INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 5, Issue 3)
Available online at: www.ijariit.com

Optimization of virtual machines in cloud environment

Rishabh Narayan Parasar
rishabhnarayansharma@gmail.com
Roorkee College of Engineering, Roorkee, Uttarakhand

ABSTRACT

The emergence of cloud computing has facilitated the provisioning of computing resources in an on-demand basis that can be swiftly allocated, released and reallocated with minimum management effort and cost. One important element of cloud is the virtual machine which encapsulates business services and acts as a resource carrier. An important task of cloud computing is to find an optimal placement scheme that can map the virtual machines to physical machines. With the increasing prevalence of large scale cloud computing environments, how to efficiently place VMs into available computing servers has become an essential research problem. These research works present a Virtual Machine Placement and Load Rebalancing Based on Multi-Dimensional Resource Characteristics in Cloud Computing Systems (VMP-LR) to improve the efficiency of VM placement.

Keywords— *Cloud computing, Virtual machines*

1. INTRODUCTION

The end of the 20th century and the beginning of the 21st century are considered to be the golden era of computing, during which, computing and digital solutions expanded their place from both commercial and academic fields to the general public. This expansion has particularly been accelerated by the twin developments of desktop computing and the World Wide Web (WWW). In these environments, the tasks are executed by applications hosted on dedicated servers. Increase in the availability of economical and sophisticated hardware/software often led to the spreading of these servers across datacentres. However, this has caused an increase in costs in terms of overhead and maintenance). Research studying this issue noted that most of the applications in these servers utilize less than 20 per cent of the resources allocated to them, as the majority of enterprise applications have dynamic workloads with resource requirements that vary according to the time, season and geographical locations. (For example, a bank web server will have high resource request rates during day time with compared to night and entertainment web server which may have high request rates after office/college hours. In this phenomenon, resource wastages or resource underutilization, have been envisaged. For example, CPU-extensive applications waste I/O (Input/Output) resources and I/O-bound applications do not utilize CPU-resources fully, thus leading to under-utilization of resources. Under-utilization of resources also arises because of resource contention, where the resources are over-provision to applications.

2. REVIEW

In today's cloud scenario, millions of users' share resources by submitting their requirements to the cloud system, where the problems of VM placement and resource sharing held a critical position. In general, solutions to these problems indicated methods by which VMs were allocated to nodes in a manner that had balanced the load on all PMs effectively and at the same time attained a specified quality of service. Research works focused on improving the problem of VM placement and load balancing has been designed with different goals that considered various aspects of the processes. For example, the goals could be set to reduce VM allocation time, reduce the number of PMs needed, reduce resource wastage and improve power efficiency.

There was a wide range of contributions in the greening of cloud data centres, among which a notable number of methodologies have addressed the problem of VM placement, load balancing and load rebalancing (consolidation). The survey of (Mishra *et al.* 2011) reviewed existing methodologies for VM placement, server consolidation, and load balancing was studied to uncover their anomalies and their causations. Similar work was reported previously by (Buyya *et al.* 2010), who discussed the open challenges in dynamic resource allocation.

3. METHODOLOGY

Advancements in hardware and software technology are motivating both users and researchers to search for techniques that challenge and improve the available industrial standards for placing VMs to appropriate PMs. This can be performed either by developing new competitive methodologies or by enriching the operations of existing methodologies as several applications require reliable models that are efficient both in the manner of finding optimal PMs, while reducing response time, saving power consumption and thus reducing energy usage. This research work takes the second type of methodology where the existing solutions are enhanced by identifying solutions to the issues still present in them.

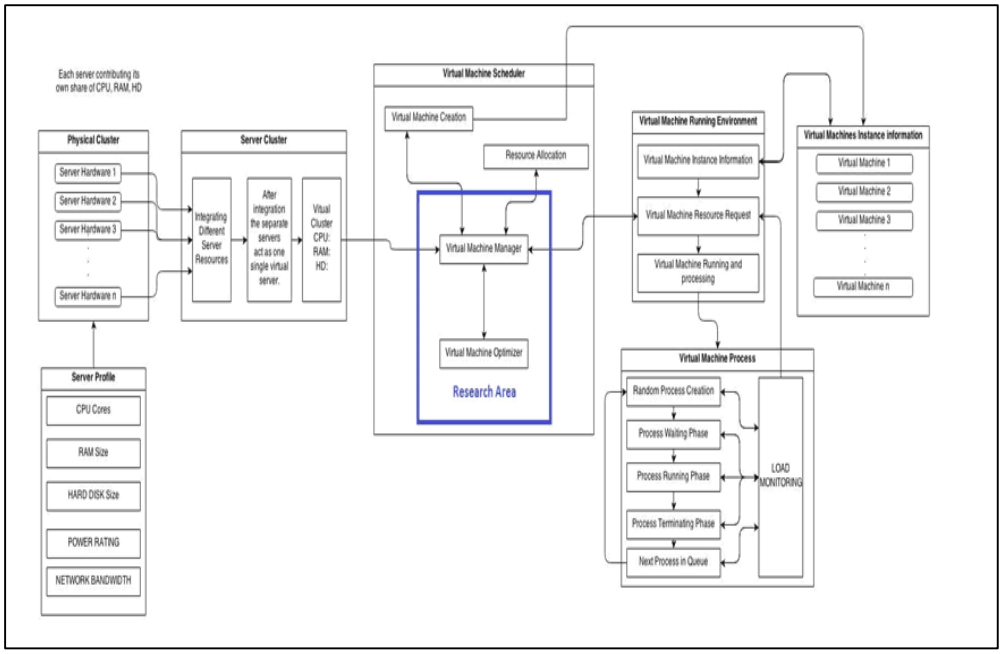

Fig. 1: Methodology

The figure shows the overall cloud framework model, where the proposed algorithm is implemented. In the cloud environment, the physical clusters are formed by adding a set of physical servers, each contributing its own share of resources (such as CPU cores, main memory, disk capacity, and network bandwidth). The users generate Virtual Machine Instances (VM requests) by submitting their resource requirements for running the applications to the computing system. These requests enter into the queue, where they wait for placement. The VM requests are handled by the virtual machine scheduler in order to find the appropriate physical machine and are allocated to the best matching PM that matches the requirements. The flow of user request to the resources is as follows:

- The registered users access the portal server and request the virtual machines with the parameters including a quantity of core, frequency, memory, storage space, OS and etc.
- Portal Server sends the request to the scheduling server.
- The scheduling server searches the physical machines to find the host to create the virtual machine according to the metadata of the physical machine, which records the operation and configuration details.
- The scheduling server chooses one optimal server and then sends the creating command of the virtual machine to its agent server.
- The scheduling server chooses the virtual machine template from the stored templates within the cloud storage administration centre and sends one request for the template to the agent server.
- The requested virtual machine image will be sent (or mapped) to the physical server based on the template, the scheduling server will start the virtual machine if the image is loaded successfully.
- If the virtual machine starts successfully, the user can access the virtual machine through RDP, VNC, ICA or SSH.

4. DESIGN

In a cloud computing environment, dynamic resource requests and their management can be effectively performed using virtualization technology. This helps to solve the problem of heterogeneity of resources and platform irrelevance. In this environment, load balancing (also handled in a dynamic fashion) helps to map (or remap) VMs and PMs according to the change in load. These advantages have made virtualization an optimal and comprehensively implemented technology in cloud computing systems. However, in order to accomplish the best performance, it is more desirable to have the VMs fully utilize its services and resources by allowing the allocation of resources (scheduling) and load balancing to guarantee optimal resource utilization. They are often used in cloud systems so as to allow the multiple users to share the resources efficiently, maintain all computer resources busy (load balancing) and to attain a target QoS (Quality of Services) many of the scheduling algorithms are job-oriented where the mapping is performed based on the user requirements. In this set-up, the resources are allocated in bundles and the scheduling algorithms are more focused on the task characteristics of the user applications and neglect the resource availability in the cloud system. In this scenario, the users obtain maximum benefit while using a specific combination of resources like CPU cycles, main memory, disk space, and network bandwidth. Resource bundling complicates traditional resource allocation models and may result in wastage of resources and long waiting time.

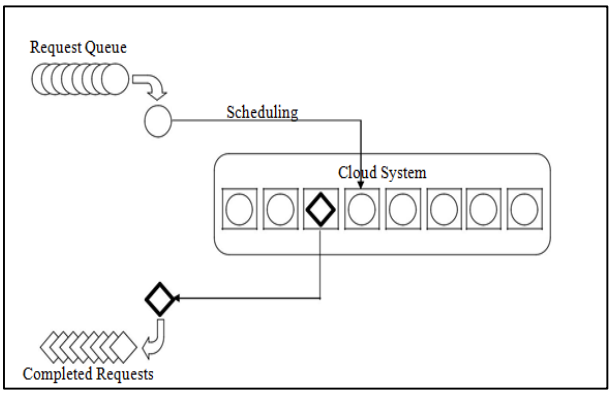

Fig. 2: Design

In order to solve these issues, this research work proposes VM scheduling and load balancing algorithm that is both user-oriented and resource-oriented. The proposed algorithm is user-oriented as it accepts resource requirements from the user and resource-oriented as it considers cloud resource repository status so as to enable dynamic resource scheduling, that

considers both resources available and network traffic. Moreover, as mentioned in chapter 3 (Methodology), the scheduling algorithm is designed to use 3-dimensional resource requirements, CPU, RAM, and network bandwidth, in order to improve the performance of the cloud system. This approach can provide multiple advantages to cloud systems including minimizing the variations during resource demands, improving overall efficiency, improving the relationship between scheduling and load balancing with each unit of time. Phase I of the research work is concerned with scheduling and load balancing. In order to optimize its operation, VMP-LR performs scheduling and load balancing in two steps, as listed below.

- VM Queuing (Linked with Resource Request Handling Component)
- Scheduling and Load Balancing (Linked with Placement Component)

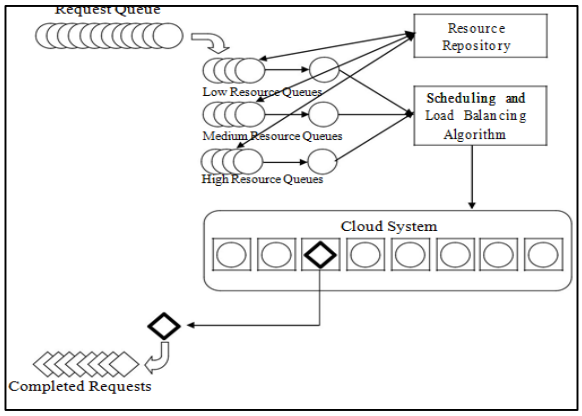

Fig. 3: Block diagram

5. CONCLUSION

Recent years have cloud computing technology emerging as a commodious resource of computation power. Cloud computing system can provide elastic scalable and on-demand resources for efficient execution of different applications. One key element of the cloud computing system is Virtualization technology. Virtualization allows sharing of hardware and software resources with multiple users and applications. An important stage of virtualization is virtual machine placement, which is the key focus of this research work.

Virtual machine placement is the process of mapping resource requests in the form of VMs to PMs in an optimal manner that improves resource utilization and reduces the time involved in providing the service desired. Traditionally, when the number of VMs and PMS is small, mapping of VMs to appropriate PMs is possible manually. However, the current scenario faces a tremendous increase in the number of VMs and PMs, which makes automation of placement task mandatory. Existing automated solutions have to evaluate several numbers of possible mappings for a given set of VMs and PMs and thus, require improved intelligent placement heuristics to narrow down the search for a solution to obtain near-optimal placement plan This work proposes a system called Virtual Machine Placement and Load Rebalancing Based on Multi-Dimensional Resource Characteristics in Cloud Computing Systems (VMP-LR) to improve the VM placement process. This work investigates challenges involved in the problem of resource placement and scheduling in cloud environments, tackles the problem using combinatorial optimization techniques and mathematical modelling. Several experiments, aiming at finding the best suitable algorithm for each queue. The queue and its designated algorithm are listed below:

(a) High Request Queue: Scheduling and Load Balancing Algorithm using Enhanced Max-Min, Ant Colony Optimization (ACO) and Artificial Bee Colony (ABC) (SLAM2A) algorithm

(b) Medium Request Queue: Scheduling and Load Balancing Algorithm based on First Fit, Best Fit and multi-level grouping Genetic algorithms (SLAFBG)

(c) Low Request Queue: Scheduling and Load Balancing Algorithm based on Enhanced Max-Min algorithm with Particle Swarm Optimizer (SLAMP)

7. REFERENCES

[1] Abdelsamea, A., Hemayed, E.E., Eldeeb, H. and Elazhary, H. (2014) Virtual Machine Consolidation Challenges: A Review, International Journal of Innovation and Applied Studies, Vol. 8, No. 4, Pp. 1504-1516.

[2] Agrawal, K. and Tripathi, P. (2015) Power Aware Artificial Bee Colony Virtual Machine Allocation for Private Cloud Systems, International Conference on Computational Intelligence and Communication Networks Pp.947-950.

[3] Agrawal, S., Bose, S.K. and Sundarrajan, S. (2009) Grouping genetic algorithm for solving the server consolidation problem with conflicts, ACM Proceedings of the first ACM/SIGEVO Summit on Genetic and Evolutionary Computation, Pp. 1-8.

[4] Al Shayeji, M.H. and Samrajesh, M. D. (2012) an energy-aware virtual machine migration algorithm, Proceedings of International Conference on Advances in Computing and Communications, Pp. 242-246.

[5] Ali, A., Belal M.A. and Al-Zoubi, M. (2010) Load Balancing of Distributed Systems Based on Multiple Ant Colonies Optimization, American Journal of Applied Sciences, Vol. 7, No. 3, Pp. 428-433.

[6] Amazon Elastic Compute Cloud (Amazon ec2) (2013) http://aws.amazon.com/ec2, Last Accessed during August 2016.

[7] Armbrust, M., Fox, A., Griffith, R., Joseph, A.D., Katz, R.H., Konwinski, A., Lee, G., Patterson, D.A., Rabkin, A., Stoica, I. and Zaharia, M. (2009) Above the Clouds: A Berkeley View of Cloud Computing, Technical Report No. UCB/EECS-2009-28, http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/ EECS-2009-28.html, Pp. 1-25.

[8] Banerjee, S., Mukherjee, I. and Mahanti, P.K. (2009) Cloud Computing Initiative using Modified Ant Colony Framework, World Academy of Science and Technology, Vol. 56, Pp. 221-224Amalarethinam, G.D. and Beena, T.L.A. (2016) Workflow scheduling for public cloud using a genetic algorithm (WSGA), IOSR Journal of Computer Engineering, Volume 18, Issue 3, Ver. V, Pp. 23-27.

