Chapter 1

Introduction

1.1 Introduction

Cloud computing has emerged as a popular paradigm for the information and communication industries due to its variety of services offered on the pay-per-use basis. Industries are shifting to Cloud to avoid management and maintenance of computing resources or to build a private Cloud to improve resource utilization and provisioning. A typical Cloud comprises of large number of data centers to handle the user’s requests. A data center is the pool of huge number of computing resources such as storage devices, processing elements, memories, network devices and more. The process of efficient management and deployment of different services on Cloud which are available to end users to the best of their capabilities is known as Cloud service provisioning (Calheiros, Ranjan & Buyya (2011)). Many service providers like Google (Googleappengine (2015)), Amazon (EC2 (2015)) and Microsoft (azure (2015)) have deployed their data centers in different locations to provide their services to the customers around the globe. Cloud delivers services pertaining to infrastructure, platform, and software that refers as Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS) respectively. The preliminary concept behind the working of Cloud is the virtualization technology that creates virtual machine (VM) over physical machine (PM) (Chiueh & Brook (2005)). The Virtual Machine Monitor (VMM) or hypervisor is software layer that creates virtual machines and the services which are made available to end users. There are many open research challenges related to the resource allocation in the domain. In this research, we aim to address the trade-off between energy consumption and performance during resource allocation in Cloud environment. Overall chapter is organized as follows. Starting with
the background of the domain we move to key motivations behind carrying out this research. Further, we formulate the problem statement and research objectives. Subsequently, various research methodologies adopted in the thesis are illustrated followed by research hypothesis, scope and limitation. At the end, we briefly outline our research contribution followed by a note on overall organizational structure of this thesis.

1.2 Background

Due to enormous computing requirement worldwide during the last few years, the number of data centers has increased considerably. This has led to an issue of energy consumption by these data centers and subsequently affecting environment and financial impacts. For instance, in Amazon’s data centers, it is identified and reported (Hamilton (2009)) that (i) expenses related to the cost and operation of the servers is 53% of total budget (ii) energy-related costs is about to 42% of total budget that includes both direct energy consumption of 19% by servers and power used in cooling the infrastructure about 23%. It has been mentioned in Gartner Report 2007, that IT industry contributes total 2% CO\textsubscript{2} emissions in environment. USEPA report in 2007 also mentioned that 1.5% of total US power consumption is used by data centers and costs $4.5 billion. Thus, it is utmost significant for service providers to think upon the two major affecting factors viz. performance and power consumption in Cloud. Efficient resource allocation policies could be used to address these factors, which should take into account the process of appropriately mapping the job on virtual resources and mapping virtual resources on physical servers. Energy-efficient allocation with performance maintenance in the Cloud is yet a challenging issue. Performance and energy-based VM allocation can be viewed as the NP-Complete problem as our aim is to reduce energy consumption and to enhance performance (Walters et al. (2008), Mishra & Jaiswal (2012), Abels et al. (2005)). One of the challenges of energy-efficient and performance oriented scheduling algorithms is the trade-off between energy consumption and performance. In this thesis, we address this issue by a method of the multi-objective optimization and propose a novel algorithm called Pareto Optimal Multi-Objective Optimization based Allocation (MOOA) for Cloud environment.
1.3 Motivation and Problem Statement

Quality of service (QoS) management is one of the key challenges posed in Cloud application which is essentially a problem of provisioning resources to the users to assure a predefined service level along with the facts such as availability, reliability and performance. Cloud infrastructure containing a large number of servers, disks, network devices makes it possible to handle millions of requests from users around the globe. However, the consumers need to pay for usage to the service providers. From the application-framework viewpoint, this paradigm has to provide the 24 x 7 availability, and hence it’s performance is considered by Service Level Agreement (SLA) between Cloud service providers and clients. Correspondingly, the heat and the carbon footprint generated by this computing technology through huge data centers are effecting the environment and they are expected to overtake the CO$_2$ emissions by the air-space vehicular transportation by 2020 (Liu et al. (2009)). As mentioned in background section, data centers are responsible for emitting tens of millions of metric tons of greenhouse gases into the atmosphere, resulting in more than 2% of the total global emissions (Barbagallo et al. (2010)). Other than environmental effect, it is identified and reported that (Hamilton (2009)) (i) expenses of the cost and operation of the servers is 53% of total budget (ii) energy-related costs is about to 42% of total budget that includes both direct energy consumption of 19% by servers and power used in cooling the infrastructure about 23%. USEPA report in 2007 also mentioned that 1.5% of total US power consumption is used by data centers and costs $4.5 billion. Thus, it is utmost significant for service providers to think upon the two major affecting factors viz. performance and power consumption in Cloud. An efficient resource allocation policy could be used to address these factors. This takes into account the process of appropriately mapping the job on virtual resource and mapping virtual resource on physical server. In particular, the following research problems are investigated:

- Detection of the efficient host to allocate VM by proposing an architecture for the same.

- Proposed Multi-Objective Optimization based resource Allocation policy (MOOA) addresses:
  1. Creation of virtual cluster including shutting down underloaded hosts.
2. Selection of efficient host-request pair.

- calculation of fuzzy based co-efficient $\lambda$ to extend resource allocation policy.

1.4 Research Objectives and Research Methods

1.4.1 Research Objectives

This thesis covers the research challenges in relation to energy and performance efficient resource allocation in Cloud environment. To deal with the challenges mentioned in the above research problems, the following objectives have been defined:

- Analyze, classify and explore the research in the area of resource allocation to achieve the systematic understanding of the existing techniques and methods available in literature.

- Analysis of algorithms for resource allocation to justify the usage of different methods available.

- Propose system architecture that
  1. Justifies how the efficient host will be selected.
  2. Develops energy and performance efficient resource allocation algorithm based on multi-objective optimization technique.
  3. An algorithm using fuzzy based approach to calculate co-efficient $\lambda$.

- Design and implement a multi-objective optimization technique based energy and performance efficient resource allocation to evaluate the performance and energy.

- Implement a multi-objective optimization technique based energy and performance efficient resource allocation by including fuzzy approach to generate co-efficient $\lambda$.

Overall objectives of our research are:
1 Introduction

- To improve resource utilization.
- To minimize energy consumption.
- To evaluate the trade-off between performance and energy consumption.

1.4.2 Research Methods

The research methodology followed in this thesis involves following steps:

- Conduct the analysis of the methods available in the domain of resource allocation to achieve the systematic understanding of the existing techniques and methods available in literature.
- Analysis of algorithms for resource allocation to justify the different methods available.
- Conduct the exhaustive study on compatibility of Cloud architecture and hypervisors to understand the real scenario of Cloud.
- Study and analysis of various multi-objective optimization techniques to identify the appropriate method for our algorithm.
- Propose and developed system architecture that includes algorithms for
  1. Detection of moderate host.
  2. Implementation of multi-objective optimization technique based energy and performance efficient resource allocation algorithm.
  3. Implementation of algorithm using fuzzy based approach to calculate co-efficient $\lambda$.

1.5 Research Hypothesis

Effective resource allocation is achieved using Fuzzy-based multi-objective optimization in Cloud datacenter which leads to:

1. Enhancement in Energy Efficiency
2. Improvement in Datacenter Performance

A more detailed process of carrying out the propositions for the hypothesis is drafted beneath:
1. selection of moderate host for more energy and performance oriented solution via proposed architecture that defines different components of node to evaluate, monitor and control dynamic utilization threshold method.

2. Propose a method of multi-objective optimization approach (MOOA) to handle two objective viz. performance and energy and introduce energy and performance oriented resource allocation algorithm.

3. Fuzzy approach is introduced to overcome the limitations of random coefficient $\lambda$ in MOOA and generate pareto optimal set of solutions/pareto front.

1.6 Scope and Limitations

During the last decade, usage of information and communication technology (ICT) has increased drastically resulting into transformation of traditional manual processes into computerized atomization. For startup companies such as small and medium enterprises (SMES), it is difficult to invest into huge computational and infrastructural resources. However, a new paradigm known as Cloud computing could be helpful to them by provisioning computing resources on rental basis in contrast to the orthodox model of procuring equipment. This attractive idea has gain lots of popularity in short spans of time just over decade resulting into numerous data center worldwide. This has raised a separate issue of electrical energy consumed by the computing resources available in the data center. To address the issue, in this research, we intend to provide a mechanism which results into enhancement in resource utilization. Another broader scope of our work is to offer a QoS to the end user by the Cloud service provider which includes (but not limited to) performance of overall system. The byproduct of enhancing the resource utilization would be reduction in number of active computing servers required. This in-turn would reduce energy required by those servers and helps maintaining environmental impacts by reducing CO$_2$ emission.

Assumption / Limitations

- We have assumed that the Cloud has sufficient processing capacity to process job at any point of time. It is further assumed that data center is having sufficient resources to execute the required tasks.
1 Introduction

- In addition, it is practically infeasible to implant real data center having numerous high configured computing nodes for testing our scheme. Hence, we have made empirical analysis in simulation environment. However, the same has been tested on a small set up in in-house laboratory.

1.7 Our Contributions

- The contributions of this thesis can be broadly divided into 4 categories:
  1. Classification and analysis of the area.
  2. Multi-objective optimization based resource allocation algorithm.
  3. Fuzzy based co-efficient calculation.
  4. Implementation of these proposed algorithms.

- The key contributions are:
  - Study of various multi-objective techniques.
  - Problem identification and formulation.
  - Proposed architecture.
  - Moderate host selection.
  - Multi-objective optimization based resource allocation policy (MOOA)
  - Fuzzy based calculation of co-efficient $\lambda$.

1.8 Outline of the Thesis

The core chapters of this thesis are organized in sequence and are derived and covering journal papers published during our research. Our work covered in the thesis is organized as follows:

- Chapter 2 presents a taxonomy and survey of energy efficient Cloud approaches which includes overview of Cloud computing, issues and challenges in Cloud covering security and privacy, interoperability and portability, reliability and availability, energy efficiency, approaches addressing energy efficiency in Cloud, different multi-objective optimization techniques and fuzzy-based approach.
• Chapter 3 presents comparative analysis of open source hypervisors and cloud architectures that describes functionality, architecture, components and working mechanism of different cloud architectures and hypervisors.

• Chapter 4 includes energy efficient VM allocation techniques that discuss the literature survey and research carried out in the area of VM allocation in Cloud. It also discusses and compares the different algorithms available on the same. The detail in the state-of-art of performance and energy efficient VM techniques are discussed.

• Chapter 5 covers multi-objective optimization technique for resource allocation approaches that includes proposed system model, system architecture and multi-objective optimization based algorithm for resource allocation.

• Chapter 6 describes fuzzy-based approach to compute co-efficient being used in multi-objective Optimization. It discusses the fuzzy approach and its applicability in multi-objective optimization based resource allocation.

• Chapter 7 demonstrates experimentation & results

• Chapter 8 depicts conclusion and future work.