Cloud Scheduling - A Survey

D.I. George Amalarethinam, Ph.D
Associate Professor & Director – MCA
Department of Computer Science
Jamal Mohamed College, Trichy
Tamil Nadu, India

T.Lucia Agnes Beena
Assistant Professor,
Department of Information Technology,
St. Joseph's College, Trichy
Tamil Nadu, India

ABSTRACT

Cloud computing is a new technological trend that supports better utilization of IT infrastructure, services and applications. It follows pay-per-use approach for its services, in which user do not need to own infrastructure, platform or applications but use them at any time wherever they need them. The realization of Cloud computing has become a reality with the support of various paradigms like Distributed Systems, Virtualization, Web 2.0, Service-oriented Computing and Utility Computing. In all these paradigms, scheduling forms an important role. In the last few decades a lot of research has been devoted to scheduling starting from High Performance Computing facilities, still there are lot of issues related to scheduling with respect to cloud scenario. In this paper, existing scheduling algorithms such as online scheduling, cost-effective scheduling, workflow scheduling etc., for cloud scenario were compared. This survey shows the need for new algorithms with energy efficiency, bandwidth usage and power efficiency are to be considered for the better cloud services.

Keywords

Cloud Computing, Virtualization, Scheduling, Scheduling Algorithms

1. INTRODUCTION

The idea of providing computing as a utility like natural gas, water, power, and telephone connection has become a reality today with the advent of Cloud computing. Distributed computing is a foundational model for Cloud computing, because Cloud systems are distributed systems. Besides the distributed nature, the extreme dynamism of Cloud systems where the provisioning is done on demand constitutes the major challenge for engineers and developers. The solutions for on demand and dynamic scaling across the entire stack of computing is achieved by a) providing methods for renting computing power, storage and networking b) offering runtime environments designed for scalability and dynamic sizing and c) providing application services that mimics the behavior of desktop application [3].

1.1. Cloud Definition

Cloud computing is an evolving paradigm. The American National Institute of Standards and Technology(NIST)[1] proposed the definition of Cloud computing environment as *"Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction".* This definition highlights that anybody can use the cloud as per their requirement whenever needed from anywhere. Another definition by Buyya et al.[2] express the utility-oriented nature of Cloud computing, *"A cloud is a type of parallel and distributed system consisting of a collection of interconnected and virtualized computers that are dynamically provisioned and presented as one or more unified computing resources based on service-level agreements established through negotiation between the service provider and consumers".*

2. KEY TECHNOLOGIES SUPPORTING CLOUD COMPUTING

Cloud computing is helping enterprises, governments, public, private institutions and research organizations to easily extend their resources or shrink it as and when needed. This is made possible by the collaboration of the key technologies like Distributed systems, Virtualization, Web 2.0, Service-oriented Computing and Utility Computing [3].

2.1. Distributed Systems

The primary purpose of the distributed systems is to share resources and to utilize them better. This is true in the case of cloud computing, where resources like infrastructure, runtime environments and services are rented to users. Distributed systems often exhibit other properties such as *heterogeneity, openness, scalability, transparency, concurrency, continuous availability and independent failures.* Cloud also includes *scalability, concurrency and continuous availability*. The major milestones which led to Cloud computing are Mainframe computing, Cluster computing and Grid computing. Cloud computing is often considered the successor of grid computing. In reality, it involves all the major technologies. Computing clouds are deployed in large data centers hosted by a single organization that provides services to others. Clouds are characterized by the fact of having virtually infinite capacity, being tolerant to failures, and being always on, as in the case of mainframes. In many cases, the computing nodes that form the infrastructure of computing clouds are commodity machines, as in the case of clusters. The services made available by a cloud vendor are consumed on a pay-per-use basis, and clouds fully implement the utility vision introduced by grid computing.

2.2. Virtualization

Virtualization is 40 years old technique, but it is always limited by the technologies. Today virtualization has become a fundamental element of Cloud computing. Virtualization can be applied to Hardware, Runtime environments, Storage and networking. Virtualization allows for the appropriate degree of customization, security, isolation and manageability which are the basic needs for delivering IT services on demand. Cloud computing systems utilize hardware and programming language virtualization, server consolidation and virtual machine migration techniques. Hardware virtualization is applied for solutions in the Infrastructure-as-a-Service(IaaS) market segment. Programming language virtualization is a technology used in Platform-as-a-Service (PaaS). The server consolidation and virtual machine migration are applied to achieve isolation and manageability while delivering IT services.

2.3. Web 2.0

The Web is the primary interface through which cloud computing delivers its services. At present, Web 2.0 comprises a set of technologies such as XML, AJAX, Web services etc., that facilitate interactive information sharing, collaboration, user-centered design and application composition. This platform strongly supports the Cloud services which needs Rich Internet Applications (RIAs) to be developed for the wider public access.

2.4. Service-oriented Computing

Service-orientation is the reference model for Cloud computing systems. It is the logical way of organizing software systems to provide end users or other entities distributed over the network with services through published and discoverable interfaces. It introduces two important fundamental concepts, *Quality of Service (QoS) and Software as a Service (SaaS)* for Cloud computing. Quality of Service identifies response time, security attributes, transactional integrity, reliability, scalability, and availability as performance metrics to evaluate the behavior of a service from different perspectives. The SaaS approach allows the delivery of complex business processes and transactions as a service, while allowing applications to be composed on the fly and services to be reused from everywhere by anybody.

2.5. Utility-Oriented Computing

Utility computing is a vision of computing, defining a service provisioning model for computing services in which resources such as storage, computing power, applications and infrastructure are packaged and offered on a pay-per-use basis. The idea of providing computing as a utility like gas, water, power, and telephone connection has become reality today with the advent of Cloud computing. Computing grids provided a distributed computing infrastructure that was accessible on demand, which brought the concept of utility computing to a new level – market orientation. Also e-Commerce technologies provided the infrastructure support for utility computing. Service-orientated computing made computer system as a utility. All these factors contributed to the development of the concept of utility computing which forms the important step in the realization of Cloud computing.

3. CLOUD ARCHITECTURE

According to the definition of NIST [1], the cloud model is composed of *Five essential characteristics* in-demand self-service, broad network access, resource pooling, rapid elasticity, and measured service**,** *Three service models* Software-as-a-Service(SaaS), Platform-as-a-Service(PaaS) and Infrastructure-as-a-Service(IaaS), and *Four deployment models* Public clouds, private clouds, community clouds and hybrid clouds.

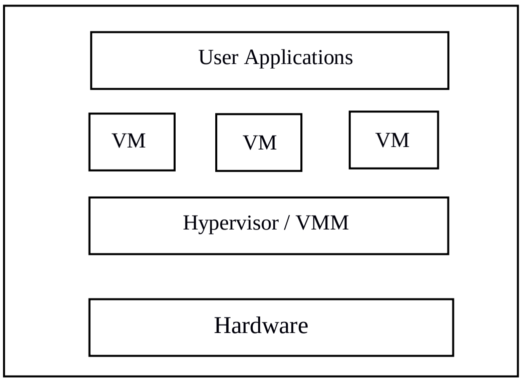

Figure -1 High-Level Cloud Architecture

The Figure-1 provides an overall view of the components of High-level architecture. At the top layer the user interface provides access to the services exposed by the Hypervisor or Virtual Machine Manager. Hypervisor is generally a program or a combination of software and hardware that allows the abstraction of the underlying hardware. There are three main modules Dispatcher, Scheduler and Interpreter Routines in the hypervisor [3].

The dispatcher constitutes the entry point of the monitor and reroutes the instructions issued by the virtual machine instances to one of the two other modules. The Scheduler is responsible for deciding the system resources to be provided to the Virtual Machine (VM). Whenever a virtual machine tries to execute an instruction that results in changing the machine resources associated with that VM, the Scheduler is invoked by the dispatcher. The interpreter module consists of interpreter routines. These are executed whenever a virtual machine executes a privileged instruction, a trap is triggered and the corresponding routine is executed. The central role is played by the Scheduler. The scheduler interacts with the other components such as pricing and billing, monitoring, reservation, QoS / Service Level Agreement (SLA) management, VM repository and VM pool manager to perform a variety of tasks. The bottom layer is composed of the physical infrastructure (processor, memory), on top of which the Virtual Machine Manager operates.

4. SCHEDULING

Cloud computing has a variety of characteristics such as Commercialization, Virtualization, Shared infrastructure, Dynamic Provisioning, Network access, Managed Metering, Self-service based usage mode, self-managed platform, Consumption-based billing, Resource pooling, Rapid elasticity and Multi Tenacity [4].

As Cloud computing is in the developing stage, researchers are interested in areas of Resource allocation and scheduling, security issues, Cloud storage and elastic scalability and programming models etc. The commercialization and virtualization properties include more complexity in the Resource allocation and scheduling. In the general scheduling problem, given a set of jobs and requirements, a set of resources and the system status, the algorithms should map the jobs to the appropriate resources. In Cloud computing, the factors such as economic considerations and efficient resource provisioning in terms of QoS guarantees, utilization and energy are playing a vital role [5]. Clouds provide the abstraction of nearly unlimited computing resources through the elastic use of consolidated resource pools. The challenge here is to how to dynamically allocate resources among VMs with the goal of optimizing a global utility function [5].

4.1. Types of Scheduling

The on-demand service of cloud leads to the need for new scheduling strategies. The new scheduling strategies to be proposed should combine the traditional scheduling concepts with new scheduling parameters such as bandwidth, energy consumption, job migration and cost for efficient scheduling. The following are the areas where new strategies can be proposed using various optimization techniques, machine learning techniques or fuzzy systems.

4.1.1. Qos based Scheduling

QoS is the collective effect of performance which determines the degree of satisfaction of a user for the service. Commonly QoS is expressed by the qualitative measures such as completion time, latency, execution price, packet loss rate, throughput and reliability. Based on these qualitative measures developing a new scheduling algorithm is a challenging problem [6].

4.1.2. Online Scheduling

The goal is to compute a schedule that specifies when and on which machine each job is to be executed. In online scheduling, the scheduler receives jobs that arrive over time, and generally must schedule the jobs without any knowledge of the future [7].

4.1.3. Resource Scheduling

The key research for cloud computing is the process of the work scheduling and resource allocation. It is mainly about how the computing resources are virtualized and with scheduler how to integrate the resources in the logical way, to focus on how to deal with data center resources virtualization, to satisfy the user needs and to maximum resource utilization rate for the service providers [8].

4.1.4. Cost-effective scheduling

Cloud users pay for what their programs actually use according to the pricing models of the cloud providers. Early task scheduling algorithms are focused on minimizing makespan, without mechanisms to reduce the monetary cost incurred in the setting of clouds. This is the new challenge to design algorithm for cost-effective scheduling along with minimum makespan [9].

4.1.5. Workflow scheduling

One of the most challenging problems in Clouds is workflow scheduling, i.e., the problem of satisfying the QoS of the user as well as minimizing the cost of workflow execution. Workflow scheduling is the problem of mapping each task to a suitable resource and of ordering the tasks on each resource to satisfy some performance criterion. The traditional scheduling methods try to minimize the execution time (makespan) of the workflows. However, in Clouds, there are many other potential QoS attributes besides execution time, like reliability, security, availability and so on. Due to the complexities of the development of a general multi-objective scheduling algorithm, many researchers try to propose bi-criteria scheduling algorithms [10].

4.1.6. Load balancing

To gain the maximum benefit from cloud computing, developers must design mechanisms that optimize the use of architectural and deployment paradigms. The role of Virtual Machine's (VMs) has emerged as an important issue because, through virtualization technology, it makes cloud computing infrastructures to be scalable. Therefore developing on optimal scheduling of virtual machines is a series issue. So, efficient algorithms are needed for task scheduling in the cloud environment with the goal of putting unused resource (virtual machines) cycles to work and distributing the load about them [11][12].

4.1.7. Capacity planning

As utility computing resources become more ubiquitous, service providers increasingly look to the cloud for an in-full or in-part infrastructure to serve utility computing customers on demand. Existing cloud computing provisioning models explore the capacity-planning problem from the service provider perspective. Client-orchestrated cloud provisioning policies with real-time demand needs are to be explored either assuming advanced resource reservation or unknown resource reservation by users [13].

4.1.8. Bandwidth-aware scheduling

Task scheduling is a fundamental issue in achieving high efficiency in cloud computing. Most existing task-scheduling methods of cloud computing only consider task resource requirements for CPU and memory, without considering bandwidth requirements. In order to obtain better performance, it is a big challenge to propose a bandwidth-aware algorithm for task scheduling in cloud computing environments [14].

4.1.9. Energy-aware scheduling

With the rapid advance of cloud computing, large scale data center plays a key role in cloud computing. Energy consumption of such distributed systems has become a prominent problem and received much attention. Among existing energy-saving methods, application scheduling can reduce energy consumption by replacing and consolidating applications to decrease the number of running servers. However, most application scheduling approaches did not consider the energy cost on network devices, which is also a big portion of power consumption in large data centers. Scheduling Algorithm for applications to minimize the energy consumption of both servers and network devices can be developed [15].

4.1.10. Gang scheduling

Gang Scheduling is an efficient job scheduling algorithm for time sharing, which is applied in parallel and distributed systems. Thus, each job requires a number of processors equal to its degree of parallelism, based on the number of tasks that should be dispatched and executed. In Cloud scenario, the use of job migration along with variable workloads, job sizes and types must be considered to better fit a real High Performance Computing into cloud computing implementation [16].

Table 1: Existing Scheduling algorithms in the Cloud Computing Environment

Scheduling Algorithm	Scheduling Parameters	Method used for optimization	Scheduling Parameters for Future enhancement
A task scheduling algorithm based on QoS-driven in Cloud Computing [6]	Priority, Completion time	Heuristic method	User privilege, expected task length and pending time
Cloud Scheduling with Setup Cost [7]	Total cost, Task delay (response time)	Heuristic method – ratio maths	Computation Time
Study on Cloud Computing Resource scheduling strategy based on the Ant Colony Optimization Algorithm [8]	Makespan	Ant Colony Optimization	Service Cost, Bandwidth, Reliability & Service completion time
Cost-efficient task scheduling for executing large programs in the cloud [9]	Makespan, Monetary cost(different pricing models)	Heuristic method	Penalty cost (for violating customer-provider contracts)
Deadline-constrained workflow scheduling in Software as a Service Cloud [10]	User deadline, Total Execution Cost	Partial Critical Paths	Different pricing models
Scheduling Virtual Machines for Load balancing in Cloud Computing Platform[11]	Response time, Processing time	Weighted Round Robin	User deadline
Intelligent Strategy of Task Scheduling in Cloud Computing for Load Balancing [12]	Load balance, Makespan	Ant Colony Optimization	Task precedence, Cost of resources
Scheduling Cloud Capacity for Time- Varying Customer Demand[13]	Blocking probability, Total unused seats(constant reliability)	Erlang-loss Queuing model (statistical prediction technique & heuristic based technique)	Reliability
Bandwidth-aware divisible task scheduling for Cloud Computing [14]	Makespan (CPU power, memory, bandwidth)	Non-linear programming	Energy consumption
Energy-aware Hierarchical Scheduling of Applications in Large Scale Data Centers [15]	Energy consumption of servers & network devices, server stability	Hierarchical scheduling method	Energy consumption with respect to cooling conditions
Evaluation of gang scheduling performance and cost in a Cloud Computing system [16]	Performance (response time, waiting time & delay), Overall Cost	Adaptive First Come First Serve (AFCFS) and Largest Job First Served (LJFS)	Job migration with variable workload, job size & job types
Efficient Optimal Algorithm of Task Scheduling in Cloud Computing Environment [18]	Makespan	Heuristic method - Generalized priority Algorithm	Large tasks with makespan

4.2. Comparative study of Cloud scheduling algorithms

Table 1 summarizes few existing scheduling algorithms in cloud computing environment. The survey made in this paper shed light on the areas where new algorithms are to be designed for the better performance of the cloud services.

Cloud operates on a pay-per use basis. There are various pricing models to support this. So the scheduling algorithms have to concentrate on the pricing policies with the other parameters. Cloud also has the deployment models such as Private, Public, Community and Hybrid. The scheduling algorithm developed for one deployment model may not be efficient for the other model. For example, parameters such as bandwidth, communication cost vary with the deployment model. The communication delay affects the makespan, which indirectly increases the cost. Therefore, there is a need for Bandwidth –aware scheduling algorithms.

Traditional scheduling algorithms had only one or two parameters to be optimized. But cloud scenario needs new algorithms with multiple criteria to be optimized to satisfy the user requirements. Also traditional job schedulers are unable to efficiently schedule workloads combining best-effort jobs and advanced reservation policies. Algorithms supporting advanced reservation with best-effort jobs can be proposed.

The high performance service is achieved through clusters of virtual machines, where applications have to be migrated to different locations in the clusters or in the cloud. Here virtual machine scheduling algorithms with cross-site load balancing can be proposed.

There are various approaches to solve the scheduling problems. They are Operation Research/ Mathematical techniques, Heuristic approaches, and Machine learning techniques. Advanced algorithms such as Genetic algorithms, Neural Networks, Simulated Annealing, Ant-Colony optimization, Cuckoo search and Fuzzy systems also can be applied. As the scheduling problem is a NP-Complete [17] in nature, most of the scheduling algorithms utilize heuristics or optimization techniques to get a sub-optimal solution. From history, it was observed that the hybrid models provide better solutions. Form this discussion, new algorithms based on cloud pricing models, energy consumption and job migrations can be developed. Also cloud deployment models raise the need for the algorithms with new parameters such as bandwidth, reliability and energy consumption. Hence future research on cloud scheduling should concentrate on better scheduling methods to enhance the performance of Cloud Computing system.

5. PROPOSED RESEARCH WORK

The popularity of cloud computing is due to its real time services. Virtualization process helps to overcome failures and supports any operating system. The attractive on-demand flexible service to the user, avoidance of initial investments, low pricing policies and reduction in maintenance cost also increases its growth. To enhance these features one of the important area needs special concentration is scheduling of resources. Classical scheduling algorithms to be revived with new parameters are identified in this survey. They are

- ✓ QoS – driven scheduling with user privilege, expected task length and pending time
- ✓ Resource scheduling strategy based on service cost, bandwidth, reliability and service completion time
- ✓ Cost based scheduling including penalty cost for different pricing models
- ✓ Cloud capacity scheduling with reliability
- ✓ Energy aware scheduling with cooling conditions
- ✓ Job scheduling through job migration considering variable workload, job size and job types

Though designing new algorithms with all these parameters increase the complexity, simple algorithms for different service models with selected parameters can be implemented to accomplish the full benefit of the Cloud computing environment. Giving importance to energy conservation increase the potential of Cloud computing in the current scenario.

6. CONCLUSION

The development of new technologies and increased familiarity of Cloud computing will lead to the establishment of a global market for trading computing utilities. Thus the vision of Cloud Computing XaaS – anything as a service can be realized. To achieve this, the user satisfaction is important, for which good scheduling framework has to be developed. The scheduling issues discussed in this paper encourage the researchers to propose new algorithms with multiple parameters for the better performance of the Cloud. In future, workflow scheduling for different pricing models that satisfies the user specified deadlines with minimum cost in IaaS will be implemented.

7. REFERENCES

[1] Peter Mell Timothy Grance , “The NIST Definition of Cloud Computing”, NIST Special Publication 800-145, September 2011.

[2] Buyya R, Yeo CS, Venugopal S., “Market oriented cloud computing: vision, hype, and reality for delivering IT services as computing utilities”, Proceedings of the tenth conference on high performance computing and communications (HPCC 2008, IEEE Press, Los Alamitos, CA). Dalian, China, pp. 5-13, 2008.

[3] Rajkumar Buyya, Christian Vecchiola, S. Thamarai Selvi, “Mastering Cloud Computing Foundations and Applications Programming”, Elsevier Publications, 2013.

[4] Vijindra, Sudhir Shenai, “Survey on Scheduling Issues in Cloud computing”, ICMOC, Elsevier Publications, Vol 38, pp. 2881- 2888, 2012.

[5] David Villegas, Ivan Rodero, Liana Fong, Norman Bobroff, Yanbin Liu, Manish Parashar, and S. Masoud Sadjadi, “The Role of Grid Computing Technologies in Cloud Computing”, Handbook of Cloud Computing, Springer Publications, pp. 183- 218, 2010.

[6] Xiaonian Wu, Mengqing Deng, Runlian Zhang, Bing Zeng, Shengyuan Zhou, “ A task scheduling algorithm based on QoS-driven in Cloud Computing”, Elsevier Publications, Vol 17, pp. 1162- 1169, 2013.

[7] Yossi Azar, Naama Ben-Aroya, Nikhil R. Devanur, Navendu Jain, “Cloud Scheduling with Setup Cost”, Preceedings of the 25th ACM symbosium on parallelism in algorithms and architecture pp. 298-304, 2013.

[8] Linan Zhu, Qingshui Li, Lingna He, “Study on Cloud Computing Resource Scheduling Strategy Based on the Ant Colony Optimization Algorithm”, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, pp. 54 – 58, September 2012.

[9] Sen Su, Jian Li, Qingjia Huang, Xiao Huang, Kai Shuang, Jie Wang, “Cost-efficient task scheduling for executing large programs in the cloud”, Parallel Computing Elsevier Publications, vol 39, issue 4-5, pp. 177–188, 2013.

[10] S. Abrishami, M. Naghibzadeh, “Deadline-constrained workflow scheduling in software as a service Cloud”, Computer Science & Engineering and Electrical Engineering, Scientia Iranica, vol 19, issue 3, pp. 680–689, 2012.

[11] Supreeth S, Shobha Biradar, “Scheduling Virtual Machines for Load balancing in Cloud Computing Platform”, International Journal of Science and Research (IJSR), India, Vol 2, Issue 6, pp. 437 – 441, June 2013.

[12] Arabi E. keshk, Ashraf El-Sisi, Medhat A. Tawfeek, F. A. Torkey, “ Intelligent Strategy of Task Scheduling in Cloud Computing for Load Balancing”, International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), Vol 2, Issue 6, , pp. 12 – 22, November – December 2013.

[13] Brian Bouterse, Harry Perros, “Scheduling Cloud Capacity for Time - Varying Customer Demand”, 1st international conference on Cloud Networking, IEEE Publications, pp.137-142, 2012.

[14] Weiwei Lin, Chen Liang, James Z. Wang, Rajkumar Buyya, “Bandwidth-aware divisible task scheduling for cloud computing”, SOFTWARE-PRACTICE AND EXPERIENCE, John Wiley & Sons Library, Vol 44, Issue 2, pp. 163–174, 2012.

[15] Gaojin Wen, Jue Hong, Chengzhong Xu, Pavan Balaji, Shengzhong Feng, Pingchuang Jiang, “Energy-aware Hierarchical Scheduling of Applications in Large Scale Data Centers”, International Conference on Cloud and Service Computing, IEEE Publications, pp. 158-165, 2011.

[16] Ioannis A. Moschakis · Helen D. Karatza, “Evaluation of gang scheduling performance and cost in a cloud computing system”, Springer Publications, Vol 59, issue 2, pp. 975–992, 2010.

[17] J. Ullman, “NP-complete scheduling problems”, Journal of Computer and System Sciences, Vol 10, issue 3, pp. 384–393, June 1975.

[18] Dr. Amit Agarwal, Saloni Jain, “Efficient Optimal Algorithm of Task Scheduling in Cloud Computing Environment”, International Journal of Computer Trends and Technology (IJCTT), Vol 9, No 7, pp. 344 – 349, March 2014.