VM Scheduling in Cloud Computing using Meta-heuristic Approaches

Mamta Khanchi
Research Scholar, Department of Computer Science & Applications, Kurukshetra University, Kurukshetra, Haryana
khanchimamta11@gmail.com
Sanjay Tyagi
Assistant Professor, Department of Computer Science & Applications, Kurukshetra University, Kurukshetra, Haryana
tyagikuk@yahoo.com

Abstract
Cloud computing promotes the provision and use of IT infrastructure, platform and applications of any kind in the form of services that are electronically accessible via internet in a dynamically scalable and metered manner. It is still in infancy, so as to reap out its full benefits, much research is required across the various topics. One of the important research issues which need to be focus for its efficient performance is VM scheduling. The goal of VM scheduling is to allocate the processing cores of a host to virtual machine that optimize one or more objectives. Scheduling in cloud computing belongs to a category of NP-hard problem due to large solution space and hence it takes a large amount of time to find an optimal solution. At present, there is no algorithm which can produce optimal solution within polynomial time to solve these problems. Some meta-heuristic based techniques have been proved to achieve near optimal solutions within acceptable time for such problems. This paper provides a review on VM scheduling for load balancing, cost effectiveness and energy saving using various meta-heuristic approaches such as ACO, PSO, GA etc.

Keywords- Ant Colony Optimization, Cloud Computing, Genetic Algorithm, Load Balance, Virtual Machine.

1. INTRODUCTION

In the recent years, distributed computing paradigms (cluster, grid, cloud) have attained much attention due to rapid elasticity, reliability, information sharing and low cost than a single processor machine. Cloud computing has come out as the most prominent distributed computing paradigm out of all others in the present scenario. It has become one of the most important IT infrastructures which is capable of delivering computational services. In this, virtual form of hardware and software are provided to the customers through internet. Cloud service providers offer computing resources (processing cores, storage etc.) at very low charges with cost equivalent to the actual consumption [1]. Cloud computing involves the features of some popular technologies such as grid computing, utility computing and virtualization. Like grid computing, cloud makes use of a bunch of computer resources same as distributed system with the purpose of addressing the user's request. Similar to utility computing, it provides the way for capturing the computing power & storage capacity so as to be allotted as measured service at low prices. The virtualization technology which is an abstract and logical aspect of physical resources and consists of large number of servers, data stores, networks, and software, virtualized a single system into number of virtual systems. The major benefit of this technology is that various physical servers can be substitute by a large resource pool and any load requirement can be supplied from that pool.

Cloud computing offer three type of services:- Software as a service (SaaS), Platform as a service (PaaS), Infrastructure as a services (IaaS)[2]. Based on the architectural model, cloud computing can be divided in to public, private, hybrid and community clouds.

This paper is organized in following sections: Section 1.1 describes the scheduling in cloud computing system. Section 1.2 discusses the virtual machine scheduling. Section 2 discuss about literature review. Section 3 describes various research issues and challenges in this field. Section 4 brings the conclusions of the paper.

1.1. SCHEDULING IN CLOUD COMPUTING

Scheduling the basic processing unit on a computing environment has always been an important issue. It play a very important role in various fields such as process, threads or task scheduling in operating system; job shop, flow shop or open shop scheduling in production environment, printed circuit board assembly scheduling, scheduling of users' task or virtual machine scheduling in grid or cloud environment.

In the layered architecture of cloud computing, scheduling takes place at three places-at application

layer, at the virtualization layer and infrastructure layer [3]. Scheduling at the application layer, is called cloudlet scheduling or task scheduling. Its main objectives are to reduce makespan & users' cost, and to increase the application performance, reliability and to achieve the user deadline constraint. Similarly, scheduling at the virtualization layer also called virtual machine scheduling, is performed with the main objective of optimal resource utilization (cpu, storage, network bandwidth etc.) for energy saving, load balancing, to minimize turnaround time, so as to improve the providers' efficiency. At last, scheduling at the infrastructure (deployment) layer is concerned with the optimal and well designed framework, service employment, data routing and efficient application transfer.

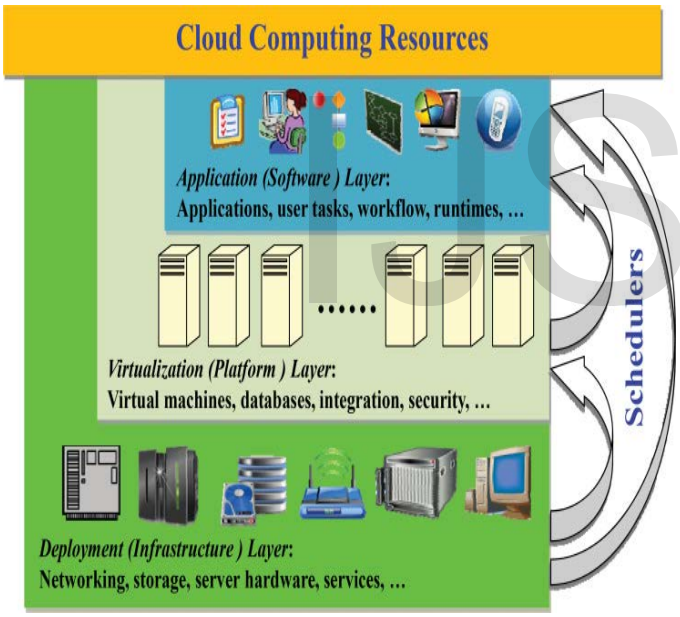

Figure1: Resource scheduling in cloud computing [3].

The major difference in scheduling cloudlets in cloud and scheduling over conventional individual processor is that processing elements are larger in numbers in a cloud environment. Optimized scheduling of tasks over those processors (i.e. solution space is large) comes under NP-hard problem. For this reason, some studies have latterly adopted meta-heuristic algorithms to find feasible solution for cloud resource scheduling.

1.2. VIRTUAL MACHINE SCHEDULING

Scheduling of resources in cloud environment is basically a resource assignment method for the user tasks. It is mainly performed for the better resource utilization. In order to provide cloud resources as utility via internet, physical resources are transformed into virtual resources. Hence how to allocate and move these virtual resources over the available physical resources has become an important research topic.

Basically, VM Scheduler decides how virtual machines are to be allocated with processing cores and how many processing cores are to be assigned to each VM. Space shared and time shared are two well-known virtual machine scheduling strategy.[4]. Space shared scheduling strategy is the process of assigning specific processing elements to a specific VM until and unless it has finished its execution. It means that if there are more VMs than available resources, the remaining VMs have to wait until enough free resources are available. While in time shared VM scheduling policy, fraction of the processing elements are shared among running VMs and every VM run concurrently.

As discussed earlier in this paper, scheduling at the virtualization layer is performed with the main objective of load balancing, for energy saving and for cost reduction:

- *Load balance oriented scheduling*: deals with assignment of virtual machines on relevant hosts and to balance the load on each host [5].
- *Energy awareness oriented scheduling*: deals with the scheduling of virtual resources for small number of physical resources so as to reduce the overall energy consumption.
-*Cost saving oriented scheduling*: aims to optimally schedule the virtual machine over the physical machine so as to reduce the overall cost.

2. LITERATURE REVIEW

In the literature, some works have been performed on the VM load balancing, while some work have been on energy conservation. There also exists work on optimal scheduling of VMs over physical resources for cost reduction Here, is the classification of these works.

2.1. Scheduling for load balancing

A GA approach was proposed for mapping of virtual machines over physical machines [6]. As several numbers of VMs can be constructed on each PM, therefore the proposed approach used a tree type encoding strategy to demonstrate this. The number of

VMs might vary during the cloud runtime i.e. creation of new VM or destruction of available VMs might took place. The computational need might also vary. When these changes took place, the approach remapped the interrelation between virtual machines and physical machine for maximization of load balance and minimization of VM reassignment. The proposed approach was evaluated in OpenNebula cloud platform and results were compared with Least load/Rotating schedule.

In [7], a multi-objective GA approach was proposed for scheduling of virtual machines over available physical machine. The processors and storage space load balance was collectively taken as a multi-objective problem and evaluated with the help of advanced non-dominated sorting genetic algorithm (NSGA-II).The cloudlet requests were offline and experimental scale was 10 VMs and 6 PMs. The approach was compared with Random, Static and Rank algorithm and reported that minimum 1.06 and maximum 40.25 speeds up of balance can be achieved using NSGA-II.

In addition to genetic evolutionary approach, some swarm intelligence approach such as ACO and PSO have also been stated to schedule VMs over available physical resources. In [8], an ACO approach was introduced to find the closest idle or under-loaded cloud resource speedily and for sharing the load of an overloaded virtual machine flexibly. For optimal identification of physical resource and load sharing of overloaded virtual machines, ants' behavior was followed in this approach.

In [5], a hybrid algorithm based on meta-heuristic approach was proposed for load balance oriented VM scheduling in cloud environment. Particle swarm approach was combined with ant colony optimization in this methodology. Previous information was used in this approach for workload forecasting of incoming cloudlet requests. In order to reduce the computation time, a pre-reject step was introduced that reject those requests that can't be fulfilled before scheduling them. The experimental analysis indicated that the proposed methodology can manage the load in a continuously changing environment & outperforms the existing methodologies.

2.2. Scheduling for cost effectiveness

In [9], an evolutionary optimal VM placement algorithm (EOVMP) was introduced. In this, a demand forecaster was used to forecast the computational needs of the cloud consumers. This hybridized approach combined the concept of GA, ACO and PSO for scheduling of VMs on physical resources efficiently. The author reported that EOVMP algorithm could provide near optimal solution for stochastic problems and prediction of demand forecaster had acceptable efficiency.

In [10], a GA approach was introduced that used topological information to schedule VM resources. Additionally, there was employment of prediction engine to take advantage of topological intelligence & for performance evaluation. The target was to decrease the total finishing time of application that automatically results in price reduction. The overall performance gain of their approach ranged from 8% to 41% when compared to completion time of H-2 and RR-S.

In [11], GA was enhanced using an evolutionary approach to solve same kind of scheduling problems. Though, the experiment results were not much elaborated.

2.3. Scheduling for energy conservation

In [12], an improved GA algorithm was designed that had the capability of assigning VMs efficiently, allowing the maximum utilization of available resources. As a result, the used physical resources had the maximum usage rate and the number of physical resources had decreased. The GA scheduler used in this approach, operate at the starting of cloud system and initiated each time while the number of virtual machine or actual machines changed. The result were compared with First fit, Round Robin and traditional GA approach and found that the speed of IGA was about two times of the original GA scheduling approach in grid computing and the resources usage ratio was greater each time.

In [13], the hybrid GA in combination with Knapsack problem with multiple fitnesses was adopted. In that, ordering of genes was done on the basis of memory size of corresponding VMs. In this manner, for an invalid chromosome, for e.g., a greedy approach could be adopted for migration of VMs to some other physical server if virtual machine memory requirement placed over that PM, go beyond the actual machine memory. The approach provided the high resource utilization and accordingly decreased the number of physical machine for efficient energy conservation.

In [14], a hybrid GA approach with local search procedure was used for decreasing the energy consumption by considering the communication network in addition to physical machine in a data center. It was concluded that the hybrid GA was

scalable and significantly outperformed the original GA.

In [15], a scenario was considered that the virtual resources as well as the physical resources would change during the cloud runtime. Hence, an encoding scheme was designed where the ith gene was considered as a triplet <VM_i, PM_j, t>, demonstrated the scheduling of ith VM over jth physical machine at time event t. The objective was to increase resource utilization. It was reported that the presented resource managing mechanism (QoS constraint base GA algorithm) have the maximum performance when compared with Haizea.

In [16], similar kind of work was carried out. In this, an ACO approach was adopted to schedule VM for decreasing energy consumption. A multi-dimensional bin-packing (MDBP) problem was used to represent resource scheduling problem. In this approach, an ant tried to select the next virtual machine (in same way as to choose an item from MDBP) efficiently over the current physical machine (like bin in MDBP) using knowledge of pheromones. The target was to put as many VMs as possible over each PM, so that overall energy can be conserved by decreasing the number of operating physical machine. The results were compared with simple greedy algorithm such as First fit decreasing (FFD) CPLEX and reported that ACO approach is more efficient.

In [17], this study was continued by using same approach for efficient VM scheduling, so as to conserve energy in a large scale cloud infrastructure.

Similar work was also carried out in [18]. The two main objectives were to decrease the overall resource wastage and energy consumption. A multi-objective ACO approach was used to solve the problem. The results were compared with max-min ant system (MMAS), multi-objective genetic algorithm and bin-packing algorithm. The results were found more efficient than the algorithms with which it was compared.

In [19], a novel ACO based approach i.e. ACO-VMP (ACO-Virtual Machine Placement) was proposed to decrease the number of PMs. It was reported that the algorithm was able to achieve more efficient resource usage than First-fit decreasing (FFD), specifically when there are large number of VMs.

In [20], a novel technique was proposed for assignment of virtual machine to relevant physical machine. The experimental results proved that proposed approach decreased the energy consumption and migration rate as well. However, that approach increased the SLA time per host.

3. RESEARCH ISSUES AND CHALLENGES

The effective performance of EC algorithms (genetic algorithm, differential evolution, differential search algorithm etc.) and swarm intelligence algorithms (ACO, PSO etc.) has increased attention in scheduling of cloud resources currently. Although lot of work has been attempted, research in this area is still a challenge. Various issues which need to be explored are as follow:

3.1Real time scheduling

Cloud resources as well as requests for these cloud resources may change dynamically in a cloud environment. A scheduling approach should be able to cope up with these dynamic changes intelligently. Existing EC algorithms for cloud resource scheduling finds the optimal solution offline. There is requirement of algorithms that could find the optimal solution online. Therefore, real time scheduling is still an issue in cloud computing.

3.2 Large-Scale scheduling

With the exponential growth of cloud computing, cloud resources, cloud consumers, cloudlets and workflow are increasing rapidly. This has made the cloud environment gigantic. Hence the size of cloud challenges the current scheduling approaches. Very huge search space of the optimization problems may be an issue in large scale scheduling [3].

3.3 Multi-objective Scheduling

Parties involved in cloud computing such as cloud service provider, cloud consumer and broker may have their own independent objectives. For example, cloud user may require high quality of services, while a service supplier may require maximizing the usage of available resources using an effective scheduling approach.

Also, different cloud consumers or the same consumer at different moment of time may demand different quality of services such as high computing speed, low cost and so on. At present, scheduling approaches for IT resources consider only one optimization objective at a time. Multi-objective scheduling in cloud computing is still a challenge and would become a more significant research topic in future.

3.4 Scheduling IT resources for Big Data and Data Analytics

Big data consists of high volume, velocity, variety and/or veracity of data. Cloud scheduling techniques can face challenges while managing and scheduling these kind of data sets. Some popular meta-heuristic approaches such as Evolutionary Computation (self adaptive differential evolution or cultural algorithm) can help to identify the cloud scheduling requirements of big data.

Scheduling the IT resources for predictive data analytics using nature-inspired EC algorithm would be a significant research topic in IT industry in future. These algorithms would help to identify the relationship between unconnected pieces of data in predictive data analytics and sentimental analytics.

4. CONCLUSION

This paper starts with introduction to cloud computing. Through describing the layered architecture of cloud computing, virtual machine scheduling in cloud computing is discussed here. Following this, previous works in VM scheduling for load balancing, cost effectiveness and energy saving with their objectives, achievements and drawbacks have been presented. Looking forward, various research issues and challenges have been discussed including real time scheduling, large-scale scheduling, scheduling IT resources for big data and data analytics. Research in this field is in its initial stage. With the advancement of big data and internet of things, some new issues may come in existence.

REFERENCES

[1] S. Sotiriadis, N. Bessis, F. Xhafa and N. Antonopoulos, "Cloud virtual machine scheduling: Identifying issues in modelling the cloud virtual machine instantiation," 2012.

[2] J. J. Wang and S. Mu, "Security Issues and Countermeasures in cloud computing," in *IEEE international conference on Grey Systems and Intelligent service (GSIS)*, Nanjing, 2011.

[3] Z.-h. Zhan, X.-f. Liu, Y.-j. Gong and J. Zhang, "Cloud Computing Resource Scheduling and a Survey of its Evolutionary approaches," *ACM*, pp. 63:1-63:33, 2015.

[4] R. N. Calheiros, R. Ranjan, A. Beloglazov, C. A. F. Rose and R. Buyya, "CloudSim: A toolkit for modelling and Simluation of cloud Computing Environment and Evaluation of resource provisioning algorithms," pp. 23-50, 2010.

[5] K.-M. Cho, P.-W. Tsai, C.-W. Tsai and C.-S. Yang, "A hybrid meta-heuristic algorithm for Vm scheduling with load balancing in cloud computing," *Springer*, pp. 1297-1309, 2014.

[6] J. Hu, J. Gu, G. Sun and T. Zhao, "A scheduling Strategy on load balancing of Virtual Machine Resources in Cloud Computing Environment," IEEE, 2010, pp. 89-96.

[7] J. Zhao, W. Zeng, M. Liu and G. Li, "Multi-objective optimization model of virtual resource schedulig under cloud computing and it's solution," Hong-kong, IEEE, 2011, pp. 185-190.

[8] X. Lu and Z. Gu, "A Load-adaptive cloud resource scheduling model based on Ant colony Algorithm," Beijing, IEEE, 2011, pp. 296-300.

[9] C.-C. T. Mark, D. Niyato and T. Chen-khong, "Evolutionary Optimal Virtual Machine Placement demand forecast for Cloud Computing," *IEEE*, pp. 348-355, 2011.

[10] G. Lee, N. Tolia, P. Ranganathan and R. H. Katz, "Toplogy-Aware Resource Allocation for Data-Intensive Workloads," *ACM*, pp. 120-124, 2010.

[11] J. J. Rao and K. V. Cornelio, "An Optimized Resource Allocation Approach for Data-Intensive workloads using Topology-Aware Resource Allocation," in *IEEE International Conference on Cloud Computing in Emerging Market(CCEM)*, Banglore, 2012.

[12] K. Tao, X. Zhang and H. Zhong, "An Approach to Optimized Resource Scheduling Algorithm for Open-Source Cloud Systems," in *2010 fifth annual ChinaGrid Conference*, Guangzhou, 2010.

[13] S. Chen, J. Wu and Z. Lu, "A Cloud Computing Resource Scheduling Policy Based on Genetic Algorithm with mulitiple fitness," in *Computer and Information Technology(CIT)*, Chengdu, 2012.

[14] M. Tang and S. Pan, "A Hybrid genetic Algorithm for the Energy-Efficien Virtual Machine Placement in Data Centers," *Springer*,

pp. 211-221, 2014.

[15] E. Apostol, I. Baluta and A. Gorgoi, "Efficient Manager for Virtualized resource provisioning in cloud systems," in *Intelligent Computer Communication and processing*, Cluj-Napaco, 2011.

[16] E. Feller, L. Rilling and C. Morin, "Energy-Aware Ant Colony Based Workload Placement in Clouds," in *The 12th IEEE/ACM International Conference on Grid Computing*, Lyon, France, 2011.

[17] E. Feller and C. Morin, "Autonomous and Energy Aware Management of Large Scale Cloud Infrastructure," in *Parallel and Distributed Processing Symposium Workshops and PhD Forum(IPDPSW)*, Shangai, 2012.

[18] Y. Gao, H. Guan, Z. Qi and Y. Hou, "A multi-objective ant colony system algorithm for virtual machine placement in cloud computing," *Journel of Computer and System Sciences, ELSEVIER*, pp. 1230-1242, 2013.

[19] X.-F. Liu, Z.-H. Zhan, K.-J. Du and W.-N. Chen, "Energy Aware Virtual Machine Placement Scheduling in Cloud Computing Based on Ant Colony Optimization Approach," in *Annual Conference on Genetic and Evolutionay Computation*, New York, USA, 2014.

[20] J. T. Christina, K. Chandrasekaran and C. Robin, "A Novel family Genetic Approach for Virtual Machine Allocation," *ELSEVIER*, pp. 558-565, 2015.