International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 1 Issue 10, December- 2012

Load Balancing The Essential Factor In Cloud Computing

Mr. Jayant Adhikari, Prof. Sulabha Patil,
Department of Computer Science and Engineering
Tulsiramji Gaikwad-Patil
College of Engineering, RTMNU, Nagpur

Abstract— Nowadays implementation of local cloud is popular, organization are becoming aware of power consumed by unutilized resources. Reducing power consumption has been an essential requirement for cloud environments not only to decrease operating cost but also improve the system reliability. The energy-aware computing is not just to make algorithms run as fast as possible, but also to minimize energy requirements for computation. This paper discusses the existing load balancing techniques in cloud computing and further compares them based on various parameters like performance, scalability, associated overhead etc. that are considered in different techniques. It further discusses these techniques from energy consumption and carbon emission perspective.

Keywords -
Cloud Computing, Virtual machine, Consolidation, Energy-Aware Scheduling, Load Balancing.

I. INTRODUCTION

Cloud computing can be classified as a new paradigm for dynamic provisioning computer services supported by data centers that usually employ virtual machine (VM) technology for consolidation[4]. Cloud computing deliver infrastructure, platform and software as services which are made available to customers as subscription based services under the pay as you go model. Using this services customers are given access to resources provided by a cloud vendor as described in their Service Level Agreement (SLA). Clouds use virtualization technology in distributed data centers to allocate resources to customers as they need them. Generally, clouds are deployed to customers giving them three levels of access: Software-as-a-Service (SaaS), Platform-as-a-Service (PaaS), and Infrastructure-as-a-Service (IaaS). The jobs can differ greatly from customer to customer.

Load balancing is one of the central issues in cloud computing [5]. It is a mechanism that distributes the dynamic local workload evenly across all the nodes in the whole cloud to avoid a situation where some nodes are heavily loaded while others are idle or doing little work. It helps to achieve a high user satisfaction and resource utilization ratio, hence improving the overall performance and resource utility of the system.

Local cloud implementations are becoming popular due to the fact that many organizations are reluctant to move their data to a commercialized cloud vendor. There are several different implementations of open source cloud software that organizations can utilize when deploying their own private cloud. Some possible solutions are OpenNebula [16] or Nimbus [18] or cloudbus[17]. This architecture is built for ease of scalability and availability, but does not address the problem of the amount of power a typical architecture like this consumes.

Organizations that wish to build local clouds do so using commodity hardware. This may mean that the cloud is made up of several different hardware sets up. Even when a cloud is initially built using one type of hardware, the nature of a cloud often means it will be expanded by adding new and different hardware throughout the course of its lifetime. The main part of power consumption in data centers come from computation processing, disk storage, networks and cooling system[15].

The rest of this paper is organized as follows. Section II need of energy-aware scheduling in clouds. Section III describes energy-aware cloud architectural elements. Section IV presents existing load balancing techniques in cloud computing. Section V describes comparison of existing load balancing technique and Section VI concludes.

II. NEED OF ENERGY-AWARE SCHEDULING IN CLOUDS

Cloud computing is a client-server architecture composed by large and power-consuming data centers designed to support the elasticity and scalability required by consumers[4]. Data center uses huge amount of data so to maintaining local cloud is costly. It consume near about 10 to 100 times more power than a office building[5]. Thus proper load balancing can help in utilizing the available resources optimally, thereby minimizing the resource consumption.

• *Reducing Energy Consumption* - Load balancing helps in avoiding overheating by balancing the workload across all the nodes of a cloud, hence reducing the amount of energy consumed[6]. To reduce power consumption of data centers, one can consolidate computation workloads onto a subset of servers, and power off servers that become idle after consolidation. The key idea is to reduce “idle power”, i.e., the power consumed by idle servers that do not have workload, so as to reduce the number of servers needed to power on.

III. ENERGY-AWARE CLOUD ARCHITECTURAL ELEMENTS

Figure 1 shows the high-level architecture for supporting energy-efficient service allocation in Cloud computing infrastructure[11]. There are basically four main entities involved:

a) *Consumers/Brokers*: Cloud consumers or their brokers submit service requests from anywhere in the world to the Cloud. It is important to notice that there can be a difference between Cloud consumers and users of deployed services. For instance, a consumer can be a company deploying a Web application, which presents varying workload according to the number of "users" accessing it.

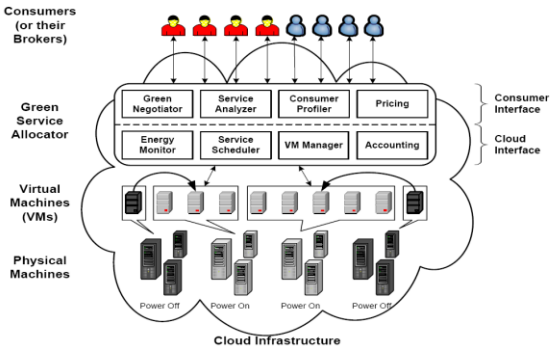

Figure 1: High-level system architectural framework

b) *Green Resource Allocator*: Acts as the interface between the Cloud infrastructure and consumers. It requires the interaction of the following components to support energy-efficient resource management:

• *Green Negotiator*: Negotiates with the consumers/brokers to finalize the SLA with specified prices and penalties (for violations of SLA) between the Cloud provider and consumer depending on the consumer's QoS requirements and energy saving schemes. In case of Web applications, for instance, QoS metric can be 95% of requests being served in less than 3 seconds[].

• *Service Analyser*: Interprets and analyses the service requirements of a submitted request before deciding whether to accept or reject it. Hence, it needs the latest load and energy information from VM Manager and Energy Monitor respectively.

• *Consumer Profiler*: Gathers specific characteristics of consumers so that important consumers can be granted special privileges and prioritised over other consumers.

• *Pricing*: Decides how service requests are charged to manage the supply and demand of computing resources and facilitate in prioritising service allocations effectively.

• *Energy Monitor*: Observes and determines which physical machines to power on/off.

• *Service Scheduler*: Assigns requests to VMs and determines resource entitlements for allocated VMs. It also decides when VMs are to be added or removed to meet demand.

•*VM Manager*: Keeps track of the availability of VMs and their resource entitlements. It is also in charge of migrating VMs across physical machines.

• *Accounting*: Maintains the actual usage of resources by requests to compute usage costs. Historical usage information can also be used to improve service allocation decisions.

c) *VMs*: Multiple VMs can be dynamically started and stopped on a single physical machine to meet accepted requests, hence providing maximum flexibility to configure various partitions of resources on the same physical machine to different specific requirements of service requests. Multiple VMs can also concurrently run applications based on different operating system environments on a single physical machine. In addition, by dynamically migrating VMs across physical machines, workloads can be consolidated and unused resources can be put on a low-power state, turned off or configured to operate at low-performance levels (e.g., using DVFS) in order to save energy.

d) *Physical Machines*: The underlying physical computing servers provide hardware infrastructure for creating virtualized resources to meet service demands.

IV. EXISTING LOAD BALANCING TECHNIQUES IN CLOUD COMPUTING

Following load balancing techniques are currently prevalent in clouds:-

A. Dynamic Round-Robin

Dynamic Round-Robin[1] method is proposed as an extension to the Round-Robin method. Dynamic Round-Robin method uses two rules to help consolidate virtual machines. The first rule is that if a virtual machine has finished and there are still other virtual machines hosted on the same physical machine, this physical machine will accept no more new virtual machine. Such physical machines are referred to as being in "retiring" state, meaning that when the rest of the virtual machines finish their execution, this physical machine can be shutdown.

The second rule of *Dynamic Round-Robin* method is that if a physical machine is in the "retiring" state for a sufficiently long period of time, instead of waiting for the residing virtual machines to finish, the physical machine will be forced to migrate the rest of the virtual machines to other physical machines, and shutdown after the migration finishes. This waiting time threshold is denoted as "retirement threshold". A physical machine that is in the retiring state but cannot finish all virtual machines after the retirement threshold will be forced to migrate its virtual machines and shutdown.

The Dynamic Round-Robin strategy uses these two rules in order to consolidate virtual machines deployed by the

original Round-Robin method. The first rule avoids adding extra virtual machines to a retiring physical machine. The second rule speeds up the consolidation process and enables Dynamic Round-Robin to shutdown physical machines, so that it can reduce the number of physical machine used to run all virtual machines, thus achieve power saving.

B. A Hybrid Approach

In order to conserve more energy, Ching-Chi Lin[1] combine Dynamic Round-Robin and First-Fit into a Hybrid algorithm. The number of incoming virtual machines is assumed to be a function of time, and follows a probability distribution (e.g., a normal distribution). Hybrid algorithm will use the incoming rate of virtual machines to guide the scheduling of virtual machines. The Hybrid method uses First-Fit during rush hours to fully utilize the computing power of physical machines, and uses the Dynamic Round-Robin to consolidate virtual machines and reduce energy consumption during non-rush hours.

C. *PALB(Power Aware Load balancing) Algorithm*

The PALB algorithm[2] has three basic sections. The balancing section is responsible for determining here virtual machines will be instantiated. It does this by first gathering the utilization percentage of each active compute node. In the case that all compute nodes n are above 75% utilization, PALB instantiates a new virtual machine on the compute node with the lowest utilization number. It is worth mentioning in the case where all compute nodes are over 75% utilization, all of the available compute nodes are in operation. Otherwise, the new virtual machine (VM) is booted on the compute node with the highest utilization (if it can accommodate the size of the VM). The threshold of 75% utilization was chosen since when 25% of the resources are available, at least one more virtual machine can be accommodated using three out of five available configurations.

The upscale section of the algorithm is used to power on additional compute nodes (as long as there are more available compute nodes). It does this if all currently active compute nodes have utilization over 75%. The downscale section is responsible for powering down idle compute nodes. If the compute node is using less than 25% of its resources, PALB sends a shutdown command to that node.

D. ESCE (Equally Spread Active Execution) algorithm

The cloud manager estimates the job size and checks for the availability of the virtual machine and also the capacity of the virtual machine. Once the job size and the available resource (virtual machine) size match, the job scheduler immediately allocates the identified resource to the job in queue. The impact of the ESCE algorithm[3] is that there is an improvement in response time and the processing time. The jobs are equally spread, the complete computing system is load balanced and no virtual machines are underutilized. Due to this advantage, there is a reduced in the virtual machine cost and the data transfer cost.

E. Task Consolidation Algorithms

Both ECTC and MaxUtil [4] follow similar steps in algorithm description with the main difference being their cost functions. In a nutshell, for a given task, two heuristics check every resource and identify the most energy efficient resource for that task. The evaluation of the most energy efficient resource is dependent on the used heuristic, or more specifically the cost function employed by the heuristic. The cost function of ECTC computes the actual energy consumption of the current task subtracting the minimum energy consumption required to run a task if there are other tasks running in parallel with that task. That is, the energy consumption of the overlapping time period among those tasks and the current task is explicitly taken into account.

F. ACCLB (Load Balancing mechanism based on ant colony and complex network theory) Algorithm

ACCLB load balancing mechanism[7] based on ant colony and complex network theory in an open cloud computing federation. It uses small-world and scale-free characteristics of a complex network to achieve better load balancing. This technique overcomes heterogeneity, is adaptive to dynamic environments, is excellent in fault tolerance and has good scalability hence helps in improving the performance of the system.

G. MCMF(Minimum Cost Maximum Flow) algorithm

The exact formulation based on the modified Bin-Packing model[8] suffers from scalability problems with large instances and increasing number of PMs as well as the length of the requests. This has been the motivation for seeking an alternate approach to the dynamic resource placement problem and this led to the Minimum Cost Maximum Flow (MCMF) algorithm. The MCMF on the contrary is not subject to the scalability issues of Bin-Packing.

H. Join-Idle-Queue

Y. Lua et al. [12] proposed a Join- Idle-Queue load balancing algorithm for dynamically scalable web services. This algorithm provides large scale load balancing with distributed dispatchers by, first load balancing idle processors across dispatchers for the availability of idle processors at each dispatcher and then, assigning jobs to processors to reduce average queue length at each processor. By removing the load balancing work from the critical path of request processing, it effectively reduces the system load, incurs no communication overhead at job arrivals and does not increase actual response time.

I.DAIRS (dynamic and integrated resource scheduling) algorithm

Wenhong Tian[9] introduce a dynamic and integrated resource scheduling algorithm (DAIRS) for Cloud datacenters. Unlike traditional load-balance scheduling algorithms which consider only one factor such as the CPU load in physical servers, DAIRS treats CPU, memory and network bandwidth integrated for both physical machines and virtual machines.

V. COMPARISON OF EXISTING LOAD BALANCING TECHNIQUE

Below table show the comparative study of different load balancing. Difference made on bass of techniques that are used in respective algorithms, advantages and disadvantages.

Table 1: Comparisons of different load balancing algorithms

Algorithm/ Technique	Description	Advantages	Disadvantages
Dynamic Round-Robin[1]	The first rule avoids adding extra virtual machines to a retiring physical machine. The second rule speeds up the consolidation process and enables Dynamic Round-Robin to shutdown physical machines,	1.Reduce power consumption. 2.Save 3% more power than power-sever strategy.	1.Perform well for low incoming rate of VM. 2.Low performance in busy server.
Hybrid[1]	Combination of Dynamic Round Robin and First-Fit algorithms	1.Reduce power consumption. 2.Easy to implement. 3.Response time is high.	1.Not Suitable for low incoming rate of virtual machine.
PALB[2]	maintains the state of all compute nodes, and based on utilization percentages, decides the number of compute nodes that should be operating.	1.Simple 2. Easy to implement. 3. save energy.	1.Low performance. and scalability. 2.Only for local cloud.
ESCE[3]	The random selection based distributed problem round robin. Selection depend on least load.	1.Response time is high. 2.Processing time also high. 3.Simple and easy to implement.	1.Consume more power. 2.Does not consider fault tolerance
ECTC[4]	Two heuristics check every resource and identify most energy efficient resource for that task.	1.Energy consumption is reduced.	1. Difficult to implement. 2.Low performance.
MaxUtil[4]	Task consolidation decision based on resource utilization.	1.Energy consumption is reduced. 2.Utilization of small no of resources.	1. Difficult to implement.
ACCLB[7]	Uses small-world and scale-free characteristics of complex network to achieve better load balancing	1.Overcomes heterogeneity 2. Adaptive to dynamic environments 3. Excellent in fault tolerance 4. Good scalability	1.Used in complex networks only. 2. Does not save energy.
MCMF[8]	It is based on a directed graph representation of the dynamic resource allocation problem.	1. Simple 2. Easy to implement. 3. Cost effective in resource utilization.	1.Cost of networking and migration is more. 2.Low performance and scalability.
Join-Idle-Queue[12]	1.First find availability of the idle processors at each dispatcher 2. Then assigns jobs to processors to reduce average queue length of jobs at each processor	1. Effectively reduces the system load 2. Incurs no communica-tion overhead at job arrivals.	1. Does not increase actual response Times. 2.Consumes more energy.
DAIRS[9]	Treats CPU, memory, and Networks bandwidth integrated for both physical and virtual machine.	1.Good performance in total imbalance level of cloud data center	1.Average running time. 2.Not energy efficient.

VI. CONCLUSION

Cloud Computing has widely been adopted by the industry or organization though there are many existing issues like Load Balancing, Virtual Machine Consolidation, Energy Management, etc. which have not been fully implemented. Central to these issues is the issue of load balancing, that is required to distribute the excess dynamic local workload equally to all the nodes in the whole Cloud to achieve a high user satisfaction. It also ensures that every computing resource is distributed efficiently and fairly. Existing Load Balancing techniques that have been studied, mainly focus on reducing overhead, service response time and improving performance etc., and some of the techniques have considered the energy consumption factors. Therefore, there is a need to develop an Energy-aware load balancing technique that can improve the performance of cloud computing along with maximum resource utilization, in turn reducing energy consumption.

References

[1].Ching-Chi Lin, Pangfeng Liu, Jan-Jan Wu. "Energy-Efficient Virtual Machine Provision Algorithm for Cloud System", IEEE 4th International Conference on Cloud Computing,81-88, 09/2011.

[2]. Jeffrey M. Galloway, Karl L. Smith, Susan S. Vrbsky. "Power Aware Load Balancing for Cloud Computing", Proceedings of the World Congress on Engineering and Computer Science 2011 Vol. IWCECS 2011, October 19-21, San Francisco, USA,.

[3]. Jaspreet kaur "Comparison of load balancing algorithms in a Cloud" 2012, International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com Vol. 2, Issue 3, pp.1169-1173.

[4].R. Yamini, "Power Management In Cloud Computing Using Green Algorithm", IEEE-International Conference On Advances In Engineering, Science And Management (ICAESM-2012), March 30, 31,2012.pp-128-133.

[5]. R. Yamini, "Energy Aware Green Task Assignment Algorithm In Clouds", Ianternational Journal For Research In Science And Advance Technology, Issue-1,Volume-1,pp-23-29.

[6]. Anton Beloglazov, Rajkumar Buyya, "Managing Overload Host For Dynamic Consolidation Of Virtual Machines Cloud Data Centers Under Quality Of Service Constraints", IEEE Transaction On Parallel And Distributed Systems, 2012.

[7]. Zehua Zhang, Xuejie Zhang, "A Load Balancing Mechanism Bassed On Ant Colony And Compel Network Theory In Open Cloud Computing Federation", IEEE-International Conference On Automation, May 2010, pp-240-243.

[8]. Makhlouf Hadji, Djamal Zeghlache, "Minimum Cost Maximum Flow Algorithm For Dynamic Resource Allocation In Cloud", IEEE-Fifth International Conference In Cloud Computing, Aug-2012, pp-876-882.

[9].Wenhong Tian, Yong Zhao,Minxian Xu, Chen Jing, "A Dynamic And Integrated Load Balancing Scheduling Algorithm For Cloud Data Center", Proceeding of IEEE CCIS Feb 2011, pp-311-315.

[10]. Zenon Chaczko , Venkatesh Mahadevan , Shahrzad Aslanzadeh and Christopher Mcdermid "Availability and Load Balancing in Cloud Computing", 2011 International Conference on Computer and Software Modeling IPCSIT vol.14 , IACSIT Press, Singapore.

[11]. Rajkumar Buyya, Anton Beloglazov, and Jemal Abawajy,"Energy-Efficient Management of Data Center Resources for Cloud Computing: A Vision, Architectural Elements, and Open Challenges", Proceedings of the 2010 International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA 2010), Las Vegas, USA, July 12-15, 2010.

[12]. Trieu C. Chieu, Hoi Chan, "Dynamic Resource Allocation Via Distributed Decisions In Cloud Environment", Eight IEEE International Conference on e-Business Engineering, Sept-2011, pp-125-130.

[13]. Sivadon Chaisiri, Bu-Sung Lee, "Optimization Of Resource Provisioning Cost In Cloud Computing", IEEE transaction on services computing, vol. 5, No. 2 June 2012

[14]. Ayman G. Fayoumi, "Performance Evaluation Of A Cloud Based Load Balancer Severing Pareto Traffic", Journal of Theoretical and Applied Information Technology ,15th October 2011. Vol. 32 No.1

[15]. Anton Beloglazov and Rajkumar Buyya, Adaptive Threshold-Based Approach for Energy-Efficient Consolidation of Virtual Machines in Cloud Data Centers, Proceedings of the 8th International Workshop on Middleware for Grids, Clouds and e-Science (MGC 2010, ACM Press, New York, USA), In conjunction with ACM/IFIP/USENIX 11th International Middleware Conference 2010, Bangalore, India, November 29 - December 3, 2010.

[16].OpenNebula http://opennibula.org/

[17]. Cloudbus http://www.cloudbus.org/

[18]. Nimbus http://www.nimbus.com/