AN ADAPTIVE QoS-BASED DYNAMIC RESOURCE MANAGEMENT IN CLOUD COMPUTING SYSTEMS

Thesis Summary

Submitted to

Babasaheb Bhimrao Ambedkar University

(A Central University)

Lucknow

For the Award of the Degree of

DOCTOR OF PHILOSOPHY

In

COMPUTER SCIENCE

By

SWATI SAXENA

Under the Supervision of

DR. NARANDER KUMAR

DEPARTMENT OF COMPUTER SCIENCE
BABASAHEB BHIMRAO AMBEDKAR UNIVERSITY
(A CENTRAL UNIVERSITY)
LUCKNOW-226 025 (U.P.) INDIA

2018

SUMMARY

Cloud computing is a utility-based computing paradigm. It provides a seamless acquisition of computing, storage and network resources to users on a payment basis, in a similar fashion like that of water, gas and electricity. Cloud computing is highly embraced by individuals and business organizations due to the advantages it offers, few of them be like-

- Zero or no opening investment for accessing computing resources like computing, memory and/or network as these are measured in metrical units and provided to cloud users on a demand basis.
- Cloud resources or services are available round the clock with negligible down-time and can be accessed from any location.
- Cloud users' privacy and security features are well-maintained. Moreover, cloud functioning is transparent to its users.
- Cloud users are charged according to their utilization of cloud resources, thereby, extending the advantage of scalability.
- Cloud services are made available on a payment basis and are categorized as software, platform or infrastructure services.

Cloud computing still lies in its infancy stage. Its potential is yet to be unleashed in new business capabilities and advancements. However, like any other advancing technology, cloud also suffers from certain issues and trials which must addressed in time to make cloud adoption a seamless process for all.

The key mechanism behind cloud computing is virtualization which creates an abstraction of actual cloud resources, so that more and more number of users can utilize the potential of this technology. These abstracted resources are enveloped in virtual machines with same functions and interfaces as that of real/actual resources. An important aspect of virtualization is that it is transparent to the cloud user. These virtual resources are offered in different sizes, performance and cost. The entire life cycle of a virtual machine is controlled a virtual machine monitor (VMM) or a hypervisor software. A physical machine or server contains numerous virtual machines and is known as a host. Virtual machines residing on a host are therefore termed as guests. Hypervisor acts as a bridge between the actual resources of a host and its multiple guests by giving a virtual operating environment to execute the guest VM tasks.

Resource management in a virtualized environment like that of cloud systems is a continuous and testing task which reflects the constant changing demand-supply graph of the virtual resources. The existing techniques for resource management are doing their jobs as per their intended purposes, but, the wide-spread adoption of cloud and its dynamic user base requires certain stringent measures towards managing cloud's assets.

The chapter-wise summary of the research is given below in brief:

CHAPTER I

INTRODUCTION

This chapter provides an introduction to cloud computing technology and its importance as a new paradigm of computing. The evolution of cloud and its classified services are described in detail. It also highlights the important attributes of cloud computing, its strengths and its weak points. A number of widely accepted and most cited

definitions of cloud computing are also given. Cloud computing is a special type of distributed computing in which any type of resource, physical or virtual, can be made available to the users, worldwide, by means of powerful technology called virtualization of resources. Virtualization is given due importance with discussion on its types and utilities. The issue of resource management techniques in cloud computing environment is discussed with an outline on the primary objectives of the research work.

CHAPTER II

REVIEW OF LITERARURE

This chapter gives details of the existing literature on cloud computing technology, available resource management techniques in cloud environment, for example job scheduling, resource auction/allocation, admission control and virtual machines consolidation issues. The rapid adoption of cloud computing by the society emphasis the need efficient and feasible resource management techniques for managing the cloud resources. An extensive study of related research papers on resource management techniques suggests that the objectives and implementation of resource management in cloud networks are very different from classical networks. For example, in classical networks resources are only physical while in cloud, resources are physical as well as virtual. Therefore, a different approach is required for cloud computing to manage resources effectively. Several reputed journals, e-books, Wikipedia, etc. are consulted for understanding the new research problems and more than 150 references are given in the thesis.

CHAPTER III

LOCATION-AWARE VM MIGRATION IN CLOUD DATACENTRES

In this chapter, a location-restricted migration mechanism of an overloaded virtual machine is presented in a clustered data centre. Server clusters are formed based on the application type hosted by VMs, and migration of VM is restricted within its home cluster. This helps in reducing the migration volume along with the total migration time to a considerable value. Selection of destination host for migration is based on the criteria of distance, which further speeds up the task at hand and also reduces energy consumption. A detailed mathematical analysis of existing live migration process is presented to point out the factors which play a crucial role in the migration performance. Sequence and class diagrams supplied help in better understanding of the proposed migration technique. The proposed migration technique is simulated on the CloudSim-3.0.3 simulator.

The content of this chapter is published in-

1. Springer series on Lecture Notes in Networks and Systems, vol. 9, pp. 119-129, 2018, ISSN 2367 – 3370.
2. Elsevier Procedia Computer Science Journal, vol. 45, pp. 823 – 831, 2015, ISSN 1877 – 0509.
3. Research Journal of Recent Sciences, vol. 3, pp. 13-20, 2014, ISSN 2277-2502.

CHAPTER IV

LOAD-AWARE VM PLACEMENT IN CLOUD DATACENTERS

In this chapter, an efficient VM placement and migration technique is proposed based on the three-tier architecture and considers load as a prime objective. We have used a multi-metrics analysis to distribute VMs evenly and to maintain a stable equilibrium

inside a data centre. We have applied Analytical Hierarchy Process (AHP) for efficient virtual machine placement using four post placement metrics which defines some of the key Service Level Agreement (SLA) parameters. The metrics considered for placement and migration are all load-centric and promise fewer migrations and SLA violations. The high points of the proposed VM placement technique are increased profits to the cloud service provider, fair and service availability to cloud users by placing a virtual machine to the best available physical machine in a dynamic fashion using the concept of clusters. It also avoids unnecessary migrations by balancing load of a cluster. Simulation results show a remarkable reduction in migrations which improves energy conservation inside the data centre. Application of AHP in balancing a data centre's load is still unexplored. The presented placement technique selects the best candidate machine for placement, hence upgrades the performance.

The content of this chapter is published in-

1. Indian Journal of Science & Technology, vol. 10(32), pp. 1-8, 2017, ISSN: 0974 - 6846.
2. BRIS Journal of Advances in Science and Technology, vol. 4(1), pp. 22-33, 2017, ISSN: 1444 - 8939.

CHAPTER V

PREFERENCE-BASED RESOURCE ALLOCATION IN CLOUD DATACENTERS

Efficient resource allocation is a major concern in utility-based systems such as cloud computing. Cloud users approach a cloud service provider to execute their tasks which require cloud resources in various measures. In return, users pay for the resources utilized by them. To cater multiple users in the same instant with varying resource

requirements, a cloud provider applies certain resource allocation techniques which must not only bridge the gap between the demand and supply but also must provide certain benefits to both the service provider and the service user. This chapter presents a resource allocation mechanism where first a market-driven auction process takes place to ensure truthfulness and profit to the service provider followed bya preferential payment process. Here, the winner of the auction is supposed to make the payment for his resource requirements by an amount which is far less than his actual bid value. The suggested resource allocation scheme is compared with the off-line VCG auction technique to register a finer performance outcome w. r. t. optimal resource distribution, fair allocation and comparatively better revenues to the service provider.

The content of this chapter is published in-

Elsevier Procedia Computer Science Journal, vol. 57, pp. 104-111, 2015, ISSN: 1877 - 0509.

CHAPTER VI

TOKEN-BASED PREDICTIVE JOB SCHEDULING IN CLOUD DATA CENTERS

Task scheduling in a distributed environment, like cloud computing, is an important function which influences the overall performance. Considerations in chapter six are a single data center where a number of users are competing for limited resources. An ideal solution must service all these concurrent jobs while maintaining quality of service parameters and ensuring optimum usage of cloud's resources. Unfortunately, dynamic user requirements and limited availability of resources often makes it difficult to satisfy every demand without compromising the quality of performance. Here, the

proposed scheduling technique ensures low turnaround time and waiting time by balancing the demand curve with allocation frequency.

The content of this chapter is published in-

Research Journal of Recent Sciences, vol. 4, pp. 29-33, 2015, ISSN: 2277-2502.

CHAPTER VII

ENERGY-EFFICIENT MULTI-CRITERIA BASED ADMISSION CONTROL IN CLOUD DATACENTERS

Job scheduling requires an intelligent selection of incoming jobs or requests so that their execution improves the efficiency of data center as a whole. Cloud data centers utilize certain admission control procedures to accept/reject a service request. Major admission control mechanisms, practiced today, consider single parameter as objective. However, there are multiple dimensions to consider while accepting or rejecting a request. This chapter presents a multi key-based admission control mechanism which ranks the incoming service requests based on various performance keys and accordingly admits the most suitable request. Simulation experiments also confirm the validity and usefulness of the scheme with respect to energy conservation in a data centre.

The content of this chapter is published in-

International Journal of Innovations & Advancement in Computer Science, vol. 5(6), pp. 110-115, 2016, ISSN: 2347 – 8616.

CHAPTER VIII

ENERGY EFFICIENCY AND OTHER SLAs

This chapter presents two aspects of modern cloud data centers- energy efficiency and cloud adoption. To make data center energy efficient, steps must be taken to limit the communication workload inside a data center and to switch off network elements when not in use. In this chapter, we put forward two models, one an energy consumption model to outline the factors which are responsible for excessive energy consumption in a data center and second an energy conservation model which tunes these responsible factors so that energy consumed is less. Proposed conservation models are applied to hierarchical and recursive architectures to verify their practicality. Simulation results mark a notable advancement in energy conservation. Also, these models are applicable to many DCN architectures.

The focus is to make cloud adoption a clean and transparent process by clearly outlining the individual rights and responsibilities of both the service provider and the cloud adopting business. In this direction, a semantic framework is introduced in this chapter which address all the risks and challenges mentioned above and provides its best possible solution. This framework establishes the trust between the two involved parties and is crucial for the success of cloud computing technology.

The content of this chapter is published in-

1. BRIS Journal of Advances in Science & Technology**,** vol. 4 (2). pp. 80-87, 2017, ISSN. 1444-8939.
2. Springer series on Advances in Intelligent Systems and Computing, vol. 434, pp. 179-187, 2016, ISSN 2194 – 5357.

CHAPTER IX

CONCLUSIONS AND FUTURE PESRPECTIVES

This chapter is devoted to the conclusions of the given research work and the future work in the area of cloud resource management. Cloud computing systems have a long way to go. Future computing world will strive on virtual resources promising seamless and continuous services. Hence, the issue of cloud resources management must be addressed in a timely and priority fashion. In the present research work, several issues of cloud resource management have been identified and independent solutions to each issue have been proposed. These are -

- VM Migration
- VM Placement
- Resource Allocation
- Job Scheduling
- Admission Control
- Energy efficiency and other SLAs

The solutions provided for each sub-problem are feasible, scalable and dynamic in nature and efficiently manage cloud resources as validated by the simulation results. As future perspectives, this work can be extended for newer DCN architectures and interoperable clouds which enable a cloud user with the flexibility of shifting his/her acquired resources between data centres and no disruption in serviceability.