DOI: http://dx.doi.org/10.26483/ijarcs.v8i8.4752

ISSN No. 0976-5697

Volume 8, No. 8, September-October 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

CHALLENGES TO TASK AND WORKFLOW SCHEDULING IN CLOUD ENVIRONMENT

Er. Amanpreet Kaur
Research Scholar, IKG PTU, Jalandhar

Dr. Bikrampal Kaur
CEC, Landran(Mohali)

Dr. Dheerendra Singh
CCET, Chandigarh

***Abstract*:** Task scheduling and workflow scheduling are two paradigms in cloud computing which differ in the extent of data involved. Workflow deals with huge scientific or business data patterns while task corresponds to a single job comprising customer or service provider application. Both requires efficient resource provisioning and utilization

In this paper, the process involved from application submission to its completion involving different phases has been discussion thoroughly along with the challenges faced by both types scheduling.

Keywords - Task Scheduling, Workflow Scheduling, Resource Provisioning, Workload, QoS.

1. INTRODUCTION

Cloud computing is the dynamic collection of heterogeneous resources including memory, storage, network bandwidth, computational power and application development software. These resources provide scientific, engineering and business applications services to its customers. Cloud datacenters are rich in versatile computing resources satisfying the needs of both cloud service provider and service consumer. In addition, cloud services also involve dynamically provisioning the sharable resources among the user request (applications). The resources are allocated and de-allocated optimally as per the application demand while considering resource availability and performance requirements based on Quality of Service parameters like energy utilization, cost, time, resource utilization and throughput.

Cloud computing is based on both dynamic and static information regarding the resources. Static information includes available datacenter information about its storage, processor cycles, memory and storage/memory allocated to VMs and throughout the cloud datacenter, this information is constant. Whereas, load allocated to hosts of datacenter is included in the dynamic information of datacenter, also, at a particular instance, the number of threads/ tasks running, running states of tasks, particular task utilizing number of CPU cycles, and tasks status during their execution. It is necessary to update in real time and at regular intervals, this dynamic information, so that the handling of dynamic requests from cloud user can be done in minimum response time. Workflows are modeled graphically as either Data flow based depicted by Directed Acyclic Graphs or Control flow based as Petri-nets. Scientific workflows are based on mainly data intensive, distributed applications without control and loop structures, so are represented by DAG while, business applications are control based represented by Petri-net.

2. RESOURCE PROVISIONING TO TASK SCHEDULING

A single resource is capable of handling multiple tasks so the order in which tasks will be executed by the resource (CPU/VM) is determined by the scheduling algorithm. A number of task scheduling algorithms are available in literature for executing dynamic tasks using distributed cloud resources and the resources are allocated in such a manner that the scheduling and load balancing algorithms must ensure minimum resource wastage while avoiding overloading/under loading of resources [1].

Following sequence of steps performed when a cloud application is submitted in cloud computing system [2]:

1**.** ***Application Partitioning:*** during this initial phase, the client application submitted to cloud system is firstly partitioned into several dependent/ independent tasks by the partitioning manager installed at the cloud server.

2. ***Task Distribution among multiple cloud datacenters***- after partitioning, the manager server distributes the tasks among different cloud datacenters at distributed locations. The decision depends on dependencies between the tasks. If the tasks belong to workflow having dependencies, then they are assigned to same cloud considering the availability of resources. In case, sufficient resources are not available at a particular cloud site, then, the tasks will be assigned to another cloud. Further, multiple tasks can be allocated to same cloud, so they queue up for the resources.

3. ***Resource mapping and task Scheduling***-

- After tasks have been allocated to suitable cloud with sufficient resources meeting the task requirements, tasks are allocated resources such that single resource is assigned to multiple tasks and ensure multi-tasking. In cloud environment a single VM can be assigned multiple tasks such that multiple VMs are hosted on a single physical machine. Further, each VM is provided by disk image by manager server before execution of tasks. Disk image is needed to provide relevant data for the task execution. Two approaches for resource allocation are:

© 2015-19, IJARCS All Rights Reserved

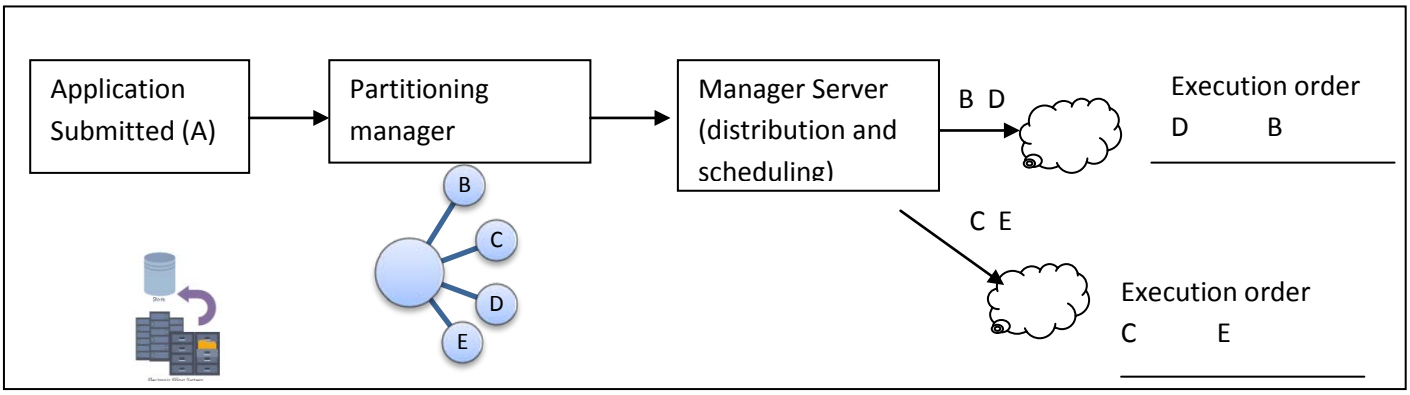

Figure 1: Steps involved from Cloud Application submission to its execution in cloud environment

a) Reservation in Advance (RA) – Advanced resources are allocated whenever they are available. This results in avoiding later delay due to non availability of resources.

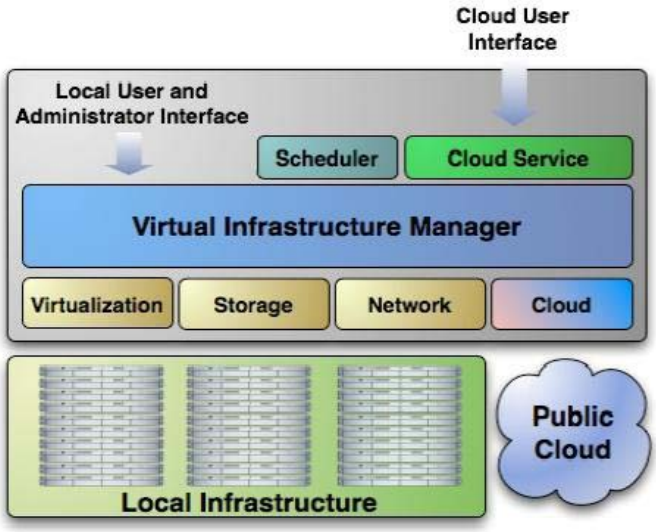

Figure 2: Cloud Architecture with Scheduler

b) Best Effort- According to this approach, the most suitable resource is searched among available resources which fulfills the requirement of the task and is allocated to it in as minimum time as possible.

If computing capacity of VM vi is C_{vi} and task t_i has computation requirement (workload) w_i, then vi is allocated to task t_i if

$C_{vi} >= \sum_{ti \in T} wi$

Multiple tasks on single resource are scheduled for execution. As shown in figure 1, an application A is submitted for execution to a cloud system which is partitioned into multiple independent/dependent tasks (B, C, D, E) by partitioning manager. These partitioned application tasks are distributed across different datacenters on same or multiple virtual servers hosted on same or different physical machines. If multiple tasks are allocated to same virtual machine, then the manager server is responsible for scheduling the tasks on same machine. However, the dependent tasks are mainly allocated to a same virtual machine to avoid communication delay among them. As shown in example, the tasks B and D are scheduled on same machine, similarly tasks C and E on different datacenter but scheduled on same virtual machine.

3.1 Role of Virtualization in Task Scheduling

Virtualization is the core technology used in cloud computing which creates multiple virtual machines (VMs) onto a single physical host (server). Hypervisors or virtual machine monitors are responsible for assigning the various VMs onto a single physical machine, hence improves the physical host utilization[5].

In other words, Virtualization Technology is used to exploit a single computational resource (host/server) to share it among different virtual machines (VMs).

Virtualization plays an important role of dividing the large computing task into smaller tasks which are allocated to

virtual machines (VMs) to across multiple physical servers. Scheduling techniques are used to allocate VMs to tasks for execution and subsequently collecting and providing back the results to cloud users. Task scheduling refers to efficient resource allocation to user requests (tasks) while following Service Level Agreements (SLAs) to achieve QoS which vary from user to user [6] and gaining maximum profit to cloud service provider. The task scheduling on VMs is NP-hard problem [7] and require optimal performance scheduling algorithms to improve quality of service in cloud environment.

Figure 2 shows a standard cloud architecture with "scheduler" at the top layer. The virtual Infrastructure Manager acts as interface between scheduler and virtualized hardware.

VMs are mapped onto physical servers and there is a centralized scheduling server which is responsible for mapping tasks onto VMs hosted on physical nodes. The physical computing nodes can be ordinary Personal Computers, servers or high performance clusters on which VMs are set-up [5].

Scheduler follows one of the scheduling strategies like Eucalyptus is based on Greedy (First fit) and rotating scheduling strategies. Greedy approach finds the first node for incoming task which best suits its requirement and deals with multiple requests (tasks) on first come first serve basis. Rotating technique stores the last position of the node which the scheduler visited and next time the scheduler begins from the last position visited rather than starting from the beginning.

3.2 Load expression for VMs on single physical machine

The virtual machines are mapped onto physical machine; the overall load of physical machine can be obtained by adding the loads of the VMs running on it (figure 3).

© 2015-19, IJARCS All Rights Reserved

Suppose the VM set comprising of 'n' virtual machines be hosted on physical machine P_i being numbered as VM_i= {vm_1, vm_2, vm_3…., vm_n}. Then, the average load on VM_j hosted on

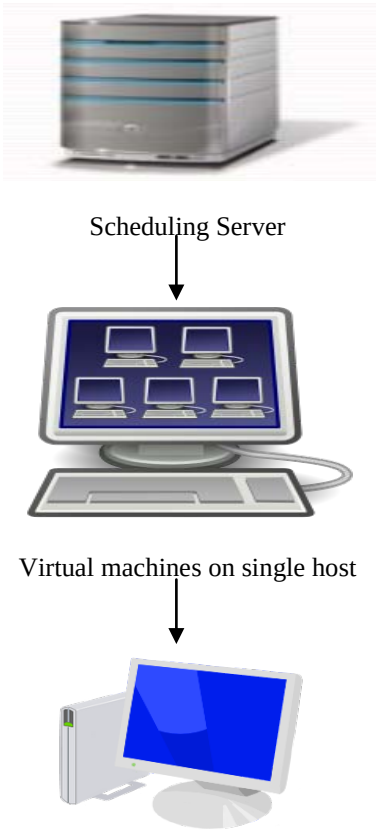

Figure 3: Mapping VMs on physical host

physical machine P_i for time period T (divided into k equal intervals of size 'Δt') is given as:

$$\textit{Average Load}\ (VM_j, T) = \frac{1}{T}\sum_{t=1}^{k} \text{load}(VMj, t) * \Delta t$$

Thus, the total load on physical machine P_i for time period T is the sum of loads on VMs hosted on it i.e.

$$\textit{Total Load}\ (P_i, T) = \sum_{j=1}^{n} \text{Avg. Load}\ (VMj, T)$$

Different scheduling approaches for cloud computing environment have been proposed by researchers. To produce optimal results various heuristic and meta-heuristic techniques have been used. The resources are allocated and utilized by VMs hosting on physical machines. In cloud computing, the demands of user are highly dynamic in nature and multi-tenancy requires isolating different users from each other and from the cloud infrastructure. The providers have to follow the SLAs and ensure the Quality of Service (QoS) requirements as mentioned by the customers. The most common QoS parameters mentioned in SLAs are high performance and throughput, better load balance, security, reliability and reduced cost, time and energy consumption [8].

Many researchers have proposed and found sub-optimal solutions towards this problem. The key parameters that are tried to optimize are minimizing makespan (total time to complete the last task on a particular resource), response time, execution time, energy consumption and operating cost while maximizing resource utilization (especially CPU, and network bandwidth) .

3. CHALELNGES TO RESOURCE MANAGEMENT AND TASK SCHEDULING

For optimal resource utilization, the proper resource management is necessary, but, as compared to traditional computing systems, Cloud computing is associated with various challenges such as:

1. **Dynamic and Fluctuating workloads-** the major challenge to elasticity in cloud computing is that the workloads are unpredictable. The workload fluctuation may occur in planned or unplanned way. In case of planned fluctuations in the workload, the situation can be predicted in advance so that resource could be allocated smoothly and in time.

2. Ensuring efficient resource utilization- the resources must be allocated instantly whenever needed, although its unplanned demand. This is called Auto-scaling in cloud computing. The incoming workload must be allocated to resources (Virtual Machines) such that the cloud service provider has to ensure efficient resource utilization. This requires optimal scheduling techniques to allocate the tasks to available machines.

3. Heterogeneous physical nodes in cloud datacenters-the tasks are allocated (or scheduled) across available nodes which are widely distributed at different location and vary in computational power, architecture, memory and even the network performance. Different tasks perform differently at different nodes [3].

4. **Increased scheduling granularity than traditional scheduling-** the size of scheduling problem has increased from simple task(process) scheduling in traditional computing systems with small data transfers to intensive VM resource scheduling and VM migrations in cloud computing environment [4].

4. CHALLENGES TO WORKFLOW SCHEDULING IN CLOUD ENVIROMENT

Workflow scheduling problem is a very dynamic and random in nature. They lack prior knowledge about randomness due to unpredictable workloads, and hence, execution time and cost factors [9] which makes this problem to be in the class of NP-hard problem, being intractable in polynomial time. Various heuristic and metaheuristic methods are used to get an optimal schedule with polynomial time complexity.

The tasks of the workflow must be executed in a specific order so that dependency constraint is handled. Workflows can be business Workflow relating to business application or Scientific Workflow. Some business workflow involves business decision taken by human decision to perform processing being passed from one participant to other.

A good scheduling algorithm must consider [11]

1. heterogeneous environment of the computing infrastructure
2. network connectivity and load among the computational sites (VMs)
3. data source(s) and sink(s) nodes.

Sophisticated algorithms are needed to schedule tasks/jobs while considering the above factors. In Cloud computing environment, one workflow task may take input data from another task executing on same VM or on different (remote) VM. Similarly, outputs can be sent to other tasks employed on same or remote VM.

© 2015-19, IJARCS All Rights Reserved

There are number of challenges faced by workflow scheduling

1. Uncertainty

The present techniques for scheduling in cloud environment follows deterministic modeling having prior knowledge about tasks and resources. However, this is not possible for cloud computing where the tasks received for computation are highly unpredictable in nature as service as provider is unaware of the amount of data and computation is required to be managed. Also, virtualization technology abstracts the cloud service provider and service users from the details of the resources available, thus adding more challenges to service provider's functionality and service users. The uncertainty about knowledge of parameters like number of computing resources available with their speed and capability, the bandwidth variations, availability of resources requires service providers as well as service users to be more concerned for ensuring minimum Quality of Service (QoS). So, researchers are working towards this challenge of reducing this uncertainty by predicting the task execution time and waiting time in queues to improve resource utilization and efficiency .

2. Quality of Service

The workflow scheduling problem in cloud environment is highly unpredictable in nature, service provider has to ensure that cloud services must be delivered with maintaining minimum QoS. On the other hand service user keeps a check on various services being received such that they follow the QoS parameters. Worse scheduling decisions is the main cause of poor QoS. As results, it will lead to long waiting and execution time of tasks, reduced throughput, and inefficient resource utilization.

However, the traditional task scheduling cannot adapt to the task scheduling for cloud environment as they do not consider the dynamic nature of computation tasks, changes in availability of computation and network resources and variations in communication delays [12].

3. Complex Integrated Architecture

Workflow management system (WMS) mainly is based on workflow DAG used to perform such comprehensive scientific applications based on large scale experiments and is used to manage, define and execute extensive distributed data represented as workflow applications.

4. Extensive data management

Scientific workflow relates to computation tasks of scientific application which involves analytical steps of large scale data analysis and mining, accessing and querying database, mathematical processing and other computational intensive applications. In other words, a workflow involves the execution of many dependent tasks which can be run in parallel [10]. Tasks are allocated to VMs in cluster of scheduling units, or jobs. A job consists of one or multiple tasks that correspond to the same workflow but different input data.

5. Load balancing

The optimal utilization of cloud resources demands uniform load must by distributed among different virtual machines so that they should not suffer from underutilization and overutilization situations.

5. CONCLUSION AND FUTURE SCOPE

Task scheduling and workflow scheduling both face the challenge of uncertainty or unpredictable workloads while the extent of data involved in them vary tremendously.

As future work, the budding researcher can work on the achieving optimal results scheduling and load balancing for various parameters like makespan, cost and energy utilization

6. REFERENCES

[1] Adhikari J. & Patil S., "Double threshold energy aware load balancing in cloud computing", Paper presented at IEEE 4thInternational Conference on Computing, Communications and Networking Technologies (ICCCNT), pp.1 – 6, 2013.

[2] Li, J., Qiu, M., Ming, Z., Quan, G., Qin, X., & Gu, Z., , "Online optimization for scheduling preemptable tasks on IaaS cloud systems", Journal of Parallel and Distributed Computing, vol. 72, pp. 666-677, 2012.

[3] Kalra, M., & Singh, S., "A review of metaheuristic scheduling techniques in cloud computing", Egyptian Informatics Journal, vol. 16, pp. 275-295, 2015.

[4] Gu, J., Hu, J., Zhao, T., & Sun, G., "A New Resource Scheduling Strategy Based on Genetic Algorithm in Cloud Computing Environment", Journal of Computers, vol. 7, pp. 42-52, 2012.

[5] Amanpreet Kaur, Bikrampal Kaur, Dheerendra Singh,"Optimization Techniques for Resource Provisioning and Load Balancing in Cloud Environment: A Review", International Journal of Information Engineering and Electronic Business(IJIEEB), vol.9, pp.28-35, 2017.

[6] Abdullah, M., & Othman, M., "Cost-based Multi-QoS Job Scheduling Using Divisible Load Theory in Cloud Computing", Procedia Computer Science, vol. 18, pp. 928-935. 2013.

[7] Abrishami, S., Naghibzadeh, M., & Epema, D. H., "Deadline-constrained workflow scheduling algorithms for Infrastructure as a Service Clouds", Future Generation Computer Systems, vol 29, pp. 158-169, 2013.

[8] Liu, J., Luo, X. G., Zhang, X. M., Zhang, F., & Li, B. N., " Job scheduling model for cloud computing based on multi-objective genetic algorithm", International Journal of Computer Science Issues, vol. 10, pp. 134-139, 2013.

[9] Zhang, F., Cao, J., Li, K., Khan, S. U., & Hwang, K., "Multi-objective scheduling of many tasks in cloud platforms", Future Generation Computer Systems, vol. 37, pp. 309-320, 2014.

[10] Prajapati, H. B., & Shah, V. A., "Scheduling in Grid Computing Environment", 2014 IEEE Fourth International Conference on Advanced Computing & Communication Technologies, pp. 315-324, 2014.

[11] Maheshwari, K., Jung, E., Meng, J., Morozov, V., Vishwanath, V., & Kettimuthu, R., "Workflow performance improvement using model-based scheduling over multiple clusters and clouds", Future Generation Computer Systems, vol. 54, pp. 206-218, 2016.

[12] Miranda, V., Tchernykh, A., & Kliazovich, D., "Dynamic Communication-Aware Scheduling with Uncertainty of Workflow Applications in Clouds", Communications in Computer and Information Science High Performance Computer Applications, pp. 169-187, 2016.

© 2015-19, IJARCS All Rights Reserved