ijiarec.com

ISSN: 2348-2079

FEB-2015

International Journal of Intellectual Advancements and Research in Engineering Computations

LOAD BALANCING AND SCHEDULING OF JOBS IN CLOUD DATA CENTERS

[1]N. Deebaa, [2]A. Jayanthi M.E, [3]Dr.S.Karthik M.E, Ph.D

ABSTRACT

Reducing the energy usage in data centers is mainly to be focused in cloud computing. Computational resource should be fairly allocated among different organization. Energy cost reduction and allocation of computational resource should be considered mainly. The time interval should be considered while performing the jobs in the data center. Server over heating should be reduced in the data centers to avoid server failure. The batch jobs should be scheduled in the data centers in the balanced order in the resource manager. An online scheduling algorithm Equally spread current execution load balancing algorithm is used to allocate the jobs to the data centers from the resource manager. Equally spread current execution load balancing can reduce the energy cost and servicing time can be reduced. It can reduce the cost and servicing time in the geographically distributed data centers.

Keyword – Cloud Computing, Load balancing, scheduling algorithm, resource manager.

1. INTRODUCTION

Cloud computing uses the computing devices which is represented in the Figure 1.1. Relatively new term for representing collection of resources which are shared, scaled dynamically.

Cloud computing are pay per use method and the user can pay according to the use of resources. This refers to both, applications service to users and servers in data centers which support those services.

Cloud computing can be defined as collection of resources (servers in data center), which are interconnected with each other and using virtualization technology can be scaled and adapted dynamically. In Cloud computing, customers can rent the cloud data centers for their use as long as they need it. Customers no need to buy a physical hardware to start their business. Customers can earn their profit by renting the cloud storage. Customers can use various applications as they need for their usage.

Author for Correspondence:
[1]PG Scholar, CSE, SNS College of Technology, Coimbatore, Tamilnadu, India. Email: deebaa.cse@gmail.com, bommals29@gmail.com.
[2]Assistant Professor, CSE, SNS College of Technology, Coimbatore, Tamilnadu, India.
[3]Professor & Dean, CSE, SNS College of Technology, Coimbatore, Tamilnadu, India.

VIRTUALIZATION TECHNOLOGY

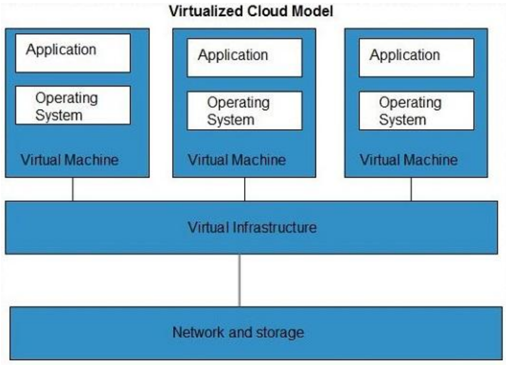


Figure 1.2 Virtualization Technologies

Virtualization , in computing is referred to the act of creating a virtual version of something, including but not limited to a virtual computer hardware platform, operating system, storage device or computer resource network.

HARDWARE VIRTUALIZATION

Hardware virtualization is also referred to as a platform virtualization. Here platform virtualization will acts as a real operating system in the computer. Executing software can be separated from the underlying hardware in the virtual machines. For example, we can use multiple operating systems at the same time in a single machine. While using a windows operating system, we can use other operating systems like Ubuntu, Linux, etc,. At the same time.

In Platform virtualization, the actual machine where the virtualization takes place is the host machine and the virtual machine is the guest machine.

Hardware virtualization can be divided in to three types which includes a) Full virtualization b) Partial Virtualizations and c) Para Virtualization.

a) Full Virtualization

Full virtualization is a technique where the whole process will done in a virtual machine environment, that is the whole simulation will done in the underlying hardware. Full virtualization needs every outstanding feature of the hardware be reflected in one of more than a few virtual machines – including the full instruction set, input/output operations, interrupts, memory access, and every other elements used by the software that runs on the bare machine,

and that is planned to run in a virtual machine. In such surroundings, any software skilled to execute on the raw hardware can be run in the virtual machine and, in any individual operating system. The clear analysis of full virtualization is whether an operating system proposed for separate use can effectively run within a virtual machine. Here in a particular machine that could be multiplexed among lots of users.

b) Partial virtualization

In partial virtualization, as well as address space virtualization, the virtual machine simulates several instances of much of a necessary hardware environment. Generally, this means that complete operating systems cannot run in the virtual machine, as it is a full virtualization, but as much as applications can run. An input form of partial virtualization is address space virtualization, in which all virtual machine will have a separate address space. Partial virtualization was most important one to create a full virtualization.

Partial virtualization is very easy to execute than full virtualization. It has frequently provided helpful, robust virtual machines, skilled of sustaining significant applications. Partial virtualization has established very much successful for sharing computer resources between multiple users.

c) Para Virtualization

In Para virtualization, the virtual machine does not essentially simulate hardware, but as an alternate, offers an extraordinary API that can only be used by modifying the "guest" OS. For this to be potential, the "guest" OS's source code must be offered. If the source code is existing, it is sufficient to go back susceptible commands with calls to VMM APIs. Then compile the OS again and utilize the new binaries.

RELATED WORK

Task scheduling, In this environment task scheduling is considered to be an important issue. A good task scheduling algorithm is to be used in the cloud computing to utilize the resource in it efficiently. Cloud task can be separated into two categories like on-line mode service and the batch mode service. [1] Online cloud task scheduling depends on virtual machine adaptive fault tolerance and load balancing can be proposed using an ant colony algorithm. The major involvement of this job is that load balancing factor is added and the decisions can be taken according to the fault tolerance of the system on the reliability of the virtual machine. The proposed scheduling strategy was simulated using the Cloudsim toolkit package. Experimental outcome prove that the proposed algorithm reached the improved load balance than Join shortest- queue (JSQ) and Modified Ant Colony Optimization (MACO) algorithms.

Cost reduction The energy cost reduction in IDCs and guarantees a service delay bound is described [3]. Internet Data Centers worn in cloud computing becomes a division of people‘s daily life. Internet data center (IDC) infrastructure support these services. As insist for cloud computing services, energy consumed by IDCs is very high. Both universities and industry have rewarded great notice to energy management of IDCs. To handle the energy management like to reduce the power cost for IDCs in deregulated electricity markets are studied. A novel two-stage design and the eco-IDC (Energy Cost Optimization-IDC) algorithm to develop the sequential variety of electricity price and dynamically schedule workload to perform on IDC servers through an input queue is proposed. The estimate outcome reveal that the projected approach drastically reduces energy cost for IDCs, guarantees a service delay bound, and increases workload fall if the service delay bound is adequately high.

High energy cost reduction To diminish the high energy cost in large data centres is becoming very important now-a-days [8]. Meanwhile, part of the computational resources required to be fairly allocated between different organizations. Latency is an additional major worry for resource management. However, energy cost, resource allocation fairness, and latency are essential but often contradicting metrics on development of data center workloads. Moreover, with the growing power density, data center operation must be reduced to avoid server overheating. An online scheduling algorithm - GreFar is proposed, which reduces the energy cost and fairness between different organizations focus to queuing delay constraints, while fulfilling the most server inlet temperature constraints. Prior information of workload arrival or electricity variations are not required in the GreFar algorithm. The cost can be reduced randomly close to that of the optimal offline algorithm is proved. Performance of GreFar algorithm is better when compared to the previous algorithm T-unaware. 16 percent of energy-fairness can be saved in the GreFar algorithm when compared to T-unaware.

MODULE DSCRIPTION

DATA CENTER SYSTEM MODEL

There are N geographically distributed data centers, each of which houses thousands of servers. Servers may be homogeneous or heterogeneous in hardware and performance characteristics. One major source of heterogeneity is that data centers operate several generations of servers from multiple vendors. Application needs, hardware innovations and prices jointly determine which type of servers to purchase. Our model accommodates both homogeneous and heterogeneous servers. Next, the state of each data center, which is time-varying and captures the randomness in the atmosphere, is noted.

Server Availability

Server availability may change over time due to different reasons such as server failures, software upgrades, influence of other workloads, etc. For example, the increase of interactive workloads may reduce the number of servers available to process batch jobs. In data center it at time t, without loss of generality, that there are Ki(t) (possibly heterogeneous) available servers are considered. In data center i, each server k, for k = 1, 2, . . .;$K_i(t)$ is characterized by three parameters: processing speed, idle power and active power.

Electricity Price

Due to the deregulation of electricity markets, electricity prices stochastically vary over time (e.g., every hour or 15 minutes) and across different locations. However, our model and analysis are still applicable when the total electricity cost is not a linear function of the energy consumption. For example, the electricity cost can be an increasing and convex (or other) function of the energy consumption. In such scenarios, the amount of other workloads such as interactive workloads also affects the energy price, and hence, need to add another component, i.e., energy consumed by other workloads during each time t, into the data center state.

JOB MODEL

Market is a central module of the Cloud Computing ecosystem. It is necessary for modifiable Cloud resource trading and on-line discussions: in public Cloud Computing model, where services are accessible in a pay-per-use model, costs for accessing the Cloud infrastructure is a main parameter to be measured in research studies. Furthermore, Cloud consumers need mechanisms to discover various Cloud providers along with their pricing policies. Thus, modeling of costs and pricing policies is an important aspect to be considered when designing a Cloud simulator. To permit the designing of the Cloud market, market-related properties connected to a data center were categorized in two layers.

Jobs of the same type can be processed by any number of servers simultaneously. In practice, however, it may be possible that only a certain number of servers can process jobs of the same type in parallel. Our model can be adapted easily to capture this fact by adding a parallelism constraint for each job type. Specifically, there need to add a constraint on the scheduling decisions such that the maximum number of servers that can be used to process jobs of the same type simultaneously is upper bounded.

SCHEDULING MODEL

Based on the recently developed Lyapunov optimization technique, an online algorithm ―GreFar‖, whose performance is provably ―good‖ compared to that of the optimal offline policy with T-step look ahead information. The intuition of GreFar is

to trade the delay for energy-fairness cost saving by using the queue length as a guidance for making scheduling decisions: jobs are processed only when the queue length becomes sufficiently large and/or electricity prices are adequately low. Before

Presenting the algorithm, to introduce ―queue dynamics, which specifies the queue length changes governed by the scheduling decisions (and job arrivals) is needed. The queue dynamics is instrumental for the scheduler to make online decisions.

SIMULATION

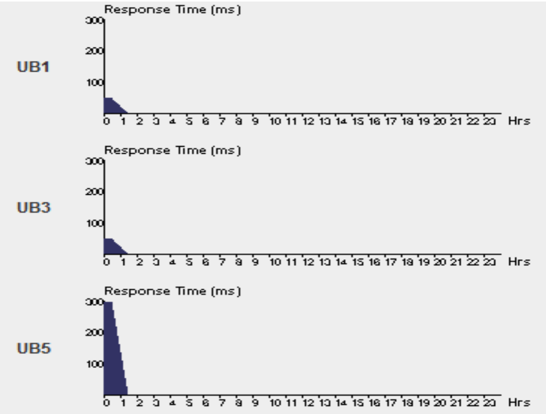
OVERALL RESPONSE TIME SUMMARY

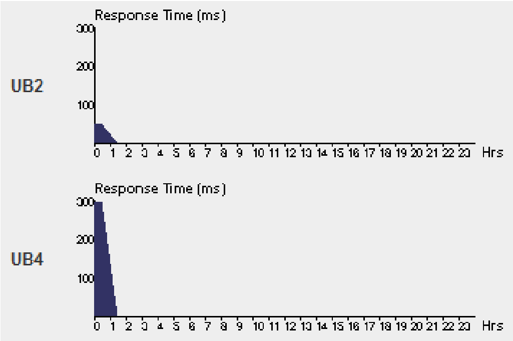
	Avg (ms)	**Min(ms)**	**Max(ms)**
Overall response time:	150.78	40.12	370.64
Data center processing time:	0.44	0.02	0.88

Response Time by Region

User base	**Avg (ms)**	**Min (ms)**	**Max (ms)**
UB1	50.00	41.11	60.61
UB2	50.28	40.12	61.87
UB3	50.01	40.38	61.63
UB4	300.49	234.14	370.64
UB5	300.17	229.62	369.14

User Base Hourly Response Times

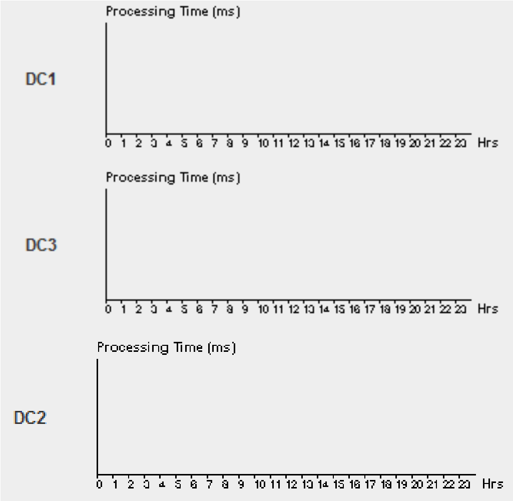




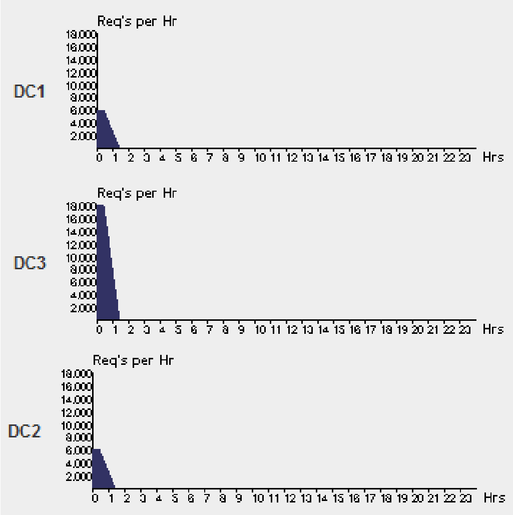
Data Center Request Servicing Times

Data Center	Avg (ms)	Min (ms)	Max(ms)
DC1	0.47	0.02	0.86
DC2	0.48	0.03	0.87
DC3	0.41	0.02	0.88

DATA CENTER HOURLY AVERAGE PROCESSING TIMES



Data Center Loading



COST

Total Virtual Machine Cost: $1.51

Total Data Transfer Cost: $0.32

Grand Total: $1.83

Data Center	VM Cost $	Data Transfer Cost $	Total $
DC2	0.502	0.065	0.567
DC1	0.502	0.064	0.566
DC3	0.502	0.192	0.693

CONCLUSION

Gre-Far algorithm is used to improve the performance of the system by reducing the energy cost, fairness of the resource allocation and queuing delay. It select the servers where the power consumption is low and schedule the arrived resources. Load balancing is not properly done in all servers present in the data center. An Adaptive Resource Allocation algorithm can be used to schedule the jobs in all servers in data center without reducing the performance of the system.

REFERENCE

[1]. Arabi E. Keshk (2014) ―Cloud Computing Online Scheduling ISSN (e): 2250- 3021, ISSN (p): 2278-8719 Vol. 04, Issue 03 , V6 PP 07-17.

[2]. Basmadjian R, Hermann De Meer, Ricardo Lent and Giovanni Giuliani (2012)―Cloud computing and its interest in saving energy: the use case of a privatecloud Journal of Cloud Computing: Advances, Systems and Applications, 1:5.

[3]. Jianying Luo, Lei Rao, and Xue Liu (2014) ―Temporal Load Balancing with Service Delay Guarantees for Data Center Energy Cost Optimization IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 25, NO. 3.

[4]. LI Hongyou, WANG Jiangyong, PENG Jian, WANG Junfeng, LIU Tang (2013) ―Energy-Aware Scheduling Scheme Using Workload-Aware Consolidation Technique in Cloud Data Centres.

[5]. Lin M, Wierman A, Andrew L L H and Thereska E (2011) - Dynamic right-sizing for power-proportional data centers ― Extended version Proc. IEEE INFOCOM.

[6]. Mosch M, Groß S and Alexander Schill (2014) ―User-controlled resource management in federated clouds Journal of Cloud Computing: Advances, Systems and applications, 3:10.

[7]. Mastroianni C, Meo M and Papuzzo G (2013) ―Probabilistic Consolidation of Virtual Machines in Self-Organizing Cloud Data Centers IEEE TRANSACTIONS ON CLOUD COMPUTING, VOL. 1, NO. 2.

[8]. Polverini M, Cianfrani A, Ren S and Athanasios V. Vasilakos (2014) "Thermal-Aware Scheduling of Batch Jobs in Geographically Distributed Data Centers" IEEE TRANSACTIONS ON CLOUD COMPUTING, VOL. 2, NO. 1.

[9]. Yi-Ju Chiang, Yen-Chieh Ouyang, and Ching-Hsien Hsu (2014) ―An Efficient Green Control Algorithm in Cloud Computing for Cost Optimization IEEE TRANSACTIONS ON CLOUD COMPUTING, TCCSI-2014-03-0116.

[10]. Yaozu Dong, Xiantao Zhang, Jinquan Dai, and Haibing Guan (2014) ―HYVI: A HYbrid VIrtualization Solution Balancing Performance and Manageability IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 25, NO. 9.