Review on Scheduling in Cloud Computing

Nora Almezeini and Alaaeldin Hafez,

King Saud University, College of Computer and Information Sciences, Riyadh, Saudi Arabia

Summary
Cloud computing has become an important and popular computing model that supports on demand services. It provides its services on pay-per-use basis. Using resources efficiently by reducing execution time and cost and increasing profit is the main goal of cloud service provider. Therefore, using effective scheduling algorithms is still main issue in cloud computing. There is no combined study of scheduling mechanism in cloud computing which describes its levels, policies, and types. This paper focuses on explaining the levels of scheduling in cloud system and presenting scheduling policies used when executing tasks. Also, it provides brief descriptions about various types of scheduling in cloud computing system.

Key words:
Cloud Computing, Scheduling, Task Scheduling, VM Scheduling

1. Introduction

Cloud computing has grown rapidly and gained considerable attention since it provides flexibility and scalability to organizations. Cloud computing is a large scale distributed system which offers a pool of computing resources to cloud consumers through the internet. There are many cloud providers which run on cloud computing environment such as Amazon, Google Engine, IBM, and Microsoft. They provide services and resources to users on the basis of pay per use at anytime from anywhere [1].

Cloud computing offers three main delivery models which are Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS). In Software as a Service (SaaS), Applications and access management tools are provided to users. Platform as a Service (PaaS) provides tools such as operating systems, databases, and network so consumers can install and develop their own software and applications. Infrastructure as a Service (IaaS) provides access to physical devices such as hardware and network so consumers can install and develop their own operating systems and applications [2].

Clouds in cloud computing is of several types based on the scalability and pooling up of the resources. Types are public, private, community, and hybrid clouds. Public clouds are available to the general public in a pay-as-you-go manner and they are owned by the cloud provider. Private clouds are operated only for a business or an organization and they are controlled by that organization or a third party. In community clouds, several organizations share the infrastructure of the cloud to support certain community that has common concerns. Hybrid clouds are combination of public, private, or community clouds [3].

Scheduling and allocation of resources and tasks are critical problems in cloud computing which many researches where carried out on it. Cloud providers must serve many consumers in cloud computing system. Therefore, scheduling is the major issue in establishing cloud computing system in order to reduce the execution time and the cost, thus maximizing resource utilization.

The main objective of this paper is to present an overview of scheduling mechanism in cloud computing system. It will discuss scheduling levels, policies, and types. The rest of the paper is organized as follows. Section 2 will explain the scheduling concept in cloud systems. Section 3 will discuss the scheduling policies. Section 4 differentiates several scheduling types in cloud computing. Finally, section 5 is the conclusion.

2. Scheduling in Cloud Computing

The concept of scheduling in cloud computing refers to the technique of mapping a set of jobs to a set of virtual machines (VMs) or allocating VMs to run on the available resources in order to fulfill users' demands [4]. The aim of using scheduling techniques in cloud environment is to improve system throughput and load balance, maximize the resource utilization, save energy, reduce costs, and minimize the total processing time. Therefore, the scheduler should consider the virtualized resources and users' required constraints to get efficient matching between jobs and resources [5]. Each scheduling technique should be based on one or more strategies. The most important strategies or objectives commonly that are used, are time, cost, energy, QoS, and fault tolerance [6].

Resources in cloud computing are scheduled at two levels [7]: VM-level and Host-level. At the VM-level, tasks are mapped for execution to the allocated VMs using a task/job scheduler. It is called Task Scheduling. At the host-level, a VM scheduler is used to allocate the VMs into

Manuscript received February 5, 2018
Manuscript revised February 20, 2018

physical hardware. This type is usually called VM Scheduling.

Task scheduling focuses on mapping tasks to appropriate VMs efficiently. Based on the task dependency, tasks can be classified as independent or dependent tasks [8]. The independent tasks have no dependencies with other tasks and have no priority order need to be followed during scheduling process. However, the dependent tasks have precedence order based on dependencies among the tasks and need to be followed during the scheduling process. Scheduling dependent tasks is called Workflow Scheduling. Fig.1 shows the levels of scheduling.

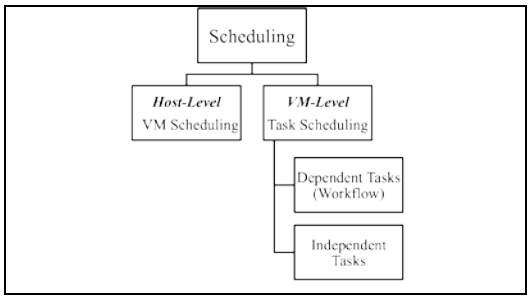

Fig. 1 Scheduling Levels

Deciding the appropriate task scheduling strategy in cloud environment is based heavily on task dependency. The main objective in scheduling the tasks is minimizing the makespan. If tasks are dependent, minimizing the makespan can be by decreasing the computation cost which is the time taken to execute each node, and the communication cost which is the time taken to transfer data between the two nodes. In contrast, if tasks are independent, they can be scheduled independently without any order [8].

VM scheduling is the allocating of VMs to run on the appropriate physical machines PMs to insure the implementation of tasks. This will improve the utilization of resources as well as managing the load balancing of all the systems. VM scheduling is important to ensure the Quality of Service (QoS) and Service Level Agreements (SLA) agreed by the cloud service providers and customers. The figure below shows the scheduling levels in cloud environment [5][9].

3. Scheduling Policies

There are two policies for scheduling in cloud computing: Space-Shared scheduling and Time-Shared scheduling as shown in Fig.2. Both policies are used for VM scheduling and Task Scheduling. In space-shared scheduling policy, only one VM/task is allowed to be executed at a given instance of a time on host/VM. In Time-Shared scheduling policy, it allows multiple VMs/tasks to multitask and run simultaneously within a host/VM. It shared the time among all VMs/tasks [10][11].

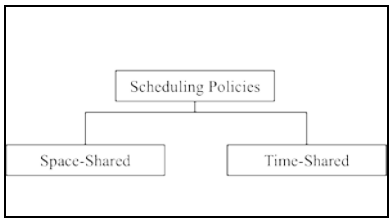

Fig. 2 Scheduling Policies

To understand the difference between these two policies and their impact on the application performance, suppose a host with two CPU cores for example hosts two VMs. Each VM require two cores and running four tasks. T1, T2, T3, and T4 will run on VM1, while T5, T6, T7, and T8 will run in VM2. Four cases show the use of scheduling policies in this example [10]:

1. Space-Shared policy is used for both VM scheduling and task scheduling. Since each VM requires two cores, only one VM can be assigned to the core in a specific time. So VM2 cannot run and use the cores until VM1 finishes. Also, each task hosted within the VM requires one core, so T1 and T2 will run simultaneously while T3 and T4 will wait until T1 and T2 are completed. The same happens for tasks running in VM2. Fig.3 describes this case.

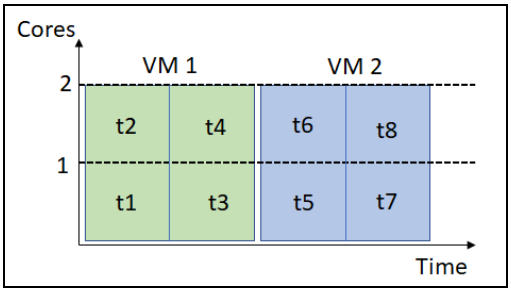

Fig. 3 Space-Shared for VMs and Tasks

2. Space-shared policy is used for VM scheduling, but time-shared policy is used for task scheduling. Hence, VM1 will be assigned first to the cores at a specific time and T1, T2, T3, and T4 are assigned to VM1 simultaneously. VM2 can run after VM1 finishes all the tasks. Then T5 to T8 are all assigned to it at the same time as shown in Fig.4.

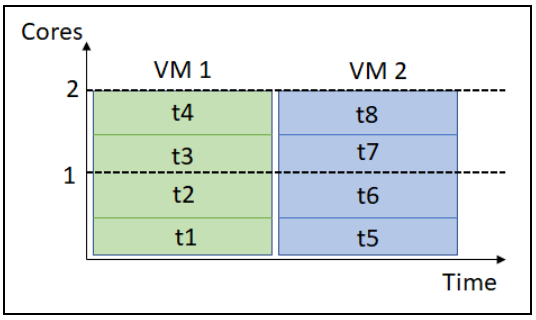

Fig. 4 Space-Shared for VMS, and Time- Shared for Tasks

3. Time-shared policy is used for VM scheduling, but space-shared policy is used for task scheduling. Hence, VM1 and VM2 shares a time slice of each core. Then each slice will be assigned only one task while others will wait until those tasks are completed as shown in Fig.5.

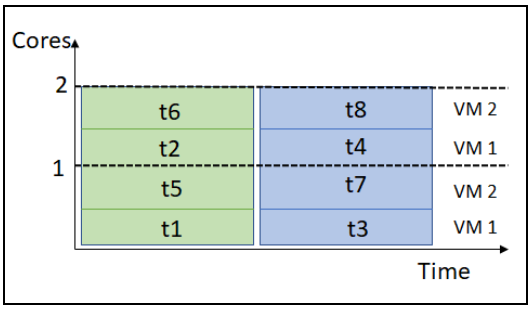

Fig. 5 Time-Shared for VMs, and Space-Shared for tasks

4. Time-shared policy is used for both VM and task scheduling. Thus, VM1 and VM2 shares a time slice of each core and all tasks share these slices simultaneously. Fig.6 presents this case.

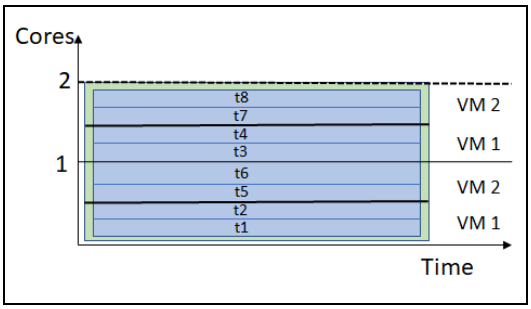

Fig. 6 Time-Shared for both VMs and Tasks

4. Scheduling Types

Different types and categories of scheduling in cloud computing system are shown in Fig.7.

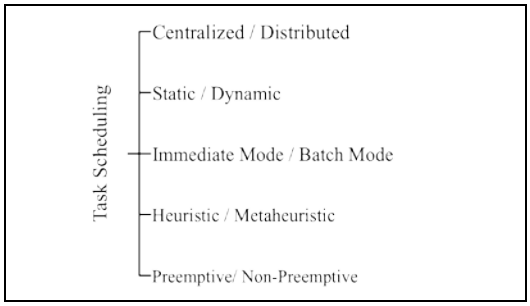

Fig. 7 Scheduling Types

These types are described in the following [12][13]:

4.1 Centralized / decentralized scheduling

When scheduling is centralized, decisions are made in a central node. It insures efficiency and ease of monitoring resources. However, it is lack of scalability and fault tolerance. Decentralized or distributed scheduling is more applied in real cloud environment although its lack of efficiency.

4.2 Static / Dynamic scheduling

In static scheduling, all timing information about tasks is available before so the execution schedule of each task is computed before executing any task. It is effective for applications that have fixed demands. Moreover, in static scheduling, the consumer makes agreement with the cloud provider for services and the cloud provider prepares the required resources before the start of required service [14].

In dynamic scheduling, the timing information about the tasks is not known at runtime. So the execution schedule of task may change as per the user demand. Dynamic scheduling incurs runtime overhead compared to static scheduling. In dynamic scheduling, the cloud provider cannot plan before usage. It allocates and remove resources as per needed [15].

4.3 Preemptive / Non-Preemptive scheduling

Preemptive scheduling allows interrupting each task during the execution and migrating the task to another resource. For example, when a task has a higher priority than another task and need to be executed although it is running in the

virtual machine. Therefore, this type of scheduling is mandatory if constraints need to be imposed such as priority, deadline, and cost.

In non-preemptive scheduling, the virtual machine cannot be taken away until the task running on it completes. It does not allow the task to be interrupted while it is executing.

4.4 Immediate mode / Batch mode scheduling

In immediate mode, called also online mode, tasks are scheduled to resources immediately without any delay. Tasks are scheduled only once and cannot be changed. On the other hand, the batch mode collects tasks into a set and are examined for mapping at prescheduled times. It is called also offline mode.

4.5 Heuristic / metaheuristic scheduling

Heuristic scheduling techniques are problem dependent that can solve specific problems. Metaheuristic techniques are high level problem-independent techniques that provide master strategies to solve general problems and can be applied to a wide range of problems. Both scheduling techniques do not guarantee that they will find the optimal solution, but they provide good solution when compared with the time spent on computation.

5. Conclusion

In this paper, we explained the concept of scheduling in cloud computing system and its levels. Also, we discussed the scheduling policies and how they affect on executing tasks. Moreover, various types of scheduling clouds have been analyzed.

Acknowledgments

This research project was supported by a grant from the Deanship of Graduate Studies, King Saud University, Saudi Arabia.

References

[1] D. C. Marinescu, Cloud computing: theory and practice. Newnes, 2013.
[2] N. Almezeini and A. Hafez, “An enhanced workflow scheduling algorithm in cloud computing,” in CLOSER 2016 - Proceedings of the 6th International Conference on Cloud Computing and Services Science, 2016, vol. 2.
[3] N. Antonopoulos and L. Gillam, Cloud Computing Principles, Systems and Applications, 2nd ed. Springer International Publishing, 2017.
[4] V. Manglani, A. Jain, and V. Prasad, “Task Scheduling in Cloud Computing.,” Int. J. Adv. Res. Comput. Sci., vol. 8, no. 3, 2017.
[5] L. Liu and Z. Qiu, “A survey on virtual machine scheduling in cloud computing,” in 2016 2nd IEEE International Conference on Computer and Communications (ICCC), 2016, pp. 2717–2721.
[6] D. P. Chandrashekar, “Robust and fault-tolerant scheduling for scientific workflows in cloud computing environments.” University of Melbourne, Australia, 2015.
[7] E. Pacini, C. Mateos, and C. G. Garino, “Multi-objective Swarm Intelligence schedulers for online scientific Clouds,” Computing, vol. 98, no. 5, pp. 495–522, 2016.
[8] R. A. J, A. B. W, and Shriram, “Article: A Taxonomy and Survey of Scheduling Algorithms in Cloud: Based on task dependency,” Int. J. Comput. Appl., vol. 82, no. 15, pp. 20–26, 2013.
[9] W. Kong, Y. Lei, and J. Ma, “Virtual machine resource scheduling algorithm for cloud computing based on auction mechanism,” Opt. - Int. J. Light Electron Opt., vol. 127, no. 12, pp. 5099–5104, 2016.
[10] H. S. Sidhu, “Comparative analysis of scheduling algorithms of Cloudsim in cloud computing,” 2014.
[11] P. J. Mudialba, “A Study on the Fundamental Properties, Features and Usage of Cloud Simulators,” in 2016 International Conference on Platform Technology and Service (PlatCon), 2016, pp. 1–5.
[12] A. P. Tikar, S. M. Jaybhaye, and G. R. Pathak, “A systematic review on scheduling types, methods and simulators in cloud computing system,” in 2015 International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), 2015, pp. 382–388.
[13] P. Banga and S. Rana, “Heuristic based Independent Task Scheduling Techniques in Cloud Computing: A Review,” Int. J. Comput. Appl., vol. 166, no. 1, pp. 27–32, 2017.
[14] Y. Chawla and M. Bhonsle, “A study on scheduling methods in cloud computing,” Int. J. Emerg. Trends Technol. Comput. Sci., vol. 1, no. 3, pp. 12–17, 2012.
[15] T. Ma, Y. Chu, L. Zhao, and O. Ankhbayar, “Resource Allocation and Scheduling in Cloud Computing: Policy and Algorithm,” IETE Tech. Rev., vol. 31, no. 1, pp. 4–16, Jan. 2014.