commands for resizing the VM and to change the power states of nodes.

Figure 2.13 depicts the life cycle of VMs and its potential states of functions [16]. The cycle starts by a client request, conveying the need for allocating a new server for a particular service. This request is being served by looking at the servers resource pool, coping with the resources with the necessities, and

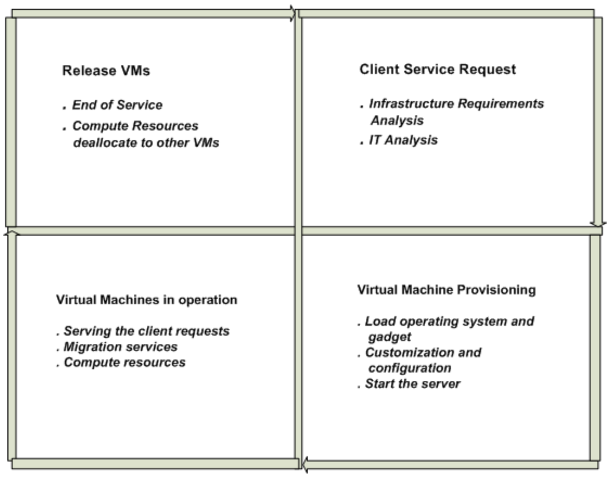

Release VMs
. End of Service
. Compute Resources deallocate to other VMs
Client Service Request
. Infrastructure Requirements Analysis
. IT Analysis
Virtual Machines in operation
. Serving the client requests
. Migration services
. Compute resources
Virtual Machine Provisioning
. Load operating system and gadget
. Customization and configuration
. Start the server

Figure 2.13: Life cycle of Virtual Machines

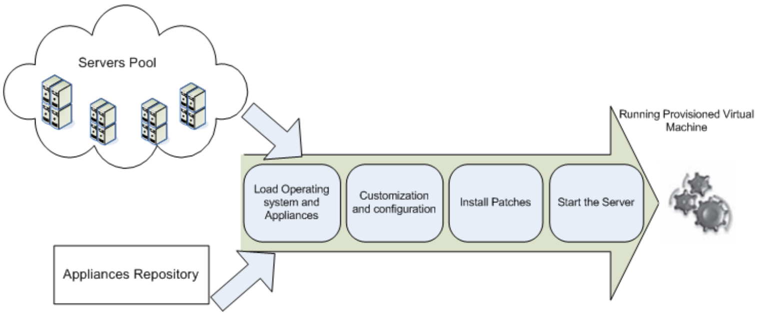

Servers Pool
Running Provisioned Virtual Machine
Load Operating system and Appliances
Customization and configuration
Install Patches
Start the Server
Appliances Repository

Figure 2.14: Virtual Machine Provisioning Process

starting the provision of the needed VM. Once it is provisioned and initiated, it is ready to furnish the required service according to an SLA, and after a time period it may be released by freeing the resources.

The virtual machine provisioning process is illustrated in Figure 2.14. The normal and common steps of provisioning a virtual machine or virtual server are as follows:

- Select a server from the available server pool (physical servers with adequate capacity) along with the appropriate operating system template needed for provisioning the virtual machine.

- Load the appropriate software such as the operating system which is selected in the first step, middleware, device drivers and the needed applications for the required services.

- Configure the IP address, gateway of the associated network and storage resources.

- Finally, the virtual server is ready to start with its newly loaded softwares.

2.4 Performance Management on Cloud

Many cloud services and cloud-oriented applications are not efficient enough for the following reasons: (1) lack of information sharing, (2) assumption of heterogeneous environments and (3) unpredictability of the environments.

2.4.1 Information Sharing

In the cloud system the providers and the users usually do not share workload information and the detailed resource states among themselves as they are often from different parties and having their own interests. Usually, providers offer a different types of resource containers without precise specifications. Similarly the specification of network I/O is specified only as being low, medium, high, and very high, without any numerical metric. The network topology and organization of the machines is not exposed. So, the users are making resource requests only through guess, which may go wrong. Similarly, the providers are unaware of users applications and workloads that are being executed in the VMs. That is the reason, providers may place VMs in an arbitrary fashion, which results inefficient resource utilization and degradation of performance. For example, if the application needs large number of inter-machine communication, the machines should be placed in the same rack to reduce the inter rack communication. On the other hand, if the application provides contents over the internet, spreading the resources over many racks will be worthy to avoid bottleneck. But, the providers are unaware of such details to make better resource allocation decisions.

2.4.2 Heterogeneous Environment

Typically, cloud service providers begin their cloud computing business with several near-homogeneous computing nodes. Over time, the cloud provider will replace some of the old computing nodes with newer nodes featuring the latest technologies. Gradually, the capability and performance of all machines in the cloud will become more heterogeneous [30]. In a heterogeneous environment, some jobs executes faster on a specific node than others. For instance, if a task is capable to exploit accelerators such as graphic processing units, then instead of running the task on regular nodes it should run the task on nodes equipped with GPUs.

2.4.3 Unpredictability

The cloud environment is highly variable and unpredictable. The players engaged in the system are frequently different parties having their business needs. The providers want to maximize the revenue with minimum investment by compacting their computing resources. In other words, cloud providers want to maximize the resource utilization. From the users point of view, it may frustrate the cloud user as placing of too many VMs on a single physical machine may cause the performance degradation. Thus, to maintain service quality, the providers may eject existing VMs or reject resource requests, which could make the environment even more unpredictable. Therefore, the performance might not be as expected and furthermore could fluctuate. This variance in performance may induce a trouble if the customer is unable to anticipate these fluctuations, order of magnitude, and duration.

2.5 Related Work

There has been an adequately discussion on performance management on a cluster. For example, in [31], the authors studied the optimization of Apache Web server. The performance management on cloud differs from the cluster mainly in the following two aspects:

- As we discussed above the VMs are heterogeneous in nature. But, the web applications deployed on a cluster must be homogeneous due to the unified software platform on each node.

- It is much more complex than the mechanism of resource sharing on a cluster as in a cloud via VM the web applications share computing resources.

RAD Lab of Berkeley focuses on the use of statistical machine learning as a diagnostic and predictive tool that would allow dynamic scaling, automatic reaction to performance and correctness problems, and generally automatic management of many aspects of these systems. In [32], Kernel Canonical Correlation Analysis (KCCA) is used to predict the execution time of MapReduce jobs in a data-intensive system running on a cloud and the work has been published in [33]. In [34], a queueing based model was proposed by the authors to predict the performance of the service exposed by the cloud. They have taken the assumption that the cloud exposes only a single service.

In [35], the authors have proposed a bin packing algorithm which enables application live placement dynamically with consideration of energy efficiency in a cloud platform. It also supports applications scheduling and live migration of VMs to minimize the number of running machines, so as to save energy. In addition, an over-provision approach is presented to deal with the varying resource demands of applications. It can be a reference model for dynamical scheduling. There are still many other researches on the performance management on cloud. All the researches can be roughly divided into two kinds: by prediction and by real-time status. As our analysis, they all have pros and cons. We design a model based on the real-time status of web applications which more exactly captures the structure of cloud and measures the performance of the VMs to decide when there is a need for VM migration.

The resource management of data centers in cloud computing environments has been discussed in [36, 37]. Nathuji and Schwan [38] have proposed an architectural model for resource management for virtualized data centers where this is handled by local and global policies [39]. On the local level, the system uses the guest operating systems power management strategies. The VMs migration to reallocate VMs are handled by global policies. Beloglazov and Buyya [7] had worked on global VM allocation policies considering strict SLA. Srikantaiah et al. and Kusic et al. [40, 41] have studied the problem of requests scheduling for multi-tiered web-applications in virtualized heterogeneous systems for resource management, while meeting performance requirements. The authors have suggested a heuristic multidimensional bin packing algorithm for workload integration to manage the optimization over multiple resources. The proposed approach is workload type and application dependent. Song et al. [42] have proposed resource allocation to applications according to their priorities in multi-application virtualized cluster. The approach requires machine-learning to obtain utility functions for the applications and defined application priorities.

Quiroz et al. [27] suggests a technique for dynamic VM provisioning in IaaS data centers based on clustering. In their a work, they need to determine the the types as well as the number of virtualized application instances. Zhu and Agrawal [43] proposed a dynamic methodology established on control theory for VM provisioning by considering the user's budget.