A Taxonomy of Scheduling Algorithms for Cloud Computing

Avneesh Vashistha[1], Rabins Porwal[2], A.K Soni[3]

[1]Department of CSE, Sharda University,
Gr. Noida, U.P, India

[2]Department of IT, Institute of Technology & Science (ITS)
Mohan Nagar, Ghaziabad, U.P, India

[3]Department of CSE, Sharda University,
Gr. Noida, U.P, India

Abstract

Cloud Computing or simply rental computing, is a new technology currently being studied in the academic world and broadly categorized as Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS).Virtualization is the backbone of cloud computing and scalable, dynamic resources can be effectively managed using virtualization technology. It is possible to remap virtual machines (VMs) and physical resources according to the change in load with the help of heuristics. In this paper we present some of the most commonly used workflow heuristics currently being used in a cloud environment.

Keywords
Virtualization, Cloud, QoS, IaaS, PaaS, SaaS

1. Introduction

The basic idea of cloud computing had first been mentioned back in 1960s by John Macarathy, when he opined that computing may someday be organized as a public utility [1].The internet is often represented as a cloud and the term "Cloud Computing" arises from that analogy. According to NIST[2], 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. Cloud Computing is broadly categorized as Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS). In IaaS, cpus, grids or clusters, virtualized servers, memory, networks, storage and system software are delivered as a service. Cloud Computing or simply rental computing, is a new technology currently being studied in the academic world [3]. The definition of the cloud computing from the Gartner "A style of computing where massively scalable IT-related capabilities are provided as a service across the internet to multiple external customers using internet technologies [4].

In the cloud computing environment, resources are shared and if they are not properly distributed then it will result into resource wastage. Another essential role of cloud computing platform is to dynamically balance the load amongst different servers in order to avoid hotspots and improve resource utilization. Thereafter, the main problems to be solved are how to meet the needs of the subscribers and how to dynamically as well as efficiently manage the resources. Infrastructure as a service (IaaS), a layer in cloud computing, the cluster of virtual machines, deployed on the cloud providers' data center.

2. Virtualization Technology

Virtualization is a software that separates physical infrastructure to create various dedicated resources. In a cloud environment, dynamic resources can be effectively managed using virtualization technology. One of the key advantage of using this technology, is the possibility to seamlessly "pack" multiple under-utilized systems into a single physical host, thus achieving a better overall utilization not only of the available hardware resources, but also on entire OS along with the application running within, can be run in a virtual machine(VM). The subscribers with more demanding service level agreement (SLA) are guaranteed by accommodating as the required services within a virtual machine image and then mapping it on a physical server. This helps to solve

problem of heterogeneity of resources and platform irrelevance. Load balancing of the entire system can be handled dynamically by using virtualization technology where it becomes possible to remap virtual machines (VMs) and physical resources according to the change in load [5].

Due to these advantages, virtualization technology is being comprehensively implemented in cloud computing. However, in order to achieve the best performance, the virtual machines have to fully utilize its services and resources by adapting to the cloud computing environment dynamically. The load balancing and proper allocation of resources must be guaranteed in order to improve resource utility [6].

3. Heuristics

A scheduling is a process that maps and manages the execution of interdependent tasks on the distributed resources. It allocates suitable resources to workflow tasks so that the execution can be completed to satisfy objective functions imposed by users. Proper and effective scheduling can have significant impact on the performance of the system. In general, the problem of mapping tasks on distributed services belongs to a class of problems known as non deterministic polynomial hard time (NP-hard) problems. For such type of problems, no known algorithms are able to generate the optimal solution within polynomial time. Solution based on exhaustive search is impractical as the overhead of generating schedules is very high. Workflow scheduling discovers resources and allocates tasks on suitable resources to meet users' requirements, while data movement manages data transfer between selected resources and fault management provides mechanisms for failure handling during execution. Workflow scheduling is one of the key issues in the workflow management.

The input of workflow scheduling algorithms is normally an abstract workflow model which defines workflow tasks without specifying the physical location of resources on which the tasks are executed. Abstract workflow model can be categorized as-deterministic and non-deterministic. Dependencies of tasks and input/output data are very well known in advance in case of deterministic model, whereas in a non-deterministic model they are only known at run time. Under same computation steps, deterministic algorithms always produce same output, given a particular input. On the other hand, non-deterministic algorithms may produce different outcomes on different runs, even on the same input, as opposed to a deterministic algorithm.

To date there are two major types of workflow scheduling- best effort based and Quality of Service (QoS) constraint based scheduling. The best effort based scheduling attempts to minimize the execution time ignoring other factors such as the monetary cost of accessing resources and various users' QoS satisfaction levels. On the other hand, QoS constraint based scheduling attempts to minimize performance under most important QoS constraints, for example time minimization under budget constraints or cost minimization under deadline constraint.

Generally, best-effort based scheduling algorithms are derived from either heuristic based or meta heuristic based approach. The heuristic based approach is to develop a scheduling algorithm which fit only a particular type of problem, while the meta-heuristic based approach is to develop an algorithm based on meta-heuristic method which provides a general solution method for developing a specific heuristic to fit a particular kind of problem [12].

Some of the commonly used [13] heuristics are:

OLB - Opportunistic Load Balancing (OLB) assigns each task randomly, to the next machine that is expected to be available, regardless of the task's expected execution time on that machine [10, 11]. The motivation behind OLB is to keep all machines as busy as possible. One advantage of OLB is its simplicity, but because OLB does not consider expected task execution times, the mappings it finds can result in very poor makespans.

MET - Minimum Expected Time (MET) assigns each task, in arbitrary order, to the machine with the best execution time for that task, regardless of that machine's availability [10]. The logic behind MET is to give each task to its best machine. This can cause a severe load imbalance across machines. However, MET is obviously not applicable to Heterogeneous Computing (HC) environments characterized by consistent ETC matrices.

MCT - Minimum Completion Time (MCT) assigns each task, in arbitrary order, to the machine with the minimum expected completion time for that task [10]. This causes some tasks to be assigned to machines that do not have the minimum execution time for them. The logic behind MCT is to combine the benefits of OLB and MET, while avoiding the circumstances in which OLB & MET perform poorly. Min-Min, Max-Min, Sufferage proposed by Maheswaren et al. [14] are three major heuristics which have been employed for scheduling workflow

tasks in Pegasus [16] and vGrADS [16]. The heuristics is based on the performance estimation for task execution and I/O data transmission. The Min-Min heuristic schedules tasks having shortest execution time first so that it results in the higher percentage of tasks assigned to their best choice than Max-Min heuristics [13].

Genetic Algorithm (GA) - A genetic algorithm combines exploitation of best solutions from past searches with the exploration of new regions of the solution space. Any solution in the search space of the problem is represented by an individual (chromosome). A genetic algorithm maintains a population of individuals that evolves by a fitness function. A fitness value indicates how good the individual is compared to others in the population. A typical genetic algorithm [18] is illustrated as follows:

1. **[Start]** Generate random population of *n* chromosomes (suitable solutions for the problem)
2. **[Fitness]** Evaluate the fitness *f(x)* of each chromosome *x* in the population
3. **[New population]** Create a new population by repeating following steps until the new population is complete
   (*i*) **[Selection]** Select two parent chromosomes from a population according to their fitness (the better fitness, the bigger chance to be selected)
   (*ii*) **[Crossover]** with a crossover probability cross over the parents to form a new offspring (children). If no crossover was performed, offspring is an exact copy of parents.
   (*iii*) **[Mutation]** with a mutation probability mutate new offspring at each locus (position in chromosome).
   (*iv*) **[Accepting]** Place new offspring in a new population
4. **[Replace]** Use new generated population for a further run of algorithm
5. **[Test]** If the end condition is satisfied, **stop**, and return the best solution in current population
6. **[Loop]** Go to step **2**

Shu-Ching Wang et al. [7] proposed a scheduling algorithm that combines OLB and LBMM that can utilize better executing efficiency and maintain the load balancing of the system.

William Voorsluys et al. present a performance evaluation on the effect of live migration of virtual machines on the performance of application running inside Xen VMs [8]. Ibrahiem M.M EL Eamry et al. [9] proposed a methodology for solving MIN-MAX problems using genetic algorithms. As a high efficient search strategy for global optimization, genetic algorithm demonstrates favourable performance for solving optimization problems. Optimization problems can be solved with genetic algorithm through efficient encoding, selection of fitness function and various genetic operations.

4. Summary

On the basis of comparison between heuristics based scheduling approaches and meta-heuristics based approaches, the meta-heuristics based scheduling approaches are advantageous in terms of producing optimized scheduling solution based on the performance of entire workflow rather than the partial of the workflow as mostly considered by heuristics based approaches. Heuristics based scheduling normally designed for a specific type of workflow applications, while meta-heuristics based scheduling approaches requires significantly higher scheduling time for producing a good quality and acceptable solution. Therefore, the heuristics based scheduling algorithms are well suited for a workflow with a simple structure, while the Meta heuristics based approaches have a lot of potential for solving large and complex structure workflows [12].

According to [17] among these five algorithm (OLB, MET, MCT, MIN-MIN, MAX-MIN) Min-Min is the most outstanding one. Although MCT usually outperforms the OLB and MET, it can induce a larger makespan compared with Min-Min. If there are several long tasks in many short tasks, Max-Min will outperform Min-Min with a balance system workload. Experimental results conducted by Maheswaran et al.[14] and Casanova et al. [15] have proved that Min-Min heuristic outperform Max-Min heuristic. However, since Max-Min schedule have more chance of being executed in parallel with shorter tasks. Therefore, it might be expected that the Max-Min heuristic perform better than the Min-Min heuristic in the cases where there are many more short tasks than long tasks [13, 14]. On the other hand, since the Sufferage heuristic consider the adverse effect in the completion time of a task if it is not scheduled on the resource having with minimum completion time[14], it is expected to perform better in the cases where large performance difference between resources.

5. Conclusion

Based on above analysis we have categorized workflow scheduling algorithms as either best-effort

based or QoS constraint based scheduling. Some heuristics and metaheuristics based algorithms which can optimize workflow execution times have been presented. Best effort based scheduling algorithms pays attention in which resource providers facilitates free access while QoS constraint based scheduling algorithms targets where SLA are established between service users and service providers.

References

[1] S. L. Garfinkel, Architects of the Information Society: Thirty- Five Years of the Laboratory for Computer Science at MIT, H. Abelson, Ed. The MIT Press, 1999

[2] http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf

[3] Michael Armbrust, Armando Fox, Rean Griffith, Anthony D. Joseph,Randy Katz, Andy Konwinski, Gunho Lee, David Ptterson, Ariel Rabkin, Ion Stoica, and Matei Zaharia, "Above the clouds: A berkeley view of cloud computing," UCB/EECS-2009-28.

[4] Cloud Computing Definition Gartner. http://www.gartner.com/newsroom/id/1035013

[5] Borja Sotomayor, Kate Keahey, Ian Foster, and Tim Freeman, "Enabling cost-effective resource leases with virtual machines," ACM/ IEEE International Symposium on High Performance Distributed Computing (HPDC 2007), 2007

[6] L. Cherkasova, D. Gupta, and A.Vahdat, "When virtual is harder than real: Resource allocation challenges in virtual machine based it environments," Technical Report HPL 2007-25, February 2007.

[7] W. Shu-Ching, Y. Kuo-Qin, L. Wen-Pin, and W. Shun-Sheng., "Towards a laod balancing in a three level cloud computing network", IEEE, 2010, Pp-108-113

[8] V. William, B. James, V. Srikumar, and B. Rajkumar, "Cost of Virtual Machine Live Migration in Clouds: A Performance Evaluation", CloudCom '09 Proceedings of the 1st International Conference on Cloud Computing, pp-254-265, 2009.

[9] M. M Ibrahiem, E. El, A. Al Dahoud, M. Mona, and E. Abd, "Solving Min-Max Problems using Genetic Algorithm", in University of Pitesti-Electronics and Computer Science, Scientific Bulletin, No. 9, 2:33-39, 2009.

[10] A. Robert, H. Debra, and K. Taylor, "The relative performance of various mapping algorithms is independent of sizable variances in run-time predictions", in 7th IEEE Heterogeneous computing Workshop (HCW '98), pp79-87, 1998.

[11] F. F. Richard, G. Michael, A. Stephen, C. Mark, H. Mike, H. Debra, K. Elaine, K. Taylor, K. Matt, L. John D, M. Francesca, M. Lantz, R. Brad, and S. H. J , "Scheduling Resources in Multi-User, Heterogeneous, Computing Environments with SmartNet," IEEE Proceedings of Heterogeneous Computing Workshop (HCW 98), pp-184-199, 1998.

[12] Yu Jia, B. Rajkumar, and R. Kotagiri, "Workflow Scheduling Algorithms for Grid Computing", Metaheuristics for Scheduling in Distributed Computing Environments Studies in Computational Intelligence, Volume-146, pp-173-214, 2008.

[13] B. Tracy D., S. Howard Jay, and B. Noah "A comparison of eleven static heurisitics for mapping a class of independent tasks onto heterogeneous distributed computing systems", Journal of parallel and distributed computing, 61:801-837, 2001.

[14] M. Muthucumaru, A. Shoukat, S. Howard Jay, H. Debra, and F. Richard F, "Dynamic matching and scheduling of a class of independent tasks onto heterogeneous computing systems," in the 8th heterogeneous computing workshop (HCW'99), San Jaun, Puerto Rico, April 12, 1999

[15] C. Henri, L. Arnaud, Z. Dmitrii, and B. Francine, "Heuristics for scheduling parameter sweep applications in grid environments," in the 9th heterogeneous computing workshop (HCW'00), April. 2000.

[16] B. Jim, J. Sonal, D. Ewa, M. Anirban, and K. Ken, "Task Scheduling Strategies for Workflow-based Application in Grids," IEEE International Symposium on Cluster Computing and the Grid (CCGrigd 2005), 2005

[17] M. Tinghuai, y. Qiaoqiao L. Wenjie, G. Donghai, and L. Sungyoung, "Grid Task

Scheduling Algorithm Review", IETE, 28:2, pp158-167, 2011.

[18]http://www.obitko.com/tutorials/genetic-algorithms/ga-basic-description.php

Avneesh Vashistha is a reseach scholar in Sharda Univeristy, Greater Noida. His research area is Cloud Computing.

Dr. Rabins Porwal is associated with ITS Mohan Nagar, Ghaziabad, a premier institution. His area of specialization of software engineering.

Dr. A K Soni is a renowned personality in academic field. Having experience of more than three decades in academia and industry.