A Review Paper on various scheduling techniques in cloud computing

Navdeep kaur
M.Tech. Student, Department of Computer Engineering, Rayat& Bahra institute of engineering & biotechnology(RBIEBT),
Mohali, India

Mandeep Kaur, *Assistant Professor, Department of ComputerEngineering,* Rayat& Bahra institute of engineering & biotechnology (RBIEBT),
Mohali, India

***Abstract*:Cloud Computing is the nascent technology which is based on pay-per-use model. It is computing paradigm where applications, data, bandwidth and IT services are provided over the Internet. Goal of the cloud service providers to use resource efficiently and attain the maximum profit. Every cloud service provider try to make best use of work resources & earns profit as much as possible. So, this leads to scheduling as a challenging issue in cloud computing. Scheduling is the process of deciding how to manage or arrange resources between a varieties of possible tasks.Scheduling is the process of arranging, controlling and optimizing work and workloads in a production process or manufacturing process. In this research paper various types of scheduling algorithms that provide efficient cloud services have been analysed . Based on the study of different algorithms, a comparison between them are presented on the basis of different perspective**

Keywords: cloud computing, scheduling, service providers, scheduling algorithms, resources.

I. INTRODUCTION

After wide discussion in IT industry, cloud computing get originated via computer network structure which represent vast internet connection as a cloud. Utmost IT Companies and market investigation firms such as IBM, Sun Microsystems, Forrester Research and Gartner had presented a whitepapers to explain the meaning of cloud computing and finally they invent a common definition that covers agreed aspects of cloud computing. The US NIST (National Institute of Standards and Technology) working definition summaries cloud computing as following:

"A model for enabling convenient and on-demand network access to a share pool of configurable computing resources (e.g., networks, applications, storage, servers and services) which can be rapidly provisioned and released with minimal management effort or service provider interaction."

The NIST [10] definition is one of the clearest and most comprehensive definitions of cloud computing and is widely referenced in US government documents and projects. This definition consists of five essential characteristics, four deployment models and three service models. The important characteristics are as following:

- On-demand self-service: Computing resources can be gathered and used at anytime without the need for manual interaction with cloud service providers. Computing resources include, storage, virtual machines, processing power etc.
- *Broad network access*: The available resources can be accessed over a network using heterogeneous devices such as laptops or mobiles phones.
- *Resource pooling*: Cloud service providers pool their resources so that they can share resources by multiple users. This is also referred as multi-tenancywhere for example a physical server may host several virtual machines belonging to different users.
- *Rapid elasticity:* A user can quickly obtain more resources by scaling out from the cloud and also they can scale back in by releasing those resources once they are no longer required.
- *Measured service*: Resource usage is measured using appropriate metrics such monitoring storage usage, CPU hours, bandwidth usage etc.

The above characteristics can be applied to all cloud but each cloud provides users with services according to abstraction at variance levels, which is referred to as a service model under NIST definition. Three most common service models are as following:

- *Software as a Service (SaaS):* This is where users simply access of a web-browser to access software that others have developed and offer as a service over the web. At SaaS level, users do not have control to the underlying infrastructure being used to host the software. The Sales force's Customer Relationship Management software3 and Google Docs4 are popular examples that use the SaaS model of cloud computing.
- *Platform as a Service (PaaS):* This is where applications are developed using a set of programming languages and tools that are supported by the PaaS provider. PaaS imparts users with a high level of abstraction that allows

them to focus on developing their applications and not worry about the underlying infrastructure. Similar to the SaaS model, users do not have control or access to the underlying infrastructure being used to host their applications at the PaaS level. Google App Engine5 and Microsoft Azure6 are popular PaaS examples.

- *Infrastructure as a Service (IaaS*): This is where users acquire computing resources such as memory, storage and processing power from an IaaS provider and use the resources to deploy and run their applications. As compares to the PaaS model, the IaaS model is a low level of abstraction that allows users to access the underlying infrastructure through the use of virtual machines. IaaS gives users more flexibility than PaaS as it allows the user to deploy any software stack on top of thoperating system. However, flexibility depends on a cost and users are responsible for updating and patching the operating system at the IaaS level. Amazon Web Services' EC2 and S37 are popular IaaS examples.

Software as a Service as the core concept behind cloud computing, suggesting that it does not matter whether the software being delivered is an infrastructure, application or platform, there is always software in the end. Although this is true to some extent, it nevertheless helps to categorize between the types of service being delivered as they have different abstraction levels. The service models described in the NIST definition are deployed in clouds, but there are different types of clouds depending on who owns and uses them. This is referred to as a cloud deployment model in the NIST definition and the four common models are:

- *Private cloud*: A cloud that is used exclusively by one organization. The cloud may be accessed by the organization itself or a third party. The St Andrews Cloud Computing Co-laboratory8 and Concur Technologies are example organization that has private clouds.

- *Public cloud:* A cloud that can be used (for a fee) by the public. Public clouds require significant investment and are usually owned by large corporations such as Microsoft, Google or Amazon.

- *Community cloud*: A cloud that is accessed by several organizations through sharing and is usually setup for their particular requirements. The Open Cirrus cloud test hub could be required as a community cloud that aims to support research in cloud computing.

- *Hybrid cloud:* A cloud that is setup using a mixture of the above three deployment models. Each cloud in a hybrid cloud could be independently managed but applications and data would be allowed to move across the hybrid cloud. Hybrid clouds allow cloud erupting to take place, which takes place where a private cloud can burst-out to a public cloud when it requires more resources.

Figure 1 illustrates an overview of the common deployment and service models in cloud computing, where the three service models could be deployed on top of any of the four deployment models.

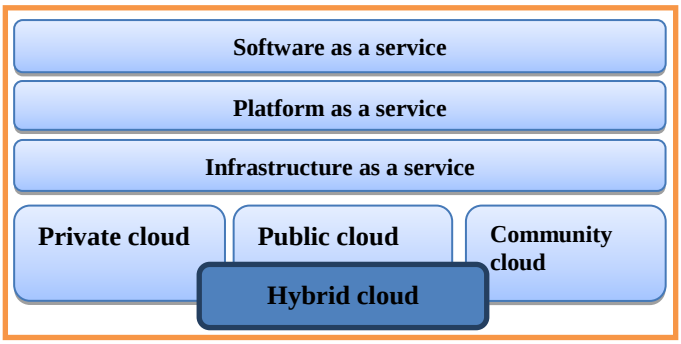

Figure 1. Deployment Models of Cloud Computing

In Cloud computing, available deployment models are:

- *Public Cloud*: Public cloud allows users to access the cloud publicly. It is access by interfaces using internet browsers. Users pay only for that time duration in which they use the service, i.e., pay-per-use.

- *Private Cloud:* A private clouds operation is within an organization's internal enterprise data center. The main advantage here is that it is very easier to manage security in public cloud. Example of private cloud in our daily life is intranet.

- *Hybrid Cloud:* It is a combination of public cloud and private cloud. It provides more secure way to control all data and applications .It allows the party to access information over the internet. It allows the organization to serve its needs in the private cloud and if some occasional need occurs it asks the public cloud for some computing resources.

- *Community Cloud:*When cloud infrastructure construct by many organizations jointly, such cloud model is called as a community cloud. The cloud infrastructure could be hosted by a third-part provider or within one of the organizations in the community

II. SCHEDULING ISSUES

Cloud computing consists of applications, platforms and infrastructure segments. Each segment performs different operations and offers different products for businesses

and individuals around the world. It suffers from various scheduling issues. Some of them are:

(1) Each virtual machine instance needs to be assigned and can only be assigned to one matching virtual machine instance vacancy.

(2) If two virtual machine instances using the same hardware resources, there should have some time interval between two usages. Because after the running of a virtual machine is completed, time is needed to process some operations like data clean and virtual machine restart.

(3) Virtual machine instance needs and the type of virtual machine instance vacancies should match each other.

(4) When large or medium vacancy is filled by small instances, we can release some new vacancies according to the proportion. For example, after a small instance fills a large vacancy, a medium-sized vacancy and two small vacancies can be produced.

III. SCHEDULING TECHNIQUES

As cloud computing has become more and more important, new scheduling systems have designed such as:

*(1) Round Robin Scheduling:*Round-robin (RR) is one of the algorithms employed by process and network schedulers in computing. As the term is generally used, time slices are assigned to each process in equal portions and in circular order, handling all processes without priority (also known as cyclic executive).

*(2) First Come First Serve Algorithm:*First-Come-First-Served algorithm is the simplest scheduling algorithm. Processes are dispatched according to their arrival time on the ready queue. If the required resource is unavailable, then the system simply waits for availability whereas our algorithm would give the resource in parts or simply put the request in a wait queue and see if the next request can be serviced. It follows a dynamic allocation towards deadline constraint or cost constraint depending on current usage. It then proceeds to allocate data to requests based on whichever category the request would fit into.

*(3) Worst Fit Algorithm:*In case of Worst Fit algorithm hosts in each list in datacenter are sorted in descending order according to remaining capacity of resources. When request of a new VM or already running VM for VM placement arrives at the cloud data center, VM scheduler finds the appropriate list and apply the binary search on the selected list to find host that is the worst fit in remaining resource capacity than the VM requirement capacity in all dimensions. If first host does not satisfy the resource requirement in all dimensions, power ON the new PM, allocate the VM on the new PM.

The update procedure is same for both the algorithms.

According to the algorithm 4 the worst case time complexity of West Fit algorithm for m number of VMs will be ($mlogn$), where n is the number of physical machines.

*(4) Best Fit Model:*In the Best Fit algorithm hosts in each list in datacenter are sorted in ascending order according to remaining capacity of resources. When request of a new VM or already running VM for VM placement arrives at the cloud data center, VM scheduler find the appropriate list and apply the binary search on the selected list to find host that is the best fit in remaining resource capacity than the VM requirement capacity in all dimensions. If no such host is available, Power ON the new physical machine and assign the VM on that PM. The Best Fit allocation algorithm calls the binary search procedure for searching the best fit satisfying host. The binary search procedure return the best fit host in remaining capacity for VM requirement capacity to the Best Fit algorithm as choosen host. Finding best fit host for VM reduces the required number of physical machine by fully utilization of the resources. The binary search strategy for searching the best fit host will reduce the allocation time of one VM in order of O($logn$).

(5) MAPE-K loop: In this algorithm the hosts are classified according to their resource availability. In general there are three types of performance parameters of any system i.e. CPU, B/W and Memory. The allocation of the VM is typically done by Mape-K loop in which the sensor to sense the VM status and create a plan

Algorithm	Description	Parameter	Tool
Independent task scheduling in cloud computing by improved genetic algorithm	In the normal genetic algorithm the initial population is generated randomly,so the diff schedules are further mutated with each other,there are very much less chances that they will produce better child than themselves. In an improved genetic algorithm, the idea for generating initial population by using the Min-Min & Max-Min techniques for genetic algorithms.	VM's & cloudlets	Cloudsim
The study of genetic algorithm-based task scheduling for cloud computing	In the proposed model,the task scheduler calls the genetic algorithm scheduling function for every task scheduling cycle. This function create a set of task schedules & evaluates the quality of each task schedule with user satisfaction & VM availability. The function iterates genetic opn's to make an optimal	Throughput,simulation time,average VM utilization,average response time,average processing cost & no. of tasks	GA-based task scheduling model
Dynamic scheduling of data using genetic algorithm in cloud computing	In dynamic scheduling task arrival is uncertain at run time& allocating resources are tedious as all task arrive at the same time. To avoid this genetic algorithm is used. Genetic algo is a heuristic method that deals with the natural selection of solution from all possible sol's . using geneticalgo the tasks are scheduled according to the computation & memory usage. This way tasks are scheduled dynamically. The execution time is also reduced by parallel processing.	Time utilisation & resource utilisation	Ubuntu enterprise cloud
Task scheduling optimization for the cloud computing systems	Describes & evaluate fuzzy sets to model imprecise scheduling parameters & also to improve satisfaction grades of each objective. Genetic algorithms with diff. components are developed on the based technique for task level scheduling in hadoop mapreduce.	Flexibility , virtualisation	Not implemented
Impatient task mapping in elastic cloud using genetic algorithm	The algorithm proposes that can find a fast mapping using genetic algorithms with "exist if satisfy" condition to speed up the mapping process & ensures the respecting of all task deadlines.	No. of jobs, time	cloudsim
A genetic algorithm for workload scheduling in cloud based e-learning	The paper presents the characteristics of a private cloud used for e-learning purposes along with a genetic algorithm that optimizes the scheduling of the e-learning workloads according to a set of conditions that are imposed by the underlying virtualization technology such as memory over-commitment & IOPS rate.	Load distribution for windows, CPU intensive, IO intensive	IBM cloudburst system

to allocate job requests to VMs.

TABLE 1. DIFFERENT REAL WORLD TECHNIQUES

However after allocation a large number of resources remain under utilized.The algorithm shows the total remaining underutilized resources is accumulated and is allocated to next requests. Let us assume that when request for VM arrived to scheduler place this VM on the satisfying host in the list in which all host have remaining resource capacity in order. In other words if a VM having CPU requirement greater than or equal to Memory requirement and Memory requirement greater than or equal to Bandwidth is placed on the satisfying host in the list, in which all host have remaining resource capacity of CPU greater than or equal to Memory and Memory greater than or equal to Bandwidth.

IV. CONCLUSION

In cloud computing environment, number of different resources are provided as a service in the form of virtual machines and these machines are scheduled by scheduling algorithm. Scheduling is the key issue in the management of application execution in cloud environment.

In this paper, existing scheduling algorithm are considered and they all are compared by using different parameters as well as tools. Mostly they all are work on to minimize the execution time, faster response time and maximum utilization of resources. Existing scheduling algorithms does not consider the load balancing, availability and reliability. Therefore, there is a need to implement such scheduling algorithm that can improve the reliability, availability and load balancing in cloud computing environment. In future, algorithm based on migration of task from one machine to another can also be introduced.

REFERENCES

[1] Joerg Fritsch and et.al. represents "A lightweighted asynchoronous high-performance message queue in cloud computing "*Journal of cloud computing: advances, system and applications"*, 2012.

[2] S. Pronk and et al., "Copernicus: A new paradigm for parallel adaptive molecular dynamics," in Proceedings of the 2011 ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis, November 2011.

[3] Jin and Loques M, "A dynamic optimization model for power and performance management of virtualized cluster",published in ACM, pp.225-233,2010.

[4] RamMohan. " The impact of virtualization in cloud computing" in IEEE INFOCOM proceedings pp1-9,march,2010.

[5] Micheal Maurer, Ivona Brandic et. al., "Self adaptive and resource- efficient SLA Enactment for Cloud Computing Infrastructure published their white papers,in 2010.

[6] G. Juve, E. Deelman, K. Vahi, G. Mehta, B. Berriman, B. Berman, and P. Maechling, "Scientific workflow applications on amazon ec2," in *E-Science Workshops, 2009 5th IEEE International Conference on*, December 2009, pp. 59–66.

[7] C. J. Woods and et al., "Grid computing and biomolecular simulation," *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*, vol. 363, pp. 2017–2035, 2009.

[8] Hu Wu, Zhuo Tang, Renfa Li, "A Priority Constrained Scheduling Strategy of Multiple Workflows for Cloud Computing," 2008.

[9] C. Hoffa, G. Mehta, T. Freeman, E. Deelman, K. Keahey, B.Berriman, and J. Good, "On the use of cloud computing forscientific workflows," in *eScience, 2008. eScience '08. IEEEFourth International Conference on*, 2008, pp. 640 – 645.

[10] A. Luckow and et al., "Distributed replica-exchange simulationson production environments using saga and migol," in *IEEEFourth International Conference on eScience, 2008*, December 2008, pp. 253–260.

[11] Lin I, Zhao Y, Raicer presents "Grid Computing Environment"in GCE (2008), pp 1-10.

[12] Rodrigo N.Calheiros and et. al., " A novel framework formodeling and simulations of cloud computing infrastructure and services", published in *Journal of Computing and InformationTechnology*, CIT-16-2008, 4-235-246.

[13] Lizheng, white paper on "An adaptive interface to scalable cloud storage,onmay 2008.

[14] Hu Wu, Zhuo Tang, Renfa Li, " A novel framework for modeling and simulations of cloud computing infrastructure and services", published in *Journal of Computing and Information Technology*, CIT-16-2008, 4-235-246.

[15] M. Joseph and P.Pandya, " Finding Response time of message queue in a real-time system, *BCS Computer Journal. 29(J):390-395*, 2002.

[16] A. Natrajan, M. Crowley, N. Wilkins-Diehr, M. Humphrey, A. Fox, A. Grimshaw, and I. Brooks, C.L., "Studying protein folding on the grid: Experiences using charmm on npaci resources under legion," in *Proceedings of the 10th IEEE International Symposium on High Performance Distributed Computing*, 2001, pp. 14–21.

[17] Kahina, "An adaptive interface to scalable cloud storage", In preceedings of Wiley Online Library

[18] Minxia K, Chockler G, Van Renesse R, "Toward a cloud computing research agenda" published in SIGACT News 40(2): 68-80.http://doi.acm.org/1556154.1556172.