CHAPTER 1
INTRODUCTION

Rapid emergence of Cloud Computing has drawn attention of lots of organizations, communities, groups and individuals towards Cloud services. Cloud services have inherent benefits like very low execution costs, hassle free services, universal accessibility and on demand availability of services on pay-as-per-use basis. Backbone of Cloud Computing is Virtualization Technology (VT) [1]. Virtualization helps in improving the manageability, utilization, and reliability of computer systems. Virtual Machine Monitors (VMMs) facilitate this by running multiple instances of Virtual Machines (VMs) on a single physical machine that are configurable, updatable, reusable and manageable. Rapid advances in cloud computing have been facilitated by recent advances in Internet backbone, high performance, and scalable infrastructure in the web and data centers [2].

1.1. WHY CLOUD COMPUTING?

Cloud Computing is the fastest growing technology due to continuous growth of internet users. This is because users can have easy access to shared pool of resources e.g. servers, networks, storage, applications at remote and geographically distant locations on demand around the globe through resource virtualization. It has three major features as follows:

- Pay as you use: Usually, business organizations need to setup their own IT infrastructure to meet the business requirements of storage, network, services etc. which requires huge business capital. Through cloud, they can get these infrastructures as a service, which saves huge upfront investment in capital by paying for only what they use.

- Flexible Allocation: The resources are never allocated permanently to any particular user. The resources are allocated to user on demand from shared pool of resources and returned back to pool after completion of its use. They can be re-allocated to a different user.
• **Virtualization:** It provides resources in a scalable manner which makes it cost effective and hence becomes very critical to cloud computing. It has three characteristics which makes it ideal for cloud computing- partitioning, isolation and encapsulation

(i) Infrastructure As A Service (IAAS) (ii) Platform As A Service (PAAS) (iii) Software As A Service (SAAS)

Major advantages of cloud computing adoption is utilization of services without owning and managing the computing resources, which reduces the burden of upfront investment and management of resources and supporting staffs. The services and resources are scalable, which can be accessed from anywhere, anytime. Though the underlying services were already in use namely managed services, SaaS, Web Services, utility computing and PaaS etc., all of which has been provisioned on the Internet using the web browsers and named as cloud services. These services are offered by several small and large vendors via the Internet to the enterprises and end users. Some major cloud computing vendors are SalesForce.com, which provides Application-as-a-Service., Workday-for ERP applications, Google Application Engine, Amazon’s Web Services, Zoho, Microsoft’s Azure, which offer services like IaaS, PaaS and SaaS in different ways. Computing resources are also provisioned on the cloud using virtualization technologies like server virtualization, storage virtualization and operating system virtualization and so on [3]. Different cloud service providers and their services offered have been reviewed along with the claims made by these vendors regarding their reliability and security issues. In future, certain additional players in the cloud industry may emerge to provision and manage virtual data centers and servers. Already few service providers like 3Tera, Liquid Q, Terremark, Layered Technologies, XCalibre are offering their solutions for deploying and managing applications in data centers. Many users may adopt hybrid systems of cloud i.e., along with public clouds, private clouds infrastructure are also set up within the premises of the organizations by using virtual data centers running within the firewalls. Hadoop may become the center of attraction as it is an open source cloud computing software platform which is currently being used by big giants of cloud computing like Yahoo, Facebook, IBM etc. [4].
1.2. FUTURE OF CLOUD COMPUTING

Today the world of computation is moving towards pay-as-per use model due to the numerous benefits provided by this model. Hence cloud computing is predicted as the preferred option for future computational world. A major goal for moving to this model is to make better utilization of distributed system resources that yields optimal throughput for tasks ranging from small to large-scale computations. Many success stories of Amazon's EC2, S3 GAE, Mapreduce, Hadoop inspired computer users to shift to Cloud Computing [5]. Cloud Computing has facilitated computations and processing of even extremely large sizes in unexpectedly less time. One major issue in cloud computing is to understand the complex cloud components and the way it works to leverage services. In cloud model, all arenas of knowledge and applications are required viz., financial issues, legal issues, administrative issues, last but not the least the technical issues. These fields of knowledge and implementation are an integral part of cloud architecture. All these arenas are required to support operations of cloud model for leveraging services.

The current cutting edge developments show that in 21st century, most computing will be done through internet services over lightweight portable devices rather than on desktop PCs. This supports the emergence of Cloud Computing in which computing and data moves to portable PCs and the data centers. Most applications will be delivered as a service over the Internet on to cloud infrastructures, viz. computing hardware, systems software and applications software, which are the services provided by cloud computing model [1]. The backbone of cloud computing emergence is because of the availability and advances in broadband and wireless networks, reduced storage costs and Internet computing technologies. User needs at peak load can be met by scalable capability of these services with reduced cost of operations and access. They need to pay only for the services used. The service provider can cater to the needs of the clients by virtualizing resources which leads to efficient utilization of resources and at the same time hiding the details of underlying complexities. Backbone of cloud computing infrastructure and services are due to advances in virtualization technologies, development of service oriented software, advancements in grid technologies, managed services and efficient power utilization innovations.
John McCarthy in 1961 predicted that “Computation may someday be organized as a public utility” which resembles with the cloud computing model [2], but at the same time addressing the significant problems and unexploited opportunities related to deployment and efficient operations and use of other computing resources. The paradigm leverages a huge change in the existing model of software and hardware infrastructures. It is believed that previously Moore’s Law suggested the increase in the clock speed per chip whereas in the current scenario, there should be increase in the number of processors and threads per chip. New systems must be designed to increase the parallelism in computing. At the same time software which support parallelism may pick good attention namely MapReduce for parallel and data-intensive computing. Attention may also shift from existing storage devices to smart solid state devices like flash, hybrid hard disks, although expensive today but new algorithms and data structures may take care of these issues in due course of time [6].

1.3. CLOUD SERVICES AND MODELS

A taxonomical classification is presented here for better understanding the cloud model and its functionalities. The design of cloud architecture is to support the services to the end users irrespective of their geographical locations, i.e., throughout the world. It appears to the end users as if there is single point access to all services. Cloud services are categorized as Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS) as defined by National Institute of Standards and Technology (NIST) [3]. These category of services are facilitated by many small and large service providers. For example, Software as a Service (SaaS) that inherently supports multi-tenancy property, is provided by Salesforce.com, Microsoft, Oracle, IBM etc. A cloud based Platform as a Service (PaaS) is provided for software developers. The developers can choose any platform of their choice for developing the complete software life cycle programs, i.e, Software Development Life Cycle (SDLC). Their programs can be developed, tested, deployed and hosted from the cloud. Some of the major PaaS service providers are Microsoft’s Azure, Google AppEngine, Amazon’s Web Services etc. Execution cost, operational complexities, administrative issues and development time is highly reduced if these services are used from the Clouds, as compared to the traditional ways of executions. Another service type provided from the cloud is Infrastructure as a Service (IaaS). These services facilitate to provide the computer
infrastructure in a flexible and scalable way on pay-as-per-use basis. Some of the major infrastructure service providers are RackSpace, FlexiScale, Joyent, GoGrid, AppNexus etc. Three service models are given by NIST, namely Private, Public and Hybrid. Private Cloud is set up in the premises of the organization. It offers the services from within the organization. These cloud services are costlier as compared to the other types of services, but at the same time data and processes are more secure, as they are managed within the organization. Hence this model supports prevention from security exposures and reduces complexity of legal hassles in Service Level Agreements as is the case in Public and Hybrid Clouds. Examples of private clouds are Eucalyptus, VPC from Amazon, VMWare, Enomaly and Intalio etc. [7].

Services offered to clients can be classified as IaaS, PaaS, SaaS and utility computing. The architectural models suggested for the clouds are divided into two parts according to the services provided.

- To provide varying computing power on demand using instances to provide services as PaaS and SaaS.

- To provide data and scalable intensive computing resources that is provided on demand.

These services are leveraged by pay-per-use model which is compatible to clusters of loosely coupled processors. Based on user’s requirements of services, cloud computing services are categorized as:

- Public clouds- available to general public
- Private clouds- restricted to a single organization or group
- Hybrid clouds- shared by multiple groups or organizations on individual basis.

Public Cloud provides open access to all public and other organizations. This is the highly used and future demanding cloud model, whereby the provisioning is done dynamically in a fine-grained, self-service manner using web technologies and services over the internet. Examples of Public cloud are FlexiScale, GigaSpaces, Zimory, Amazon’s Elastic Cloud Compute (EC2), Windows Azure, SymmetriQ and SunCloud etc. Other cloud service type is known as Hybrid cloud whereby the Private and Public both types of services are used together to optimally use the services for varying application requirements. Private cloud
services are used for secure data and processes, whereas the public Cloud is used where comparatively less or minimal security exposures are required. For example, Elastra, Cloud Backup from Aisgra, Carpathia and Right Scales etc., [8].

The cloud computing comprises of processing power, network and storage devices. All these are abstracted in three different layers namely-

- **Application Layer** - The topmost layer which provides software applications as a service.
- **Platform Layer** - The middleware abstraction which provides the PaaS to develop, deploy, test, host and manage all integrated services.
- **Infrastructure Layer** - The lowest layer to deliver the computing and storage facilities in a standardized way, namely servers, storage devices and other networking devices like switches, routers, which altogether handle the processing during varying loads.

### 1.4. VIRTUALIZATION IN CLOUD COMPUTING

Virtualization is the major technology being used to provision the resources viz., software virtualization, hardware virtualization, and network virtualization. These virtualization strategies are implemented at hardware levels, operating system levels and at application levels. Full virtualization, Para-virtualization, System virtualization and Application virtualization techniques are being used depending upon the application requirements. Classic benefits of virtualization include improved utilization, manageability, and reliability of computer systems. Several users with different OS requirements can share more easily a virtualized server. OS upgrades can be staged across VMs to minimize downtime and failures in guest software can be isolated to the VMs in which they occur [9]. These benefits have been considered valuable in high-end server systems. Recent academic research and emerging new VMM-based products suggest that the benefits of virtualization have wider appeal across a broad range of both server and client systems. Virtualization, which is the backbone of heterogeneous distributed infrastructure, has become more adaptable. Cloud computing model is based on customized environment created on top of physical infrastructure using Virtual Machine technologies. The major benefits of using VM technology are server consolidation, by which major portion of the workloads can be assigned to only few underutilized machines. By virtualization, the applications can run in virtually partitioned machines which avoid
interference from other applications and APIs, hence increasing the security of separately running applications. Overall performance of individual applications also gets a boost in terms of quality of service guarantees. The existing VM based system can only manage the clusters within a site, as the resource managers. Virtual Infrastructure Engines (VIEs) let users create customized virtual clusters by sharing of the physical machines’ availability at that site. The new Virtualization Technology reduces the security concern by using virtualization software, which is used to realize the InterGrid architecture. It simplifies the deployment of execution environments on different sites without requiring access to hosts’ operating systems at the computing sites. Out of the three cloud computing services, namely SaaS, PaaS and IaaS, only IaaS is considered which provides storage and computing resources as a Service. Amazon’s EC2 (Elastic Compute Cloud) allows the users to deploy VMs on its infrastructure on payment basis for the used resources namely computing, storage and network resources [3].

1.4.1. Current Trends in Virtualization Technology

Virtualization-related research has made its way not only into the cloud field but also into the other scenarios, resulting in creation of a new class of rootkits. VMs are software entities that emulate a real machine’s functionality and execute under the control of a hypervisor that virtualizes and multiplexes low-level hardware resources [10]. There are two types of hypervisors: non-hosted, which run directly on top of the hardware, and hosted, which are integrated with a host operating system (OS). The presence of a hypervisor makes VMs subject to a level of visibility and control that’s hard to achieve with real machines. These small size, isolation, and mediation power of an ideal hypervisor over VMs make it an interesting candidate for a trusted computing base, with applications in security research fields such as intrusion detection, integrity protection etc. in distributed and cloud computing environment. All these cloud service models were practically possible because of the re-invented Virtualization Technology (VT).

According to Gartner’s hype cycle for Cloud Computing for the year 2014, the trend shows that VT is at its slope of second phase productivity. It implies that lots of work is going on for enhancing and developing VT, so that it may be used in an economic and efficient way, whereby the usage cost for the end users is highly reduced. VT provides a software solution for resource abstraction and leveraging the services in a consolidated and isolated manner.
Virtualization is inculcated at different software layers depending upon the application type, its usages and service demands by users. Though at the back-end, it is essentially used to optimally provision the resources. Some virtualization techniques that are adopted are Server Virtualization, Storage Virtualization, I/O Virtualization and Network Virtualization etc. Server Virtualization refers to mapping of single or cluster of physical servers to logically partitioned cluster of servers whereby the logical clusters of servers are created that belong to different physical clusters. The advantage of doing so is that the sizes of clusters can be dynamically varied as per the user’s demand, making it highly scalable in nature.

Various commercial and freeware software solutions are designed by different organisations known as VMMs. These VMMs facilitate to run multiple isolated VMs onto a single physical machine, thereby consolidating the workload. Some commonly used VMMs are Xen, Virtual Box, VMWare, KVM, Denali, VMMs from Microsoft etc., [11].

1.4.2. Advantages of using Virtualization Technology

Cloud computing uses some fundamental services, e.g., discovery for promoting reusability, data replication and consistency for performing atomic transactions, load balancing to avoid computational overloads onto some set of machines, whereas other machines within the same computational setups are free or less loaded. Various applications are provisioned dynamically at the run time. Similarly, the applications can also be de-provisioned on-the-fly without changing the system and network configurations. Load balancing also has an impact on energy consumption and optimal resource utilization of underlying system resources, either directly or indirectly. This further affects the cost of operation and contributes towards green computing [12]. Additionally, the hardware infrastructure and their individual components are also protected from over load. It also contributes to scalability feature when there is sudden increase or decrease in the load. It is predicted that all our content will be in the virtualized environment, encrypted by means we trust and control and duplicated across carriers and providers. So, from inexpensive, easy-to-secure computer, which can be purchased from a local store on payment basis, one can access anything of choice, without possessing any of it other than the images appearing on your screen when you need them. Vast amounts of computing power and shared data sets can be used by millions of people around the globe. One can pay only for what is being used. The service providers are highly competitive, when
one service provider increases prices, reduces services, or becomes intolerable to the users’ needs, one can instantly and transparently migrate to other Cloud Service Providers (CSPs). Users need to worry about running complex IT infrastructure and building a lot of customized applications [13].

1.5. CHALLENGES AND ISSUES IN CLOUD COMPUTING

There are various challenges and issues in adopting cloud service model, e.g., technical issues related to data center management, data governance, resource allocations, trust development, fault tolerance, energy consumption. In addition, there are other challenges too, e.g., issues related to administration, billing, accounting, privacy, legal and service level agreements etc., Some of the major technical challenges are mentioned here.

1.5.1. Resource Allocation

Efficient allocation and scheduling of physical resources among multiple host Operating Systems are required to be addressed. Now a day, Virtualization is inherently supported by processor manufacturers, e.g, Intel supported Virtualization Technology [14]. The same has been extended to different Cloud Platforms provided by different Cloud Service Providers at software levels [15]. In our work, open source VMM has been chosen so that additions/modifications can be implemented and tested. Resource provisioning for VMs are done by VMM. This facilitates to develop, implement and analyze various resource allocation and scheduling techniques. Virtualization and Resource allocation strategies can be implemented at hardware, operating system and network levels. Full virtualization, Para-virtualization, System virtualization and Application virtualization techniques are being used by the organizations, depending upon the application requirements [16][17][18]. Xen is commonly used VMM in industries, academia as well as for research and development [19].

Cloud service resources are allocated and provisioned in such a way that it can be automatically and dynamically provisioned without manual intervention. This feature hides the complexity of the cloud’s internal systems from the application developers and end users [20]. One major issue in cloud computing adoption is security of data. This is because the data is frequently required to be moved from data centres to the end user and vice versa. Developing and building a trust on the CSPs is a major design challenge. This inherently
involves performance, system design, data governance, deployment and risk management. CSPs are competing to establish and implement best possible security practices and policies to satisfy the service users. Various encryption and decryption policies are implemented before the data is moved on to the communication channels. Most advanced algorithms are used e.g., AES-128, 192, 256, and RSAs etc. Various privacy and protection mechanism are adopted to user’s data.

1.5.2. VM Migration

VM migration and Process migration are the two migration techniques that are generally adopted to balance the system loads. There is an inherent difference in the operating concepts of VM migration and process migration. Process migration demonstrates a functionality of transferring a process running on one machine to the other. In practice, process migration is difficult and complex as it takes care of legacy applications and at the same time it also moves Process Control Blocks (PCBs) to maintain complete isolation on other machine. These can be overcome by using a VMM based migration [21]. VMMs such as VMware use hardware abstraction to encapsulate the complete Operating Systems(OS) environment in such a way that it can be suspended from one machine and resumed on the other one provided there are inherent similarities in system architecture of OSes. VM migration [22] supersedes Process Migration, except in some cases that occur due to the narrow interface between a virtualized OS and the VMM. VM migration has the advantage of transferring internal memory states in a consistent and efficient way [19]. VM Load Balancing is crucial characteristic of system virtualization. VMMs can allocate and shift the running applications dynamically to other physical machines as and when the load increases on any particular machine [23].

1.5.3. Trust Awareness and Fault Tolerance

In Virtualization supported environments, user requests are forwarded to the VMs. Selection of VM depends on the requirements of the tasks. Hence it is an important issue to identify the trusted VMs onto which the requests can be forwarded. Trust can be defined as an entity based on reliability and firm belief based on attributes of the entity. Firm belief is not a fixed value associated with the entity. Rather, is subject to the entity’s behavior and is applied only within a specific context at any given time [25]. This means that the firm belief is a dynamic value that varies with time. Based on the past reputation, a default trust value for grid computing is
taken into account for calculating the trust value [26]. Past experiences involve the calculations based on previous transactions and trust values. These scheduling algorithms are implemented in real clouds. For example, in Eucalyptus, Greedy and Round-robin algorithms are implemented. Amazon uses trust and reliability based scheduling algorithm [27][28]. All algorithms mentioned above do not take into consideration fault rate of the datacenters. Since datacenters or nodes are the resource providers and if the resource provider is itself faulty, then the expectations of QoS by users cannot be fulfilled. Hence trust based framework and fault aware scheduling techniques are the need of the hour.

Fault tolerance is one major issue because failure in cloud model is inevitable. Various CSPs have history of major failures, downtimes, network breaks and many other failure issues. Hence reliability, outage and failure is one major issue. Various fault tolerant recovery and redundancy measures are taken to minimize the problem, e.g. Recovery-Oriented-Computing (ROC), Google File Systems, Server replication measures are adopted in Big Data centres [29].

1.5.4. Energy Aware Computation

Since large data centers and computations are involved in cloud model, hence energy consumption is one of the major design issues. Lot of work is going on to develop new tools, devices, techniques and algorithms to minimize the energy consumption. Cloud federations are also done by various cloud service providers and cloud brokers to combine the data and identities of users across different and multiple systems [30]. This helps in identifying the users across a single or multiple portals by providing a card to access these portals. Various authentication and authorization measures are also taken at the machines that are placed on the clouds. These are done by single or multiple sign-on when the users try to access these machines [31].

Additionally, since data centers require large amount of power, there is a need to address the Energy issue while maintaining QoS. Energy Aware Computation in Cloud environment is one major cause of concern for researchers and Cloud Service Providers. Some heuristic algorithms and Task Consolidation based [32][33][34] on energy consumed by each task were suggested to reduce the energy usage. Some Hotspot and Coldspot based migration techniques were suggested to reduce the energy consumption [35]. Static and Dynamic VM migration
techniques were suggested that forcibly migrate the VMs on less loaded physical machines. Dynamic scheduling algorithms for reducing the energy consumption were suggested. These algorithms were based on the levels of imbalances in the system resource consumptions [36]. Consolidating the VMs onto the optimally running physical machines and switching off the remaining physical machines based on work-load awareness were suggested to save energy [37][38]. Scheduling techniques based on cost constraints, deadline aware and other energy aware parameters for homogeneous and heterogeneous processors were suggested [39][40].

1.5.5. Scalability
As the number of sites increases, monitoring and reporting about their operations becomes more complex. Various issues are required to be taken care of viz. All the phases of development, debugging, testing and their performances are to be monitored. Hence a good mechanism should be there especially for SLAs and managements. Few examples of monitoring mechanisms used in cloud systems are SimpleDB, HyperQ and SimpleQueue that are offered by Amazon Web Services (AWS). Another important service management is providing a transparent system for metering and billing. This has a great impact in increasing the trust of Cloud Service Provider (CSP). Various types of pricing strategies are adopted by different cloud vendors, depending upon the type of service they provide. For example, data storages are billed on minutes, hourly, monthly and yearly basis. Computational costs are calculated on cycles/sec, cycles/hour or per day basis and so on. Different cloud vendors adopt different techniques and strategies for billing but in a justifiable manner as the end users may have a comparative study of cost for any given service by various CSPs [41].

1.6. RESEARCH OBJECTIVE
The main objective is to explore Resource Allocation Strategies in Cloud Computing environment for optimal utilization of Resource with respect to cost, time, trust, fault tolerance energy consumption and reliability. Efforts are made to maintain or enhance Quality of Service while implementing above strategies. The main objectives are mentioned below:

- To develop systems parameter based approaches for VM Migration and Load balancing for optimum resource utilization.
- To develop trust and fault aware allocations strategies for trusted servers.
➢ To develop fault aware load balancing algorithm for data centers using Content Delivery Networks (CDNs).
➢ To develop energy aware scheduling technique for optimum resource, time and cost utilization.

1.7. ISSUES ADDRESSED AND THEIR APPROACH

The resource allocation issues addressed in our work can be classified based on two broad approaches:

1.7.1. System Parameter based Approach

System based parameters like temperature based trigger is used for VM Migration and Network File System (NFS) based load balancing is done by migrating loads on idle VMs. The need to migrate VMs arises because of the overloading of physical machine that runs multiple VMs. Consequently, there is overheating which may degrade the performance and may also lead to faulty operations and system crashes. Hence a conditional load balancing is required for better manageability of the cluster of servers.

A two-fold technique for optimizing system performance is proposed: First, A Trigger based VM Migration technique that gets activated when CPU temperature increases beyond an upper threshold value, called Hotspot. Based on the Hotspot threshold, a VM can be live migrated to another available physical machine identified based on best threshold. Second, NFS based dynamic load balancing strategy is proposed for better system resource utilization. This is achieved by selecting the most suitable VM for load allocation. This method computes the load on various VMs and then finds the VM which is most suitable for the upcoming load. A dynamic load balancing algorithm is developed and implemented on Xen VMM for VM load balancing based on NFS that is efficient and requires less computational overhead as compared to normally used Migration strategies and other traditionally adopted load balancing techniques. This results in better resource utilization.

1.7.2. Quality of Service (QoS) based Approach

Three broad approaches are used for improving the Quality of Service. These are Trust Aware, Fault Aware and Energy Aware approaches:
• **Trust Aware**: A trust based model is proposed to address the above issues by VMM features that vary from datacenter to datacenter. Further, these trust values are used in load balancing and Scheduling algorithms to improve the QoS for end users with better resource utilization. Additionally, a QoS aware honey bee algorithm is advanced to provide higher QoS to the user and at the same time cost minimization parameter has been taken into consideration to facilitate user in his decision making process for choice of service. A trust based framework and algorithm is proposed which forwards the requests to trusted data centers and tries to avoid untrusted data centers. A Trust Aware Cloud framework VIMCLOUD based on trust values of data centers has been suggested. Trust value is calculated for each data center. Incoming requests are allocated to data centers having higher trust value. The proposed trusted cloud service framework ensures that user execute only on cloud data centers that run trusted VMs having known integrity and reliability in terms of trust value. The performance of VIMCLOUD is better than existing models as evident from the results observed. Further, a QoS aware Honey Bee scheduling algorithm is proposed for cloud IaaS. In addition to existing parameters, i.e., VM initiation time, price per hour usage, VM processing speed and bandwidth, fault rate of data centers is also taken into consideration which leverages significant improvement in allocating the requests to trusted data centers and at the same time the number of failed requests is reduced.

• **Fault Aware**: A fault aware load balancing algorithm for Content Delivery Networks (CDNs) is suggested that improves both the QoS and reliability. The effect of network failure on QoS and reliability of the system is studied in presence of high request rate and network traffic. Performance of existing load balancing algorithm is investigated in the presence of faults and is compared with the suggested algorithm. The performance of the suggested algorithm is also compared with other existing algorithms used in CDNs. Performance of the proposed algorithm is compared with existing techniques, i.e., Queue Length based Load balancing algorithm (QLBLBA), Random Algorithm (RAND), 2RC algorithm, Least Loaded (LL) algorithm and Round Robin (RR) algorithm. The experimental results demonstrate that the proposed algorithm provides better robustness and resilience to fault without affecting the QoS. The experimental results establish that suggested algorithm provides better robustness and resilience to fault without affecting QoS. Further, a dynamic fault model is proposed and
implemented which takes care of changing failure probability with load and provides better result as compared to the existing static fault aware models. A Fault Aware algorithm is proposed to reduce request failures.

- **Energy Aware**: Combination of energy aware, makesapn time and processing cost based algorithms are optimally used for better resource, time and cost utilization.

Energy aware scheduling techniques for Cloud environment are proposed. Software based scheduling and testing is implemented with DVFS for minimizing the processing cost, makespan time in Energy Aware environment. Simulation was carried out using CloudSim with combination of various QoS parameters e.g., processing cost, makespan time along with energy aware VM allocation policies. A comparison of these algorithms is given with the currently used energy aware algorithms. All above mentioned algorithms and techniques significantly contribute towards energy savings, but the Quality of Service parameters were neglected. A software based scheduling can be done with Dynamic Voltage and Frequency Scaling (DVFS) technique [42]. This minimizes the processing cost, makespan time in Energy efficient way. Energy can be saved without compromising the QoS i.e., it also minimizes the processing cost and the makespan time. Simulations were carried out using CloudSim for Quality Of Service (QoS) parameters alongwith the combinations of energy aware VM allocation policies. A comparison of these algorithms is shown with the normally used existing algorithms based on the processing cost, makespan time and Energy Utilization parameters that show better results. Software based scheduling and testing is done with DVFS based experiments for minimizing the processing cost, makespan time in Energy Aware environment so that in addition to energy saving, the QoS is not compromised. Simulations are done using CloudSim with combinations of various QoS parameters along with the combinations of energy aware VM allocation policies. A comparison of First Come First Serve (FCFS), Min-min, Max-min, Minimum Completion Time (MCT), Data Aware and Round Robin algorithm is shown based on the Processing Cost, Makespan Time and Energy Utilization parameters. A combination of Max-Min scheduling algorithm for cloudlet or task scheduling with Minimum Used Host scheduling algorithm for VM allocation gives the most efficient environment in terms of processing cost, makespan time and energy consumption maintain the QoS.
1.8. EXPERIMENTS/SIMULATION DETAILS

The experimental and simulations details carried out are as follows:

- VM migration is done based on the hotspot and coldspot thresholds. A trigger is initialized and appropriate VM based on above criteria is selected for migration causes the machine to fire the trigger and hence migration. We restricted the cumulative CPU usage to certain maximum cutoff for CPU temperature to $50^\circ C$ implying “Over load”, and similarly, for minimum cut off for CPU temperature say $10^\circ C$ implying “Under load”. The ASHRAE (American Society for Heating, Refrigerating and Air-Conditioning Engineers) standard cold spot for a typical data server is 64.4 degree Fahrenheit or 18 degree Celsius. And hotspot depends on the quality of hardware use but generally 85+ degrees Celsius is considered highly critical. Additionally, a dynamic load balancing strategy and algorithm is developed and implemented on Xen VMM for VM load balancing based on NFS.

- Our experimental work mainly comprises of implementation and analysis load balancing strategy. A comparative study of system performance in three conditions are noted, i.e., normal physical executions without using VMs, VM based execution without using load balancing technique and VM based execution using load balancing technique is made. Three types of executions are performed, namely, system running on a single operating system, system in which Xen hypervisor is implemented without using load balancing and system with Xen hypervisor installed and load balancing is done. When a single operating system is installed, the performance of system decreases when load increases, while if a hypervisor is installed then initially the performance is lower than native performance due to certain overheads, but as load increases performance becomes better. Performance is further improved when load balancer is used.

- A trust based cloud framework VIMLCOUD is proposed. It guarantees that the trusted user will get high QoS by running their VM on trusted datacenters with high reliability. A honey bee cost efficient algorithm is proposed which performs better than the existing scheduling algorithms. CloudSim is used for simulation of cloud IaaS. Honey bee scheduling algorithm is implemented in CloudSim. By default, CloudSim does not have the support for finding the
fault rate at datacenter, so the fault rate parameter of datacenter is added which considers the faults occurring at the datacenter while allocating the requests to the VMs.

- A fault aware algorithm for balancing the load is proposed which performs better than existing load balancing algorithms that are generally used for CDNs in fault aware environment. The proposed algorithm proves to have a lower failure count and failure probability as compared to existing QLBLBA, RAND, 2RC, RR and LL algorithms. Additionally, the average queue lengths and average failure counts for the proposed and above mentioned existing algorithms are compared. Proposed algorithm shows better performance in term of average queue length and average failure count.

- In Energy Aware environment, based on QoS scheduling technique, it is found that combination of Min-Min and Min Used Host is the most effective in terms of Cost and Makespan Time but combination of Max-Min and Min Used Host is most energy efficient. From the results, we also found that when Min-Min gives least cost and time then Max-Min was second most efficient with a difference of approximately 0.1% and when Max-Min gives most energy efficient environment, Min-Min consumes maximum energy with a difference of approximately 3%. Therefore, if we use the combination of Max-Min scheduling algorithm for cloudlet or task scheduling and Minimum Used Host scheduling algorithm for VM allocation, then we get the most efficient results in terms of Processing Cost, Makespan Time and Energy Consumption.

1.9. ORGANIZATION OF THE THESIS

The thesis is organized in 7 chapters.

**Chapter 1** introduction and **Chapter 2** gives review of relevant literature.

**Chapter 3** presents a Trigger based VM migration technique from a Hot CPU to an available normal CPU based upon threshold value of a temperature termed as Hotspot. A dynamic NFS based load balancing technique has been advanced for efficient resource utilization. User requests are allocated to VMs so as to optimally balance the load on VMs.

In **chapter 4**, an open source trust based framework has been proposed. Trust based scheduling and load balancing features can be implemented and tested using this framework. The proposed trusted Cloud framework supports IaaS, named VIMCLOUD. It ensures high
QoS to users by running their VMs on trusted data centers. The proposed Honey Bee scheduling algorithm considers fault rate of data centers in addition to normal QoS parameters resulting into significant improvement in cloud IaaS services.

**Chapter 5** presents a Fault Aware Load Balancing algorithm for Content Delivery Networks (CDNs). The proposed Load Balancing Algorithm provides improved QoS and reliability of service request allocations even in presence of high network traffic as evidenced by low rate of request failures. A dynamic fault model has been proposed and implemented which takes into account dynamically varying failure probability with load. Use of this model gives better results in comparison to existing static fault models. Simulation studies of load balancing using the proposed algorithm for Content Delivery Networks has been carried out and results presented in the chapter.

In **Chapter 6**, Energy aware software based scheduling policy is proposed. This policy results in reduction of processing cost and makespan time. It makes use of Dynamic Voltage Frequency Scaling (DVFS) and has been simulated using CloudSim. The performance of the algorithm has been compared with existing algorithm i.e., First Come First Serve (FCFS), Min-min, Max-min, Minimum Completion Time (MCT), Data Aware and Round Robin (RR) algorithm on the basis of processing cost, makespan time and energy utilization.

**Chapter 7** gives the contributions made by this work. It also gives directions for possible future work in the area of this thesis.