CHAPTER 1

INTRODUCTION TO CLOUD COMPUTING PARADIGM

Delivery of Information and Communication Technologies (ICT) services as a utility has recently received significant consideration through Cloud computing. Cloud computing technologies will provide scalable on demand pay per use services to customers through distributed data centers. Still this paradigm is its infant stage and many challenging issues have to be addressed. Accordingly, in this chapter, firstly cloud computing paradigm basics are introduced and various cloud computing services are discussed subsequently. Next this chapter presents High Level Architecture for providing cloud services along with the security issues at different layers. This chapter elaborates functionalities performed at different layers and also the security requirements at different layers. This chapter also explores the different research issues in cloud computing environment.

1.1 Cloud Computing Paradigm

In Cloud Computing [1] scalable resources are provisioned dynamically as a service over internet in order to assure lots of monetary benefits to be scattered among its adopters. Different layers are outlined based on the kind of services provided by the Cloud. Moving from bottom to top, bottom layer contains basic hardware resources like Memory, Storage Servers. Hence it is denoted as Infrastructure-as-a-Service (IaaS). The distinguished examples of IaaS are Amazon easy Storage Service (S3) and Amazon Elastic Compute Cloud (EC2). The layer above IaaS is Platform-as-a-Service (PaaS) which mainly supports deployment and dynamic scaling of Python and Java based applications. One such an example of PaaS is Google App Engine. On top of PaaS, a layer that offers its customers with the capability to use their applications referred to as Software-as-a-Service (SaaS). SaaS supports accessing user’s applications through a browser without the knowledge of Hardware or Software to be installed. This approach has been proven to be a universally accepted and trusted service. Internet and Browser are the two components required to access these Cloud services. IaaS applications access requires more internet bandwidth where as web browser may be sufficient with reasonable internet bandwidth is sufficient to access SaaS and PaaS applications.
The word “cloud” was a euphemism for everything that was beyond the data center or out on the network. There are several definitions of a cloud assumed by different categories of cloud users. It is mostly described as software as a service, where users can access a software application online, as in Salesforce.com, Google Apps and Zoho. It is also described as in the form of infrastructure as a service, where a user does not own infrastructure but and rents it over time on a server and accesses through a site such as Amazon Elastic Compute Cloud (EC2).

Another form of a Cloud is Platform as a service in which certain tools are made available to build software that runs in the host cloud. Basically a cloud is built over a number of the data centers, which reflects the Web’s context for loosely coupled systems (i.e. two systems don’t know about each other), and provides the ability to have virtualized remote servers through standard Web services to have large computing power. Cloud paradigm also serves as a business model apart from technology. Through the business model, the cloud makes a new form of computing widely available at lower prices that would have been considered impossible.

1.2 Cloud Computing Services

The basic services of Cloud have been considered as the following.

**Platform as a Service (PaaS):** PaaS is an deployment and development platform for applications provided as a service to developers over the Web. Third party renders develop and deploy software or applications to the end users through internet and servers. The cost and complexity of development and deployment of applications can be reduced to a great extent by developers by using this service. Thus the developers can reduce the cost of buying and reduce the complexity of managing the required Infrastructure. It provides all of the services required to build and deliver the web services to support the complete life cycle of web applications entirely from the Internet. This platform consists of infrastructure software, databases, middleware, and development tools.

**Infrastructure as a Service (IaaS):** is a delivery model associated with Hardware and Software as a service. Hardware such as Storage, server and network along with supporting software such as operating system, virtualization technology and file system. It is an
evolution of traditional hosting to allow users to provide resources on demand and without require any long term commitment. Different from PaaS services, the IaaS provider does very little management of data other than to keep the data center operational. Deployment and managing of the software services must be done by the end users just as the way they would in their own data center.

**Software as a service (SaaS):** SaaS allows access to programs to large number of users all the way through browser. For a user, this can save some cost on software and servers. For Service provider’s, they only need to maintain one program, this can also save space and cost. Naturally, a SaaS provider gives access to applications to multiple clients and users over web by hosting and managing the given application in their or leased datacenters. SaaS providers also runs their applications on platforms and infrastructure provided by other cloud providers.

### 1.3 Overview of High-level Cloud Architecture

Figure 1.1 provides a High Level Architecture [2][97] of a cloud computing architecture along with the security issues at different layers. This section elaborates functionalities performed at different layers and also the security requirements at different layers.

#### 1.3.1 User Layer

Different types of users like customers, application programmers, and administrators interact with cloud software through the user layer. This layer consists of two sub layers.

**Application Sub Layer:** The cloud applications are visible through the user layer to the end-users of the cloud. Normally, the applications are accessed through web portals by the users and from time to time required to pay amount to use them. The overhead of software maintenance is done by this sub layer and also the ongoing operation and support costs. Furthermore, it moves the computing tasks from the user terminal to servers in the Datacenters where the cloud applications are deployed. This in turn minimizes the requirements on the hardware required from the user’s point of view, and permits them to gain higher performance. This approach supports efficient processing of CPU-intensive and
memory-intensive workloads of the users without any huge capital investments in their local machines.

Thus this sub layer even simplifies the work with respect to code upgradation and testing, while protecting their intellectual property from the service providers point of view. Developers can add new features through patches easily without distributing the end users as the cloud application is deployed at the provider’s computing infrastructure rather than at the user machines. Configuration and testing of a application is less complicated using this sub layer functionality, since the deployment environment becomes restricted in the provider’s datacenter. in terms of profits margin to the provider, continuous flow of revenue is supplied through this sub layer, which brings more profits over a period of time.

In spite of all the benefits and advantages of this sub layer functionality, a number of deployment issues hinder with its broad acceptance. More specifically, the availability and
security of the cloud applications are the two major challenges that has direct impact on Service Level Agreements (SLAs). In addition, managing the availability is a monarchy that providers and users of SaaS has to deal with due to possible network outage and system failures. Additionally, the migration of the user data and integration of legacy [8] applications to the cloud is another challenge that is also time-consuming for the adoption of SaaS.

**Programming Environment Sub Layer:** The users of this layer are cloud application developers responsible for development and deployment of applications on to the cloud. The cloud service providers support development environment with necessary set of defined APIs. Developers interact with the environments through the available APIs, which accelerates the deployment and support the with scalability support.

Google’s App Engine [3] is one of the example systems in this category, that supports python runtime environment and APIs for to interact with Google’s cloud runtime environment. Through this approach implementation of automatic scaling and load balancing becomes easy for the developers in developing their cloud application for a cloud programming environment. Through this approach integration with other services (e.g. email, authentication, user interface) with PaaS-provider becomes easy. Hence, to a large extent the additional effort require to develop cloud applications can be reduced and is managed at the environment level. In addition, the developers possess the capability of integrating the other services with the applications as and when necessary. This results in making the development of a cloud application a simple task and also speed up the development time. In this connection, Hadoop [4] supports deployment environment on the cloud would be considered as cloud programming, as the application developers are offered with a development environment which is also refererred as Map/Reduce framework for the cloud. The process of the development of cloud applications becomes easy through these cloud software development environments.

**1.3.2 Cloud Service Management Layer**

This layer provides management of the applications and virtualized infrastructure for business solutions. This layer is responsible for providing virtualized resources for services such as service level management, policy management metered usage, license management, and disaster recovery. This layer supports scaling of applications through dynamic provisioning by allocating the resources to applications on demand results in minimizing the
underutilization of resources. Key components of Cloud Service Management layer are listed below.

**SLA Monitor:** When a customer first submits the service request, the requests are interpreted by the SLA Monitor to evaluate QoS requirements to determine whether to accept or reject the request. It is also responsible to monitor the progress of the submitted job. If any violation of the SLA is observed by the SLA monitor, it has to act immediately for corrective action.

**Resource Provisioning:** Availability of VMs and resource requirements are tracked through this mechanism. It manages the different requests coming to the virtual servers by the creation of multiple copies of VMs. The resource provisioner is self adjusted dynamically such that the processing is completed as per the requirements even at peak loads.

**Sequencer & Scheduler:** Based on the information from SLA Monitor and Resource Provisioning, the sequencer arranges or prioritizes the jobs based on objectives of service provider. Scheduler makes effective resource allocation by having the latest status information from the Resource Provisioning regarding resource availability and workload processing.

**Dispatcher:** The resources selected and assigned by the scheduler to the process are controlled by this module. It involves switching of context, switching of user, hopping to the proper location in the user program to restart that program, dispatch latency (i.e. the time required by the dispatcher to stop and start a process). It is also responsible for the start of the execution of selected service requests on the allocated Virtual Machines.

**Accounting:** It maintains a record of the actual resources usage by service requests in order to compute the final cost and charge the users. In addition, resource allocation decisions can be improved through the use of historical usage information.

**Metering:** Billing of the users is based on the usage of the system. Usually, billing is based on the usage of CPU per hour or rate of data transfer per hour. This mechanism also provides information about pricing policies and different types of services to customer. The customer
needs to select the level or quality of service by providing QoS service requirements without
the need to know how cloud provides the service.

**Load Balancer:** This mechanism contains algorithms for mapping virtual machines onto
physical machines in a cloud computing environment, for identifying the idle virtual
machines and for migrating virtual machines to other physical nodes. Whenever a user
submits an application workload into cloud system, one can create a new virtual machine.
Now the mapping algorithm of Load balancer will generate a virtual machine placement
scheme, assign necessary resources to it and deploy the virtual machine on to the identified
physical resource. Unmanaged and forgotten virtual machines can consume Datacenter
resources and cause energy waste. Another algorithm of Load balancer will identify idle
virtual machines and shut them off. In the process of optimally placing the virtual machine
onto the destination, we need to relocate the existing virtual machines. For doing this
operation, virtual machine migration algorithm of load balancer is invoked. In summary Load
Balancer will have the following three sub modules.

- **Migration Manager:** It triggers live migration of VMs on to physical servers
depending on information provided by the VM Mapper. It turns a server ‘on’ or ‘off’.

- **Monitoring Service:** This module collects parameters like status of application,
workload, utilization of resources, power consumption etc. This service works like
global information provider that provides monitoring data to support intelligent
actions taken by VM mapper. The status information is utilized to arrest the sprawl of
unmanaged and forgotten virtual machines.

- **VM Mapper:** This algorithm optimally maps the incoming workloads (VMs) on to
the available physical machines. It collects the information from monitoring Service
time to time and makes decision on the placement of virtual machines. The VM
mapper searches the optimal placement by a genetic algorithm provided in the next
chapter.

**Policy Management:** It is mandatory for organizations to define clear and unambiguous
definitions of governance, policy (i.e. regulatory), security, privacy [11] etc to make sure that
SLAs are not violated when applications are operated on the cloud. In order to deal with
business within a cloud, cloud consumers and providers are to be aligned on guaranteed
SLAs and equivalent pricing models. As the cloud capabilities are being improved (such as
virtual supply chains) policy-driven interactions that are fully abstracted need to be supported across clouds. It has become a main challenge for the cloud providers in modeling and extending policies in order to provide integrated services across distributed and heterogeneous business processes and infrastructure. Policy management has conventionally been fixed within and across organization boundaries of enterprise IT platforms and applications. Hence, globally expansion of businesses requires applying new methods to combine and complement the policies within and across external process networks and value chains.

**Advance Resource Reservation Monitor:** This is the mechanism to guarantee QoS in accessing resources across Datacenters. By reserving resources in advance, users are able to complete applications that are critical with respect to time such as parallel workflow applications that are realtime in nature require a number of resources in near future to run. The prediction of future demand and usage can be done by the provider more accurately. With this information, the provider can maximize revenue at various times to by applying policy management to determine pricing. Users will able to decide in advance in resource reservation as per the needs and net expenses as these costs are publicized in advance. To successfully plan and manage their operations it is essential for the enterprises to have prior knowledge of expected costs. Guaranteed supply of resources also helps enterprises to contemplate and target future expansion more accurately and confidently. Hence, enterprises be able to scale up or scale down their resource reservations based on short-term, medium-term, and long-term commitments. A mechanism for Negotiating the advance resource reservation using the alternate offers protocol is explained in [5].

**Security and Identity Management:** Cloud environments must control a identity [9] [13] and security infrastructure in order to enable elastic provisioning and to implement security policies across the clouds. It is necessary to control and ensure that sensitive information is protected against SLAs, as resource provisioning is done outside the enterprise’s legal boundaries by the cloud provider. The issues that need to be addressed by the cloud provider before convincing the end users to migrate from desktop applications to cloud applications are safety and security of confidential data stored on the cloud. Also the other issues are user’s authentication and authorization, up-time or down time and performance of the applications. Finally, data backup and disaster recovery to provide reliable SLAs for their cloud services.
**Autonomic Management:** Each element in the cloud service architecture is embraced with autonomic management capability. The autonomic managers determine task assignment to each of the available resources, reassignment of the tasks during workload execution based on the overall progress of the submitted requests. Also autonomic managers adaptively assign tasks in the workload to execution sites to meet given objective(s) to minimize total execution time and optimize for QoS targets. These objectives are imposed by an SLA. Autonomous system that can schedule a job to run at a certain date and time, dynamically trigger workload on unexpected business events, trigger workload via a web service, monitor for workload status information, get alerts on successful and unsuccessful completions, generate reports on job history, use conditional logic, provide a business process view, linking disparate processes together, schedule maintenance tasks in the applications, the database, etc.

**Green Initiatives:** Green Computing refers to the practice of implementing policies and procedures that reduce the impact of computing waste on the environment through the improved usage of computing resources. This movement has greater environmental concern involves every department in an enterprise. Green IT is the driving factor for the IT industry [6]. One particular area of interest in Green IT is “Datacenter”. IT departments are looking to reduce their carbon footprint of Datacenters to save the environment. Constraints on available Computing power, cooling capacity and physical space in an enterprise Datacenter facility impose serious limitations on ability to deliver key services. In these circumstances, “going green” in the Datacenter is not just about social responsibility, it is a business imperative.

To have a green datacenter, balanced utilization of power, cooling capacity and efficient infrastructure are the key components. In order to establish a green data center, it is important to understand how these components in a Datacenter have traditionally been deployed and to know the initiatives to be taken to make the datacenter green. Now a days renewable energy sources such as solar and wind with partial or complete power are chosen by many of the businesses. Of all these, “Energy Efficiency” provides the greatest potential for quick return on investment, ease of implementation, and financial justification. There are several successful green Datacenter initiatives to help the enterprises to overcome the energy and capacity limitations, operational vulnerabilities, and constraints that limit today’s Data center [7].
1.3.3 Datacenters Layer

Datacenters Layer is at the bottom of the cloud service architecture. Normally big enterprises with huge hardware requirements in need of subleasing Hardware as a Service (HaaS) are the users of this layer. The HaaS providers operate, manage and upgrade the hardware on behalf of their consumers in the duration of the lease or contract. This helps the is enterprises, as it relieves them from upfront investment in building and managing Datacenters. Meanwhile to maximize profits, HaaS providers have the cost-effective infrastructure and technical expertise to host the systems. As enterprise users have predefined business workloads, SLAs in this model are stricter due to severe performance requirements imposed. The HaaS providers materialize the profits from the economy of scale of building huge Datacenters infrastructures with gigantic floor space, power, cooling costs as well as operation and management expertise. A number of technical challenges need to be addressed by the HaaS providers in operating and managing their services. The major challenges are efficiency, ease and speed of provisioning to large scale systems. Datacenter management, scheduling, and power-consumption optimization are the other challenges that arise at this layer.

Virtual Machines: Virtual machine is the fine grain unit of computing resource. cloud users will have the flexibility on their VMs for performance and efficiency as they have super-user previliege in accessing to their Virtual Machines. The users can customize the software stack of Virtual Machines. Frequently such services are referred as Infrastructure as a Service (IaaS). Virtualization is the primary technology in a cloud environment that supports the users with extraordinary flexibility in configuration of the settings without disturbing the physical infrastructure in the datacenters of the providers. The concept of IaaS has become possible through recent advances in OS Virtualization. Multiple application environments will be supported by each of the virtual machine that runs a different operating system which is referred as guest operating system.

Virtual Machine Monitor: It is a hardware abstraction layer that acts as an interface between virtual Machines and hardware. It coordinates the access of resources by all virtual machines. VMM enables organizations to speed up the response to business dynamics through consolidation of computing resources thus results in less complexity in management of computing resources. Improving the resource utilization and reducing power consumption are key challenges to the success.
**The Hardware:** The hardware offers basic computing resources such as CPU resource, Memory resource, I/O devices and switches that form the backbone of the cloud.

### 1.4 Research issues in Cloud Computing Environment

This section explores the different research issues in cloud computing environment.

#### 1.4.1 Security Issues

In this section we present a review of the security issues in Cloud Environment.

**Availability of Service:** Most of the Organizations are some cautions of cloud computing and sufficient availability is existing in Utility Computing services. In this view, all available SaaS products have a high standard. Users expect high accessibility from cloud facilities; it attracts large base of customers with an opportunity to transfer their business to Cloud in critical situations. It is possible to accept for different companies to provide independent software stacks for them, but it is very hard for single organization to justify maintain and create more than one stack in the name of software dependability.

And one more availability problem [10] is Distributed Denial of Service (DDoS) attacks [15] [16]. Attackers make use of large bonnet’s to reduce the profits of SaaS providers through DDoS by making the services unavailable [19]. A long botnet attack may be difficult to maintain, due to the fact that longer attacks may easily be uncovered and defend against. Hence these attacking bots could not be immediately re-used for other attacks.

**Data security:** One of the crucial topics of research is Data security in a cloud. As the Cloud service providers do not have permission for access to the physical systems of data centers, they have to depend on the infrastructure provider to get full data security. In a cloud environment, the service provider can only specify the security setting remotely but not sure whether those settings were implemented fully or not. In this Process, the infrastructure provider must reach the following objectives: (1) *confidentiality*, for secure data access and transfer, and (2) *auditability*, whether security settings of applications have been tampered or not.
Unencrypted data in placed in a local data center in a cloud is not secured when compared to the encrypted data. Confidentiality is can be achieved through cryptographic [12] protocols. Remote attestation techniques can be used to achieve auditability and it could be added as an extra layer apart from the virtualized guest Operating System. An additional logical layer may be added to maintain confidentiality and auditability.

Trusted platform module (TPM) is required for remote attestation of non-forgable proof of system security. In virtual environment, Virtual Machine can change dynamically the location from one to other. It is very difficult to construct trust mechanism in every architectural layer of the cloud. Migration of Virtual Machine should take place only if both source and destination servers are trusted [20].

1.4.2 Data Issues
In this section we exploit some of the Data issues in Cloud Environment.

**Data Lock-In:** Software stacks have better interoperability between platforms, but customer’s feels difficult to extract their programs and data from one location to run on another. Some organizations are concerned about extracting data from a cloud due to which they don’t opt for cloud computing. Customer lock-in appears to be prominent for Cloud providers. Generally cloud users are more bothered about price change, consistency problems, or even to providers leaving out of business. SasS developers could take the benefit of deploying the services and data across multiple clouds supported by different providers so that failure of a single service provider does not affect the customer data. The only fear is that they are much worried about the cloud pricing and flatten the profits. Two solutions were offered to relieve this alarm. First, the quality is an important issue when compared to the price as customers may not only be fascinated towards the lowest cost service. The other issues is data lock-in justification. APIs standardization leads to a new model for Private Cloud and a Public Cloud with same software infrastructure usage. This option Available in “Surge Computing,” due to high workloads it is not easy to run extra tasks in private clouds compare to the public cloud [21].

**Data Transfer Bottlenecks:** Applications maintain to be converted into additional data-intensive. Data placement and transport may be complicated if the applications are moved across the boundaries of clouds. Cloud providers and users have to feel about to minimize costs on the concept of the traffic and the implications of placement at each level of the system. By shipping transfers to disks high cost of data transfer bandwidth may be reduced. It
is one more provision to keep maintains data in the cloud. Data availability in the cloud may not block enablement of a new service. If data is archived in the cloud then vending of Cloud computing cycles with the new services become possible, such in all your archival data by creating searchable indices or group the images according to who appears in each photo by performing image recognition on all your archived photos.

A third opportunity is quick minimization of network bandwidth cost. One can estimate that one-third of the cost is the fiber cost and two-thirds of the cost of WAN bandwidth is the cost of the high-end routers. Some researchers are exploring lower cost with controlling of routers in centralized manner instead of the distributed approach. If this technology were deployed in WAN then WAN costs dropping more quickly.

Traffic management and analysis: Data traffic Analysis is one of the important challenge for today’s data centers. Nevertheless, there exist various techniques for analyzing the traffic measurement and data analysis in Internet Service Providers (ISPs) [94] networks and enterprises. But extending them to data centers is not so easy. The challenges that should dealt with are: In a cloud setting as the density of links is much elevated than that in ISPs or enterprise networks, that results in a worst case scenario to deal with while extending existing methods. Secondly, most of the existing methods can scale upto a few hundred end hosts while computing traffic matrices where as a modular data center may have quite a few thousand servers. Finally, the flow patterns assumed by applications in existing internet may not hold in applications deployed on data centers as they scale up or scale down dynamically. Further, the existing methods work for tight coupled setting of application’s use to network, computing, and storage resources where as loosely couple setting followed in a cloud. Presently there exists less work done on analysis and measurement of traffic in a data center. Greenberg et al. [22] has reported characteristics of data center traffic by taking flow sizes and concurrent flows which are used in design of network infrastructure. Benson et al. [23] presented a study on data center traffic at the edges through the examination of SNMP traces from routers.

Reputation Fate Sharing: Virtualization may not be fit for reputations. A single customer’s bad behavior may affect the reputation of the cloud all together. “Trusted email” services are hosted in smaller ISP’s setup with little cost and with little experience creates reputation-guarding services. From the cloud provider point of view the legal issue that need to be taken care is the transfer of legal liability and maintainability by customer.
1.4.3 Performance Issues

**Virtual Machine Migration:** Virtual Machine migration support lot of advantages in a cloud through load balancing across data centers. Through migration of VMs robust and high response is achieve in data centers. Actually migration of VMs was adopted from process migration. Through VM migration hotspots may be avoided even though it is not simple and straight forward. Detecting workload hotspots and unexpected workload changes requires initiating a virtual machine migration. During migration the process in execution should be transferred effectively. By considering resources and physical servers, the transfer can be done in consistency for applications.

**Server consolidation:** Server consolidation [18] serves as an efficient approach in minimizing the energy consumption that results in best use of resources. During server to consolidation all the VMs running on various under-utilized servers are migrated through live Virtual Machine migration approach and all the servers that were unused will be set to energy-saving state. This technique does not impact the performance of the applications running on VMs. As the individual VMs may vary time to time, resource usage will also change from time to time that results in change of footprint of a server. Resource congestion may be resulted due to sharing of resources such as CPU, I/O, network bandwidth and memory cache among VMs on server. This fluctuations of VM footprints helps in effective server consolidation. As and when the resource congestion occur, then the system react quickly through server consolidation.

**Performance Unpredictability:** In a cloud computing environment [14] sharing of I/O is difficult than sharing CPU and main memory. To improve the efficiency of operating system and architecture one of the solution is to virtualize the I/O interrupts and network channel. Technologies such as PCI express are very difficult to virtualize in a cloud environment. As an alternative, Flash memory might be a smart solution with minimized I/O interference. Through this the possibility of sharing I/O per second with more storage capacity can be achieved through multiple virtual machines. Virtual machines with different I/O workloads can exist together in a single physical machine without the any interference.

In high performance computing [24] the unpredictability problem refers to scheduling of virtual machines of various classes with batch processing programs. This problem to in HPC is not in the use of clusters as the processing majority of parallel computing applications
at present is done in huge clusters using the message-passing interface (MPI). HPC applications require that all the threads of a program should run simultaneously and today’s virtual machines and operating systems do not guarantee a programmer with a visible way to ensure this. Gang Scheduling [25] can be used to overcome this difficulty in a cloud computing environment.

**Scalable Storage:** The important properties are of a cloud are unlimited capacity on-demand, no initial investment, short term usage. This is an open research problem, is not only create storage system [26] to meet these issues. And consider cloud advantages of programmer expectations and scaling up and down randomly on-demand in this regard to resource organization for durability, high availability and scalability of data.

**Bugs in Large-Scale Distributed Systems:** Another challenging issue in Cloud Computing is recovering from errors in large scale data centers. The debugging of these bugs have to be completed in large scale in data centers as these bugs cannot be reproduced in smaller configurations.

One approach may be to rely on virtual machines in a Cloud environment. SaaS providers developed their applications without the support of VMs. Virtuals machines may not be opted by SaaS providers due to the fact that VMs were popular recently and they may not provide the required performance. Since Virtual Machines are crucial in Utility Computing, the level of virtualization determines the success or failure of the applications in as large distributed system.

**Scaling Quickly:** The approach of Pay-as-you-go id definitely applicable to network bandwidth and storage on the basis of byte count of data used. Computation may be differed based on the level of virtualization. Google AppEngine automatically scales in response to load up and down, and users are charged by usage of the cycles. Amazon Web Services charge based on the number of instances occupied in an hour without taking into consideration that whether the machine is idle or not.

Another opportunity is to scale quickly increase and decrease resources without violating service level agreements automatically in response to change in load in order to save money. About two-thirds of the power used by a idle computer is compared with a busy computer. Currently, in datacenters in cloud environment receive a great deal of negative attention but careful use of resources could reduce the impact.
Cloud providers already have a careful accounting mechanism for consumption of resources. Some of the metrics of accounting in the concept a cloud are: per-byte and per-hour costs, development difficulties.

**Latency:** Latency is one of the important research issues on the Internet. Any degradation of performance of an application in a cloud results in performance degradation in terms of the result to the client. The latency is allows understanding where and how the applications are running with both the smart platforms and intelligently planned infrastructure. In future, capacity of a cloud and cloud based applications will increase rapidly hence latency also increases. latency in a Cloud computing environment must be improved through the improvement of latency in the desktop PC which is the largest bottleneck in the usage of memory and storage.

### 1.4.4 Energy management

Increasing the energy efficiency is one of the major challenge in a cloud computing environment. In a recent study [27], it has been estimated that the cost of powering and cooling accounts for 53% of the total operational expenditure of data centers. IaaS providers are to expected to reduce energy consumption under massive pressure. In data centers, the aim are only to reduce energy is cost. Data centers are recently received considerable attention on designing energy-efficient data centers. This approach leads to several directions. For example, slowing down of CPU speeds and turning off partial hardware components [28] to improve energy efficient through hardware architecture has become common practice. Server consolidation [29] and energy-aware job scheduling are two solutions to minimize power consumption by setting the unused machines to energy saving state. Current research focus on network protocols and infrastructures that are energy-efficient [30]. A key challenge is to achieve a good trade-off between energy savings [95] and application performance. Recently some of the researchers are working on the solutions for improving the performance and reduce the power consumption in a cloud environment that is dynamic in nature.

### 1.4.5 Software Frameworks

Cloud computing provides a convincing platform for hosting significant data-intensive applications. In this category, Hadoop an open source initative provides scalable and fault-tolerant data processing by through Map Reduce framework concept. Any MapReduce job is highly dependent on the type of the application [31] and shows the relation
between performance and resource consumption. In cloud setup that uses Hadoop, all nodes may have heterogeneous characteristics. All the applications may be run through the initialization of Virtual Machines. Therefore, optimization of performance as well as cost is possible for a MapReduce application through the careful selection of configuration parameters and along with an efficient scheduling algorithm [32]. The bottleneck of resources availability can be reduced by executing applications at different intervals. The challenges that need to be addressed in designing a hadoop based cloud includes modeling of performance of Hadoop jobs in all the possible cases and scheduling conditions dynamically. Energy efficient MapReduce [33] is the other approach to turn Hadoop node into sleep mode after it has finished its work while waiting for new assignments. Some researchers are still working on a trade-off between performance and energy-awareness.

1.4.6 Novel cloud architectures

In commercial clouds large data centers are operated in a centralized fashion. This design results in high manageability and economy-of-scale. The limitations of this approach are: constructing large data centers such requires huge initial capital investment and high energy expense. Study on size of the datacenters [34] suggests that small size data centers may result in more savings than big data centers in many ways as listed below: 1) less power is consumed by a small data center, hence it does not require a powerful and expensive cooling system. 2) geographically building the large data centers are expensive than small data centers. 3) Due Geo- diversity of large data center Content delivery and interactive gaming which are time-critical services results in slow performance. For instance, Valancius et al. [35] presented a feasibility study on hosting of video-streaming services through application gateways. The other related research direction is on using deliberate resources (i.e. resources supplied by users) in deploying cloud applications. If the cloud application are a mixture of voluntary and dedicated resources then they are much cheaper to operate. These kinds of clouds are more suitable for non-profit applications such as scientific computing. This architecture proposed the new design challenges such as frequent churn events and management of resources that are heterogeneous in nature.

1.4.7 Software Licensing

Commercial software licenses that exists nowadays usually have the control on which computer the software should run. Initially charge and annual maintenance charge which are fixed will be paid by the users. But for Cloud applications, this licensing approach may not be
suitable as many cloud service providers depends on open source software [36] for providing services. The primary solution is to change the licencing approach that suits the needs of cloud based applications and needs of the customers.

The primary challenge for software companies to sell the Cloud applications through sales support services. The approach of Pay-as-you-go may not be economically viable with a single purchase. This challenge can be overcome through offering discounted plans for bulk usage to promote the bulk purchases and use.

1.4.8 Task Assignment

Cloud computing is a new paradigm for enterprises that can effectively facilitate the execution of tasks. Task scheduling is an important issue which is greatly influencing the performance of cloud computing environment. The cloud service providers and consumers have different objectives and requirements. Meanwhile, the availability of resources and the load on them dynamically varies with time. Hence, scheduling resources in Clouds is a complicated problem.

Task scheduling algorithm is a method by which tasks are matched, or allocated to data center resources. Due to conflicting scheduling objectives generally no absolutely perfect scheduling algorithm exists. A good scheduler implements a suitable compromise, or applies combination of scheduling algorithms according to different applications. A problem can be solved in seconds, hours or even years depending on the algorithm applied. The efficiency of an algorithm is evaluated by the amount of time necessary to execute it.

1.5 Chapter Summary

In this chapter we presented cloud computing basics and services offered by cloud computing paradigm. This chapter explored various research issues in cloud computing environment. One such challenge is mapping a set of tasks onto available cloud resources generally known as task scheduling. Next chapter provides precise elucidation of multi objective task scheduling problem, and its mathematical modeling.