**CHAPTER 1**

**INTRODUCTION**

**1.1 OVERVIEW**

Cloud computing and energy efficiency [16] are two important trends in the IT industry at the moment. The paradigm of cloud computing relieves users from the burden of managing their own information technology infrastructure. Third party clouds enable organizations to focus on core business aspects instead of relying on complex computer infrastructure and maintenance. In recent years, more and more Internet services (e.g., e-commerce, content distribution, gaming, social networking) have been deployed over the cloud infrastructure. The cloud infrastructure provides reliable, elastic and cost effective services. It enables ubiquitous access to shared resources. Cloud computing is the result of the evolution and adoption of existing technologies and paradigms. Cloud computing provides enterprises and users with the computing capabilities to store and process data on-the-go. The shared resources can be provisioned with minimal effort over the Internet. A CSP normally operates multiple geo-distributed datacenters, which are located in different regions which provides cloud services to users from different regions. User requests can be routed to any datacenter via the intermediate load balancers.

**1.2 CLOUD COMPUTING**

Cloud computing is an Information Technology (IT) paradigm that uses a network of remote servers hosted on the Internet to provide ubiquitous access to shared resources. It provides provisioning and network access to a shared pool of physical or virtual resources like computing power, storage, applications and other information technology resources on demand. It also enables sharing of higher level services with minimal management effort, thus providing elasticity and broad network access to multiple users. Cloud computing provides a method for consumers to take advantage of the cloud platform without having deep knowledge or expertise about it. Cloud users can focus on core business aspects instead of information technology obstacles. Improved management of resources and lesser maintenance enable enterprises to minimize infrastructure costs and adjust resources to meet highly dynamic and wildly fluctuating user demand. It also provides small organizations with the computing power and resources that are generally only available to larger enterprises. Cloud computing is largely based on Service Oriented Architecture (SOA) which helps to provide integrated services as solutions. Cloud computing is mainly enabled by the concept of Virtualization. Cloud computing adopts its many aspects from other models like Client-Server model , Grid computing ,Fog computing , Green computing , Peer-Peer computing.

**1.3 CLOUD COMPUTING CHARACTERISTICS**

Cloud computing displays the characteristics of Reliability, Productivity, Security, Multitenancy, Resource Pooling , Device and Location Independence , Scalability and Elasticity. Cloud platforms are typically made up of multiple sites which store redundant information for better disaster recovery. Cloud computing also enables accessing information stored on remote servers using any device regardless of the location the user is in. Sharing of resources with multiple users enables improving utilization and efficiency for underutilized systems. Further, since data is stored in remote servers, they can be accessed simultaneously, thus improving productivity. Cloud computing can help to provide improved security to data in spite of existing security concerns, since Cloud Service Providers (CSPs) can devote more resources to secure data than enterprises would. Scalability and Elasticity ensure that resources can be scaled up or down based on usage.

**1.4 CLOUD COMPUTING SERVICE MODELS**

Cloud Service Providers (CSPs) provide cloud computing services based on three different models which are : infrastructure , platform ,software.

**1.4.1. Infrastructure as a Service (IaaS)**

IaaS service providers enable consumers to use various high level APIs to access low-level details of underlying network infrastructure. Network infrastructure like physical computing resources, data partitioning, scaling, etc ,. Hypervisors or Virtual Machine monitors can be used to run VMs as guests and provide ability to scale services up or down. IaaS is a cloud platform service wherein the user can deploy arbitrary software, applications or operating systems and manage them without controlling the underlying cloud infrastructure.

**1.4.2. Platform as a Service (PaaS)**

PaaS vendors offer consumers a development environment on top of their cloud infrastructure. Vendors provide toolkits and standards for distribution and payment to the consumers which typically are operating systems, programming language execution environment, database, web servers , etc.,. Cloud users can develop their customized applications using the development environment or deploy their existing applications on top of the cloud infrastructure. Cloud providers enable computer and storage resources to scale automatically based on application demand so that users need not scale them manually.

**1.4.3. Software as a Service (SaaS)**

SaaS vendors provide cloud users the capability to use the provider’s applications running on top of the cloud infrastructure. Cloud users can access the applications through thin client interfaces such as browsers or through program interfaces. Cloud infrastructure such as network, servers, operating systems, storage cannot be controlled by the user except user-specific application configuration settings. SaaS software is also called “on-demand software” and usually accessed by a subscription fee or pay-per-use basis.

**1.5 CLOUD COMPUTING ARCHITECTURE**

The cloud architecture typically consists of the following components which are shown in figure 1.1 :

* Front end platform (client computing devices)
* Back end platform (servers, storage)
* Cloud based delivery
* Network (Internet, Intranet, Intercloud)

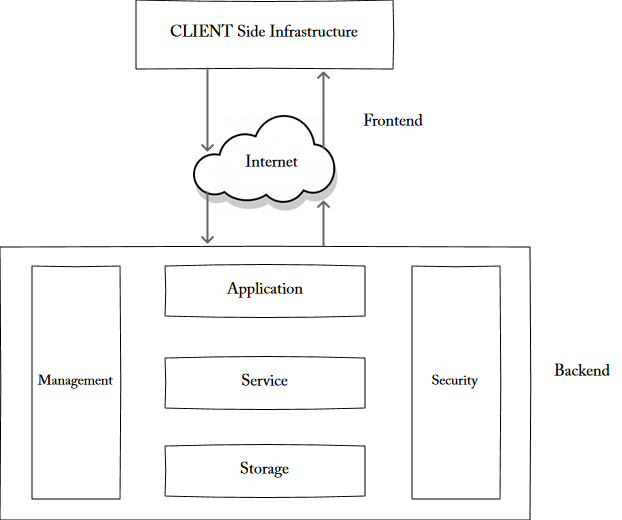


Figure 1.1 – Cloud Computing Architecture

The front end platform of cloud architecture consists of client side computing devices like thin / thick client devices, smart phones, tablets, and other computing devices. The client computing devices access cloud services using browser, virtual session or other middleware.

The back end platform of cloud architecture consists of remote servers and storage which form the basis for cloud computing. The cloud storage is accessible by multiple clients who can interact through a middleware. It is agile, reliable, flexible and scalable. Cloud storage is generally deployed in one of the four configurations:

* Private Cloud
* Public Cloud
* Community Cloud
* Hybrid Cloud

Cloud based delivery is captured in one of the three configurations:

* Infrastructure as a Service (IaaS)
* Platform as a Service (PaaS)
* Software as a Service (SaaS)

The cloud network layer provides high bandwidth data and application access, agile on-demand access to data which can freely move between different servers or clouds and the much needed network security which is used to achieve isolation and secure access to data and applications.

**1.6 VIRTUALIZATION**

Virtualization [5] involves sharing of a physical instance of resource among multiple users to better serve user requests. Virtualization is the enabling technology for cloud computing. It transforms a single physical “host” device into multiple “guest” machines/devices that can be easily managed and controlled. Using virtualization, multiple independent computing devices can be created which can be easily scaled and which can efficiently utilize idle computing resources. Virtualization enables increased cloud flexibility and security. The hypervisor or virtual machine monitor is responsible for managing the virtual machine running on the physical machine. Virtualization can help to speed up operations and increases utilization. Virtualization is of different types namely hardware virtualization, operating system virtualization, server and desktop virtualization.

**1.7 RENEWABLE ENERGY IN CLOUD COMPUTING**

Traditionally, cloud datacenters are powered by electricity generated from fossil fuels (brown energy). With the increase in cloud datacenters due to the growth of cloud computing, power consumption and carbon emission has increased tremendously. Large scale CSPs consume millions of megawatts of power which translate to about 1.5 percent of the world electricity. In order to reduce the carbon footprint, datacenters have started migrating towards green cloud datacenters which use renewable energy [30]. To reduce the carbon footprint, both renewable energy as well as energy efficiency which enables electricity to be used effectively, in cloud datacenters is required. Renewable energy usually involves solar and wind energy which typically translate to 15-20% electricity. Energy efficiency can be achieved by partitioning workloads to reduce overloaded servers (load balancing) and sleeping, rate adaptation of servers.

**1.8 LOAD BALANCING IN CLOUD COMPUTING**

Load balancing [12] is one of the most important issues in cloud computing. Cloud Load balancing involves distributing workloads among multiple hosts or computing devices. Cloud load balancing is advantageous in that it can transfer loads globally instead of just local servers. Cloud load balancing also routes user requests to the closest regional server in case of a server outage. Load balancing also improves response times by routing requests to the best performing data centers and provides for better resource utilization and performance of the system by balancing loads across hosts. Several methods have been proposed for load balancing which involve nature inspired algorithms [27][20], dynamic methodologies [7], clustering [15], energy aware scheduling [16],etc.,.

**1.9 LYAPUNOV OPTIMIZATION**

Lyapunov optimization [23] is generally used to achieve optimal control in queuing networks with the use of a Lyapunov function. Lyapunov functions are scalar functions used to achieve stability / equilibrium in dynamic systems. Lyapunov functions yield a scalar measure that can be used to determine the level of system stability as the scalar value increases as the system moves towards non-equilibrium states. The Lyapunov function takes a multidimensional input vector (factors of the system) as state variable. This is used to evaluate the function which depends only on the current system state thus achieving sub-optimality.

Lyapunov drift is used to stabilize the queuing network by minimizing the time average network penalty function. This is achieved by defining a set of virtual queues which is used to determine the lyapunov drift.

**1.9.1 DEFINITION OF LYAPUNOV FUNCTION**

Let **V : Rn --> R** be a continuous scalar function.

**V** is a Lyapunov-candidate-function if it is a locally positive-definite function, i.e.

* **V(0)=0**
* **V(x)>0** for every **x** belongs to **U \{0}**

Where,

**U** - neighborhood region around **x=0**.

**1.10 NATURE INSPIRED ALGORITHMS**

Nature inspired algorithms [27] are Meta heuristic algorithms designed to optimize multi objective functions and solve NP-hard problems for large number of variables, dimensions, etc.

Nature inspired algorithms are classified into evolutionary and swarm intelligence based algorithms. These algorithms use recombination and mutation operators to optimize the complex problems. Nature inspired algorithms help to overcome the curse of dimensionality by optimizing search for an optimal solution on a potentially infinite search space. A wide variety of nature inspired and swarm optimization algorithms are applied to cloud load balancing. Since cloud load balancing is an NP-Complete problem, nature inspired algorithms can be applied to achieve near–optimal solution.

**1.10.1 APPLICATION OF NATURE INSPIRED ALGORITHMS**

Swarm intelligence algorithms help to model the collective behavior of natural swarms. This can be used to optimize the number of task migrations by making the submitted / migrated cloudlets pass information to other cloudlets about the VM/ Datacenter. The predefined migration thresholds can be used to determine migrations of cloudlets.

**1.10.2 BACTERIAL FORAGING OPTIMIZATION ALGORITHM**

The key idea of Bacterial Foraging Optimization Algorithm [28] is mimicking chemotactic movement of virtual bacteria in the problem search space. The process, in which a bacterium moves by taking small steps while searching for nutrients, is called chemotaxis. The bacteria perceive the chemical gradients in the environment and moves away or towards some specific signals. It is applied to monitor balance of loads across hosts.

The algorithm consists of 3 stages:

* Chemotaxis
* Reproduction
* Elimination Dispersal

**PARAMETERS OF BFOA**

* **S** – Total number of bacteria in the population
* **Nc** - Number of chemotactic steps
* **Ns** - Number of swim steps
* **Ned** - Number of elimination dispersal steps
* **Ped** - Probability of elimination dispersal
* **J(i,j,k)** – Cost of ith bacterium in jth chemotactic and kth elimination dispersal steps.

**1.10.3 MODIFIED BACTERIAL FORAGING OPTIMIZATION ALGORITHM**

Since the reproduction step eliminates one half of the bacterial population while reproducing only the “healthy” half of the population, it can result in certain bacteria losing their optima since they are assigned only the values pertaining to the healthy half of the population. Hence, in modified bacterial foraging optimization algorithm, the reproduction step is eliminated and the bacteria are allowed to retain their positions which may be altered based on their probability in the elimination dispersal step. This way, convergence can also be achieved since the bacteria do not stay in their positions for too long and are dispersed over the search space.

**1.11OBJECTIVE**

The objectives of the project are to develop a dynamic load balancing methodology that can be deployed to balance loads across hosts in the system and to develop a power management methodology that can be used to effectively integrate renewable (green) energy with brown energy. This is used to manage the time-average eco-aware power cost of green cloud datacenters while still maintaining user QoS(Quality of Service) constraint. User requests would be routed to optimal datacenters by intermediate load balancers using modified bacterial foraging based load balancing algorithm. This is used to reduce the finish times of cloudlets while optimizing the number of migrations and operating power of datacenters.

**1.12 PROBLEM DEFINITION**

The problem of load balancing and power management with renewables is modeled as a stochastic optimization problem with the following aspects:

* The problem considers N geo distributed datacenters with user requests routed to any datacenter.
* All hosts are homogeneous and capable of servicing any request.
* The variation of operating power of the datacenter with time is modeled as a stochastic process that is optimized by Lyapunov optimization framework to reduce the time average power cost.
* Power management methods are devised to integrate renewables to the existing power supply from the grid and also to reduce the operating cost for a datacenter.
* Load balancing also influences the amount of power consumed at a datacenter. Hence optimal load balancing can reduce the number of task migration since they increase the operating power of a datacenter.

**1.13 ORGANIZATION**

The following chapters in this thesis are organized as follows:

* Chapter 2 discusses about the literature survey undertaken for the project.
* Chapter 3 explains the proposed system and the various techniques involved in the proposed approach.
* Chapter 4 explains the Implementation and result discussion.
* Chapter 5 provides the conclusion.
* Chapter 6 summarizes the references.