CHAPTER 1

INTRODUCTION

1.1 CLOUD COMPUTING

Cloud Computing has been rapidly and radically changing the dynamics of Information Technology (IT) consumption. Cloud computing is an emerging technology that enables the user to avail the IT resources on pay for use basis. It facilitates the user to access the IT resources, without spending on capital expenses, through the internet in an on-demand manner. The core of cloud computing proposition is being able to buy compute technology as a service in an on-demand, elastic, and pay-as-you-go model. The National Institute of Standards and Technology (NIST) definition of cloud computing is that “Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources such as networks, servers, storage, applications, and services that can be rapidly provisioned and released with minimal management effort or service provider interaction” (Mell & Glance 2011). The characteristics of cloud computing as listed by NIST are on-demand self-service, broad network access, resource pooling, rapid elasticity, and measured service.
1.2 DEPLOYMENT MODELS OF CLOUD ENVIRONMENT

The adoption and deployment of cloud computing platforms have many attractive benefits such as reliability, quality of service, and robustness (Zhang et al. 2013). Cloud can be classified into four categories, such as public, private, community, and hybrid.

1.2.1 Public Cloud

As the name suggests, public cloud is something that is probably out of premises managed by a third-party service provider, which is made available to the 'public'. Anybody can use a public cloud, be it a small to medium sized organization or a big organization. It emphasizes the pay-as-you-go model which implies that the enterprises have to pay only for those services that had been utilized by them, which is usually in the form of a monthly bill that will be generated by the service provider. Public cloud is envisioned to provide boundless computational resources via the internet to all enterprises. One of the key features of public cloud is that it's moldable setting that lets the users configure their resources and also to choose what to hide and what not to hide from their clients. Most prominent instances of public cloud are Amazon Elastic Compute Cloud (EC2), IBM's Blue Cloud, Sun Cloud, Google App Engine and Windows Azure Services Platform.

1.2.2 Private Cloud

Private cloud is a model of cloud computing that evolves with a definite and shielded cloud setup where only the stated customer can benefit. It is a datacenter model, which is engineered in a way such that it can be inherited by a single enterprise. A private cloud provides elasticity, scalability,
better computerization, and supervision. The objective of a private cloud is to utilize the offerings of the cloud architecture without giving up the security and control over the sensitive data. A private cloud is similar to that of a public cloud, but it discards the exception over the control on the data. Private cloud is expensive when compared to the public version, which makes it unavailable for a lot of small and medium scale industries. They are more stringent when it comes to security and acquiescence, and keeping resources within the firewall. The resources of the private cloud are always drawn from distinct pool servers which can be hosted either internally or externally, but can be tapped from either a private connection or via an encrypted public connection, which strikes a contrast to the public cloud providers where all their resources lie in a shared pool of servers. Moving data into the private cloud offers flexibility and convenience to people since they do not have to fear about third-party data management’s trustworthiness (Mell & Glance 2011). Creating these private clouds enable IT departments to focus on innovation for the business, reducing both capital and operational costs and automating the management of complex technologies (Devi et al. 2010). Nowadays, the private cloud is getting popular e.g. virtual classroom, virtual desktop to make better utilization of resources. Most prominent instances of private cloud are OpenStack, Ubuntu Enterprise Cloud (UEC), CloudStack, and OpenNebula.

1.2.3 Community Cloud

Community cloud infrastructure is shared by several organizations and supports a specific community that has shared concerns such as mission, security requirements, policy, and compliance considerations. It may be managed by the organizations or a third party and may exist on premise or off premise.
1.2.4 Hybrid Cloud

Hybrid cloud infrastructure is a composition of two or more clouds such as private, community, or public that remains unique entities but is bound together by standardized or proprietary technology that enables data and application portability. Hybrid cloud is an infused concept of both public and private cloud setup in order to carry out versatile tasks within the same enterprise. It is an environment where an organization handles few of its data in-house, while outsourcing rest of its data via the public providers. For instance, an enterprise can use Amazon’s elastic cloud to manage its archived data while it maintains its confidential data like customer information in-house with the help of a private cloud. This approach lets the consumers enjoy the benefits of scalability and cost-effectiveness from a public vendor without compromising on the critical information to the third-party vulnerabilities. This method proves to be cost-efficient than that of the private cloud set-up.

1.3 SERVICES OFFERED BY CLOUD COMPUTING

Cloud computing is available majorly in three service models. The service models of cloud computing include Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS) (Mell & Glance 2011).

A service is defined as fine-grained reusable resources such as infrastructure or business processes available from a service provider; this is now what is popularly called —as a service.
1.3.1 Software as a Service

Software as a Service (SaaS) is a kind of application that is available as a service to users; it delivers software as a service over the internet, eliminating the need to install and run the application on local computers in order to simplify the maintenance and support. The benefits of SaaS are lower cost, user familiarity with web availability and reliability. SaaS provides software that is deployed over the internet. With SaaS, a provider licenses an application to customers either as a service on demand, through a subscription, in a “pay-as-you-go” model, or at no charge when there is an opportunity to generate revenue from streams other than the user, such as from advertisement or user list sales. SaaS is a rapidly growing market as indicated in recent reports that predict ongoing double-digit growth (Cooter 2010 & Williams 2010).

1.3.2 Platform as a Service

Platform as a Service (PaaS) model enables the deployment of applications without the cost and complexity of buying and managing the underlying hardware and software layers. A customer can deploy an application directly on the cloud infrastructure using the programming languages and tools supported by a provider without managing and controlling that infrastructure. A customer has the control over its applications and hosting environment’s configurations. PaaS brings the benefits that SaaS bought for applications, but over to the software development world. PaaS can be defined as a computing platform that allows the creation of web applications quickly and easily and without the complexity of buying and maintaining the software and infrastructure underneath it (Rackspace 2011).
1.3.3 Infrastructure as a Service

*Infrastructure as a Service (IaaS)* provides the virtualized computing resources to the users which can be adjusted according to the user’s claim. Instead of purchasing the hardware outright, users can purchase the computing resources based on consumption like electricity. IaaS delivers computer infrastructures like processing power, storage capacity, and network to customers; instead of building datacenters, purchasing servers, software or network equipment, a customer buys the resources as a fully out-sourced service; a customer does not manage the underlying infrastructure but has full control over the operating systems and the applications running on it. IaaS models often provide automatic support for on-demand scalability of computing and storage resources (RackSpace 2011). The proposed research work focuses on the IaaS cloud which facilitates the customers to work with the servers in a virtualized environment.

![Service Models of Cloud Computing](image)

**Figure 1.1** Service Models of Cloud Computing
Figure 1.1 shows the major three service models offered by the cloud computing. From the Figure 1.1, it is observed that the Infrastructure as a Service is the foundation for the other two layers. Table 1.1 lists some of the major players in each service model.

Table 1.1 Some of the Major Players at Each Service Model

<table>
<thead>
<tr>
<th>Layers</th>
<th>Major players</th>
</tr>
</thead>
<tbody>
<tr>
<td>SaaS</td>
<td>Gmail, Google Docs, Salesforce, Dropbox</td>
</tr>
<tr>
<td>PaaS</td>
<td>Google App Engine, Windows Azure, Force.com</td>
</tr>
<tr>
<td>IaaS</td>
<td>Amazon EC2, Rackspace, Eucalyptus, OpenNebula, Nimbus, OpenStack etc.</td>
</tr>
</tbody>
</table>

1.4 CONCEPT OF VIRTUALIZATION

Cloud computing is the extension of grid computing, parallel computing, and distributed computing. In contrast with the above-mentioned technologies, virtualization is the key concept of cloud computing by which the Central Processing Unit (CPU) cycles and bytes are provisioned as a commodity. With the help of virtualization, the users can avail any services from the cloud resources through the internet. 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 of the datacenter at the provider’s site.

Virtualization is a technical innovation designed to increase the level of system abstraction and enable IT users to harness ever-increasing
levels of computer performance (Buyya et al. 2008). Cloud provides rapid provisioning of virtual machines on the physical machines and allows scaling up and down quickly with the help of virtualization. Different kinds of servers can be run on the same platform known as server consolidation, making the better utilization of resources. Thanks to virtualization and consolidation techniques, applications can be embedded in virtual machines and deployed on demand on a reduced number of nodes (Ristenpart et al. 2009). The advantage of virtualization as stated in (Devi et al. 2011) is its consolidation of workloads to reduce hardware, power, and space requirements.

Figure 1.2 Concept of Virtualization

Figure 1.2 shows the concept of virtualization. As depicted in Figure 1.2 virtualization introduces two level schedulers, one from the virtual machine monitor, where virtual CPUs are mapped to the physical CPUs and the other from the guest operating system, where the tasks are scheduled to
virtual CPUs. Some of the key virtualization technologies are Kernel Virtual Machine (KVM), Xen, Microsoft Hyper-V, VMware ESX, etc.

1.5 VIRTUAL MACHINE PLACEMENT IN CLOUD ENVIRONMENT

Cloud framework changes the traditional computational infrastructure to virtual infrastructure. In the cloud, the services are not tied up with the physical resources. Virtual machine provisioning manages the virtual infrastructure with ease, offers these resources to the cloud users based on their claim. Since virtualization is the core technology for cloud computing, virtual machine placement becomes a hot topic. It processes the virtual machine request from cloud users and places the Virtual Machines (VM) to appropriate hosts (also referred to as a Physical Machine or PM).

Figure 1.3 Virtual Machine Provisioning
Figure 1.3 shows the process of virtual machine provisioning. For launching the virtual machine, provisioning services selects the proper server in the cloud datacenter which satisfies the resource requirements as well as any specific constraints specified by the placement policies like reduced power consumption, better resource utilization, etc. This problem of mapping of VMs to PMs is one of the most challenging problems in cloud computing environment (Pedram & Hwang 2010) (Pelleg & Moore 2000) (Pelley et al. 2009), while the virtualization at the operating system such as VMware (Nez et al. 2012) level is entering the mainstream. The challenge is, therefore, to provide an efficient approach to analyzing the mapping problem by taking its entire complex behavior into account.

This virtual machine placement problem in the cloud environment is a combinatorial NP-hard problem since it involves enormously huge search space with a respectively large number of possible solutions. An NP-hard problem means non-deterministic polynomial-time hard which cannot be solved to optimality within polynomially bounded computation time. The solution space for the scheduling problem is so large, thus, it takes a long time to find out the optimal solution. In the cloud environment, it is preferable to find a suboptimal solution in a short period of time. Meta-heuristics based techniques have been proved to achieve near optimal solutions within a reasonable time for solving NP-hard problem which is very hard to resolve.

Particle Swarm Optimization (PSO) is a meta-heuristic algorithm as it makes few or no assumptions about the problem being optimized and can search very large spaces of candidate solutions. In PSO, the particle represents the ‘birds’ or ‘bees’ or any individuals who manifest the social behavior and interaction with each other in a group. It was originally designed by Kennady & Eberhart (1995). Each particle is associated with the position and velocity; proceed through the multidimensional search space. The velocity of a particle
is updated in each iteration using the best position of the particle and the best position of the swarm. Each particle is called as a candidate solution. The candidate solutions are flown in the problem space to find out the optimal solution. The candidate solutions are changing its position according to its own experience and the experience of its neighbors. There are two important terms in PSO, the particle’s position: where the particle is compared with other particles in the search space. The particle’s velocity: how fast the particle moves from one position to other and in what direction. The other term is the fitness value of the particles, which can be used to find the optimal solution.

As virtual machine mapping is a combinatorial optimization problem, it can be solved using swarm intelligence methods. This research work proceeds to apply meta-heuristic approach Enhanced Particle Swarm Optimization (EPSO) for VM to PM mapping with the objective of minimizing power consumption and resource wastage of the datacenter.

1.6 VIRTUAL MACHINE CONSOLIDATION IN CLOUD ENVIRONMENT

With the help of virtualization, a server can be configured to host a multitude of VMs. Virtualization helps to scale up and down the resource capabilities of the VM and also helps to consolidate servers by selective migration of VMs between hosts. This consolidation approach reduces the number of hosts by either switching off or transitioning to low power sleep state then un-used, thus reducing the overall power consumption of the datacenter. Majorly, only a fraction of computer machine’s potential power is used and many system runs with low average system load. Many hardware resources and power are exhausted. Therefore, instead of physical computers or hosts that are only partially used, in a few powerful hosts, many VMs can be packed by balancing them. During server consolidation, all the VMs
running on various under-utilized servers are migrated through live VM migration approach and all the servers that are unused will be set to power-saving state. Live VM migration ensures that the service downtime is minimized as well as the period of migration.

Datacenters are recently receiving considerable attention on designing in a power-efficient manner. This approach leads to several directions. For example, slowing down of CPU speeds and turning off partial hardware components to improve power efficiency through hardware architecture has become common practice. Server consolidation and power-aware VM mapping are two solutions to minimize power consumption by setting the unused machines to power saving state (Mohan & Shriram 2016). In the proposed research work, dynamic consolidation of servers is implemented in cloud environment with the objective of minimizing the power consumption, resource wastage and number of VM migrations in the datacenters.

1.7 MOTIVATION

The users access the computing resources by creating virtual machines on the physical machines. Virtual machines are an abstract unit of computation in the cloud. While mapping the virtual machines efforts should be made to utilize the resources efficiently for the benefit of the providers. From the provider’s point of view, the major concern is to maximize profits by minimizing the operational cost. In this regard, minimizing the power consumption and resource wastage in cloud datacenters becomes a crucial issue as it involves in the operating cost.

This research work focuses on the IaaS cloud which facilitates the cloud users to work with the computational resources in the virtualized
servers. Apart from the benefits gained by the cloud computing, the pitfall is a huge amount of power consumed by datacenters which majorly contributes to its operational cost. For example, in 2004 to dissipate 1W consumed by a High-Performance Computing (HPC) system at the Lawrence Livermore National Laboratory (LLNL), 0.7Watts of additional power is needed for the cooling system. 2004’s top ranking machine, the Japanese Simulator, required 12Mega Watts of power to operate, roughly the amount required to power a small town (Petrini et al. 2004).

Electricity costs for powering servers form a major cost of operation in the datacenter and it has been estimated that in near future, operational expense over a period may be even more than the capital expense (Mohan & Shriram 2016). According to Hamilton (2013), Vice President, Cloud Computing Services, Amazon, the actual cost of power consumed by the servers plus the cost of cooling the servers in a year is 34% of the total cost of ownership of a datacenter, whereas the amortized server costs in a 10-year lifetime of a datacenter is 54% of the total cost. This implies that if concentrating on getting maximum throughput from servers by consolidation, it will be able to reduce datacenter power consumption and cooling costs and that will be the biggest savings for a datacenter operator. For many organizations, the cooling and power costs of running a server for two years are equivalent to the price of its purchase. Due to the environmental and more important economical gains, researchers, as well as leading entrepreneurs, pioneers, and technological companies, have begun work targeting different aspects of the power consumption problem.

If the resources are not properly scheduled, there may be a chance for failure of virtual machine creation. Even though the resources are available in the datacenter, but scattered across the host results in failure of virtual machine creation. Most of the time servers operated at 10-50% of their full
capacity, leading to extra expenses on over-provisioning, and thus extra Total Cost of Acquisition (TCA) (Barroso & Holzle 2007). In this regard, minimizing the resource wastage in cloud datacenters becomes a crucial issue as it involves the operating cost.

Hence, this research work focuses on the development and evaluation of combinatorial optimization approach PSO to virtual machine placement and consolidation problem with the objective of minimizing the power consumption, resource wastage and number of VM migrations in the datacenters.

1.8 RESEARCH OBJECTIVES

Power consumption and resource wastage minimizations are the two major criteria concentrated in this research work. In particular, the following objectives are delineated.

- To explore, analyze and classify the research in the area of power-aware and resource-wastage aware VM allocation and dynamic consolidation to gain a systemic understanding of the existing techniques and approaches.
- The meta-heuristic algorithm EPSO is applied for solving the virtual machine placement problem in the cloud computing environment with the objective of minimizing the power consumption and resource wastage while satisfying the resource requirements constraints of the virtual machine.
- Power-aware dynamic consolidation is done. The algorithm performs dynamic rescheduling of VMs for making the cloud environment to be power efficient.
• When to switch off and on the physical machine is analysed. To save energy, dynamic consolidation should be combined with dynamic switching off the power states of the hosts for addressing the problem of eliminating the power consumption in the idle state.

• Selection of VM for migration is done. When servers are overloaded, migration of VMs should be performed on the physical machine. The problem is determining the subset of VMs to migrate that will be most beneficial for the system reconfiguration. So, resource wastage aware VM migration is proposed which minimizes the number of VM migrations.

The parameters considered for the proposed algorithms are power consumption of datacenters, resource wastage of datacenters, and the number of virtual machine migrations.

1.9 METHODOLOGY USED

The research methodology in this thesis consists of several consecutive steps which have been summarized below.

1. Theoretical analysis of virtual machine allocation and consolidation algorithms is conducted to obtain theoretical performance estimates and gain insights on designing such algorithms.

2. Over last few years, the field of metaheuristic and its ability in solving a complicated problem is one of the hot research areas. Hence, this research work takes up the present investigation on VM placement using the meta-heuristic approach.

3. Virtual machine allocation and consolidation algorithms are developed based on the insights from the conducted competitive analysis and
derived system model to minimize the power consumption and resource wastage of datacenters.

4. The proposed algorithms are evaluated through discrete-event simulation using the MATLAB toolkit for reducing the power and resource wastage of the datacenter. As the target system is an IaaS, it is necessary to create an infinite set of computing resources to the users, it is also important to evaluate the proposed algorithms on a large-scale virtualized datacenter infrastructure. Hence it is very difficult to conduct repeatable experiments on the large scale; simulations have been chosen as the initial way to evaluate the performance of the proposed algorithms.

5. In order to test the power consumed by servers and datacenters when running a full range of applications, Standard Performance Evaluation Corporation (SPEC) set up an industry-standard power-performance benchmark (Fourth Quarter SPEC Results 2009). The real data on power consumption for the hosts are taken from the results of the SPEC benchmark. The data for virtual machines is taken from the Amazon EC2 instances (2012).

1.10 CONTRIBUTION OF THE THESIS

The contribution of this research is summarized as follows:

1. A taxonomy and state of the art in the virtual machine allocation and dynamic consolidation are studied.

2. Formulation of power-aware VM mapping using Enhanced PSO. Power-efficiency of the host differs according to its CPU power-efficiency and speed. The proposed algorithm minimizes the power consumption of datacenters by efficient mapping of virtual machines. The performance of the proposed algorithm is compared with the
existing Self-Adaptive PSO in terms of power consumption of the datacenters.

3. Power-aware dynamic consolidation: The power management techniques such as system sleep states and dynamic consolidations are used in the proposed research work to prevent the system from consuming more power. The system’s states are taken as defined by Advanced Configuration and Power Interface (ACPI) specification (Ware et al. 2010). The proposed algorithm performs dynamic rescheduling of VMs for making the cloud environment to be power efficient and fulfilling the Service Level Agreement (SLAs). The performance of the proposed algorithm is compared with the existing Modified Best Fit Decreasing (MBFD) algorithm in terms of power consumption of the datacenters and number of VM migrations.

4. Resource wastage aware VM allocation: The proposed algorithm takes advantages of Enhanced PSO to minimize the resource wastage in the datacenters. The performance of the proposed algorithm is compared with the existing PSO algorithm in terms of resource wastage of the datacenter.

5. VM selection for migration: When the hosts are overloaded, VM migration is done for maintaining the SLAs. For doing the migration, the decision to choose the VM for migration is essential. The proposed algorithm minimizes the number of migrations by choosing the VM which contributes more in the resource wastage of the host. The performance of the proposed algorithm is compared with the existing Black-box algorithm in terms of number of VM migrations.

6. Managing Resources in Private Cloud: Chrome extension for the open source private cloud OpenStack is developed which can be utilized by the cloud administrators to manage the resources in the datacenter. It is an add-on to OpenStack Horizon dashboard for giving alerts message to
the administrator when there are over-utilization and under-utilization of resources in a host. The experiments are done in a multi-node test bed using OpenStack private cloud.

1.11 THESIS SCOPE

The scope of this thesis is summarized in Table 1.2.

<table>
<thead>
<tr>
<th>Target Systems</th>
<th>Heterogeneous IaaS Cloud</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Heterogeneous server cluster</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Virtualization</th>
<th>Virtualized datacenters</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Virtualized server cluster</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>System Resources</th>
<th>CPU</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RAM</td>
</tr>
<tr>
<td></td>
<td>Network bandwidth</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Goal</th>
<th>Improve energy efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Improve resource utilization</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Power saving techniques</th>
<th>Dynamic Power Management (DPM) with sleep state transitions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dynamic VM Consolidation</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Workload</th>
<th>Spec 2006 Benchmark for hosts</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Amazon EC2 Instances for virtual machines</td>
</tr>
</tbody>
</table>

1.12 ORGANIZATION OF THE THESIS

In this chapter, the methodology, objective and scope of the research work, the VM allocation and consolidation problem in cloud
environment used in this research work are discussed. The remainder of this thesis is organized as follows:

Chapter 2 presents the literature survey on VM allocation and server consolidation algorithms and also provides relevant background information about the algorithms in the cloud computing environment. Then, it describes and reviews the different existing VM allocation and server consolidation algorithms developed for the cloud computing environment.

Chapter 3 presents the Power Aware Enhanced Particle Swarm Optimization algorithm to place the VM on the PM. It demonstrates the advantages of using EPSO algorithm, such as faster convergence for obtaining the better solution than existing algorithms.

Chapter 4 deals with the proposed Power-Aware Dynamic Consolidation algorithm. Followed by the description of the algorithm, it presents the results obtained by various experiments.

Chapter 5 gives the VM mapping using Resource Wastage Aware Enhanced PSO algorithm with the objective of minimizing the resource wastage of the datacenter. And also gives the comparative analysis of the proposed algorithm with the existing PSO algorithm in terms of resource wastage of the datacenter.

Chapter 6 discusses the VM selection for migration algorithm when the hosts are overloaded. The proposed algorithm is compared with the existing Black-box algorithm in terms of number VM migrations.

Chapter 7 discusses the add-on software solution for OpenStack platform. An extension is made to Horizon – the OpenStack dashboard project in order to give an alert messages to the cloud administrators about the under
and over utilization of resources in the hosts. The real-time test bed using OpenStack private cloud has been established and the virtual machine placement algorithms used in that are studied. In the future work, the proposed algorithms will be implemented in the private cloud setup.

Chapter 8 concludes the thesis by summarizing the contributions. It also provides insights on future work that could be carried out based on the work presented in this thesis.

1.13 SUMMARY

This chapter gives a brief introduction to the VM allocation and consolidation problems in the cloud computing environment. The motivation, objectives of the proposed research work and organization of the thesis are also presented. The next chapter deals with the literature review and the need for the proposed research work.