Power Curtailment in Cloud Environment Utilising Load Balancing Machine Allocation

Javadpour, Amir; Wang, Guojun; Rezaei Badafshani, Samira; Chen, Shuhong

Published in:
IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation

DOI:
10.1109/SmartWorld.2018.00237

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2018

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Download date: 05-10-2019
Abstract— Cloud computing is a new technology to provide services users on the Internet. Companies providing services in this area are trying to offer attractive solutions and services business with the slogan desire to use this technology to reduce operating costs and increase their endurance. The increasing use of cloud computing with great benefits, also faces challenges as the most important of which can be noted consumable supplies. Load balancing algorithms are several parameters that can be used by the utility and productivity of the most important features to regard power consumption will be given. In this paper, a method based on network load balancing to optimize the energy consumption by using Dynamic Voltage and Frequency Scaling (DVFS). Load balancing algorithms and power energy management are absence maximum of workloads in the data center for service. This algorithm given time intervals by monitoring utilization rates of server hosts, the hosts have the additional burden reduction and load them into hosts with little time has passed.

Keywords- Dynamic Voltage and Frequency Scaling (DVFS), load balancing, Energy consumption, virtual machine.

I. INTRODUCTION

Cloud computing has developed so significantly during recent years, such that there are chances it is introduced as the fifth utility (after water, electricity, gas and telephone). This public service, like all others, provides the foundation for a computing service which can meet the needs of the society. This service will be distributed similar to other public utilities and every person can use it according to his needs regardless of where the service is hosted or how it is distributed[1].

Several computing patterns have been proposed so far; cluster computing[2], grid computing[3], Intrusion Detection in Cloud Environment based on Machine Learning Algorithms [4], mobile computing, health care monitoring in cloud [4]–[8] and the latest is cloud computing. The nomenclature for "cloud" stems from the fact that users and businesses can access the cloud computing from [9]–[11]result, the computing world is moving from software that can only run on a single computer, to services which are able to serve millions of people around the globe. Cloud computing is the latest solution of this type and provides software as a service for commercial use. Service providers benefit from a wide range of users and the users benefit from the cost reduction of cloud computing in comparison with in house software maintenance and installments[12]–[16]. However, due to the very high sensitivity required by some of the commercial uses of the software which are hosted by cloud, it is vital for service providers to provide service guarantees to users. Usually this guarantee is established by a quality of service contract between service provider and the user. Until two decades ago, computers were connected to each other in a cluster to simulate a supercomputer and create a very high processing power. This technique was so prevalent and was utilized in most IT departments. Clustering made it feasible that many computers were connected and communicate by means of specific protocols. The ultimate goal was to distribute the processing load among the involved machines. It did not make so much difference to the user as where the software was running. In other words, several clusters which might be in different areas are connected and comprise a net. One of the obstacles of migrating from cluster to grid model was data residency. Virtual machines can enhance performance of user applications without interrupting service to be transferred to another host. Also, it is possible to dynamically increase or decrease the amount of resources allocated to a client. Cloud providers to profit from providing services to gain customers and cloud users, such as organizations and enterprises, for the sake of profits due to the reduction or elimination of costs related to infrastructure maintenance services as in-house gain [17]. One of the major challenges in the field of cloud computing on the server cloud environment is maintaining optimal load balancing energy. Cloud Computing distributed servers include virtualization, distributed computing, networking, software and services on the Internet [18]. In the development of modern data centers in which there is a great desire to design centers, applications, providing them with the least dependence on hardware infrastructure, implemented and easily share resources. It is also desirable to allow the application of a set of resources to another without having to stop the implementation of the program. This characteristic of cloud computing infrastructure that virtualization technologies and live transfer (migration) environments, cloud computing (cloud computing) are met. Using these technologies increase the productivity of resources, we can control and manage the resources, improvements also imposed on them. The power and capabilities of virtualization technology can be used to transform data centers into cloud computing infrastructure and services centers to provide flexible and reliable IT equipped [11]. Due to the distributed nature of net, computing nodes may be anywhere around the globe. Paul wallis explains data residency consequence of such model as follows [19]:

"..."
This paper attends to reducing the power consumption of virtualized datacenters of cloud computing infrastructures. In section two, relative prior works are summarized. In the proposed method section, new algorithms are introduced to reduce the power consumption and increase efficiency and reliability. In the simulation section, the simulation results of a datacenter which uses the proposed algorithm, indicates improvement in power consumption and an increase in reliability as well.

II. RELATED WORKS

Today, most of the computer software utilized in business or science sectors requires a high performance computing infrastructure. This need requires very large scale datacenters which have high power consumption. Although improvements have been made to the hardware power efficiency, the overall power requirements are still pretty high due to the heavy demand for computing resources. Apart from the exorbitant cost building over designed datacenters to encounter load peaks while having a low average efficiency of the devices, results in high power consumption. In addition, lack of cooling systems or their failure might lead to excessive heating of the datacenter and reduce its reliability as well as device life time. Besides, a high power consumption, causes more CO2 gas to be emitted and eventually severs the greenhouse effect [20].

In order to improve efficiency, several methods have been employed so far. These methods include: betterment of practical software algorithms and using high efficiency hardware [21].

By virtualizing technology, several virtual machines can be created on a physical server. Hence, the required hardware reduces and the efficiency increases. Cloud computing uses virtualization technology to provide resources according to the demand by users. Using this technology in cloud computing reduces power consumption due to the below reasons:

Better exploitation of resources [22];

Elimination of physical dependency – virtual machines can be moved and run in a location with lower energy price.

Variable scale of resource usage – usage of resources can scale up or down according to the demand.

In article [23] power consumption reduction in virtualized cloud datacenters has been studied. In this paper, the authors propose an efficient resource management system for the cloud datacenter which reduces the operational cost and guarantees the quality of service. Its power saving is implemented by continuous control of virtualized machines using the current utilization method. The main tool used in this process is live transfer of virtualized machines of cloud computing. Live transfer is a capability to transfer virtual machines among physical hosts with small overhead that provides the flexibility of a dynamic resource allocation based on the current demand, current resources and policies.

In [11], the method are used to reduce network power using the time scheduling and DVFS method. This architecture is proposed to manage power consumption in virtualized datacenters which has local and global scopes. In local level, the power management capabilities of the guest's operating are used. On the other hand, in the global level, the global policies of utilizing live transfer are applied. In [24], efficient allocation of virtual machines' power in virtual asymmetric computing environments has been studied. In the propose method of this paper, min, max and share parameters of the virtual machine manager (VMM) are used which show the minimum, maximum and relative value of the CPU which each of the VMs can share respectively. This method is only useful for enterprise solutions or private clouds where SLA value is not important and there are no requirements with this regard. In [25], a symmetric and scalable framework is proposed to create cloud applications which can endure by component ranking technique. Component ranking is based on the logic that the components called by cloud more frequently are of a higher priority. The goal is to identify the important components of the applications on the cloud where on the contrast the target of methods like Google Page Ranking and SPARS is to search a webpage or reusable codes respectively. As mentioned earlier, a framework to improve reliability is proposed in this paper which is elaborated in this section. Figure 1 depicts FTCloud architecture which consists two main parts: ranking and selecting the optimized reliability method. FTCloud development procedure is as follows:

The initial architecture of an application is provided by the system designer and a graph element of the cloud program will be generated.

- An element sorting algorithm is applied for calculating the importance of each element of the cloud.
- The more important elements are determined.
- Efficiency of various fault tolerance strategies is evaluated and the best option is selected for every important element.
- The improved architecture is generated and the results of sorting are returned back to the system designer.

The reliability of cloud modules does not solely rely on the system itself. It also depends on the node or host which is running the applications on the cloud and also the unpredictable internet. This issue is more critical for voluntary resource clouds. In this paper a byzantine fault tolerance framework (BFTCloud) is proposed which focuses on creating reliable systems in voluntary resource cloud environments[26]. BFTCloud guarantees the reliability of the systems until the time when f resources out of 3f+1 resources have failed. This fault might be due to resource crashing, malware activity etc. The voluntary resource cloud infrastructure is depicted in figure 2 and includes the below modules[26]: Computing resources are nodes or hosts in the cloud which are prone crash. In contrast to the powerful and efficient nodes which are managed by major cloud providers, the underlying nodes of this cloud is provided by the volunteers which are usually very dynamic and very cheap and have low computing power and low reliability. The inter-relations of the modules are not reliable. In contrast to the large and centralized cloud infrastructures which maintain a high speed and reliable inter-connection among the modules, underlying nodes of the cloud are constructed by voluntary resources which usually have unpredictable and unreliable links. Faults or connection errors.
such as time-out could affect voluntary resource cloud’s applications very severely. Primary node selection: after receiving the request from the cloud module, a node from the cloud is selected as the primary node. This node is chosen according to quality of service requirements and based on the primary node selection algorithm. Replica node selection: again, in this phase, a set of nodes are selected by application of replica node selection algorithm while quality service requirements are considered. After this selection, the primary node sends the execute request to all of the replica nodes. The set of replica nodes and primary node form the BFT altogether.

- **Execution request**: in this phase, all of the BFT elements, run the request locally and return their results to the cloud module. After collecting all the responses from BFT in a specific timeframe, the consistency of the nodes is judged.

- **Updating the primary node**: in this phase, the faulty primary node in BFT group is detected and then substituted by the new selected node.

- **Updating the replica node**: in this phase, faulty replica nodes in the BFT group are detected and then substituted by proper nodes from the cloud after considering the information generated by the execution phase.

In designing and developing modern datacenters, there is a great tendency to design centers which have the minimum dependency of the applications on the hardware infrastructure. As a result a high level of sharing among resources can be maintained. In addition, it is favorable to be able to transfer an application from a set of resources to another without halting the process. These are features of the cloud computing infrastructure mainly conceived by virtualization technology and live migration in cloud computing environments [27]. Utilizing these technologies, not only increases the efficiency of resources, but also makes it possible to optimize them through control and management methods. Virtualization technology can be used to convert datacenters to cloud computing infrastructure and use them to provide flexible and reliable IT services. Using this technology, the hardware resources of the datacenter are virtualized and can be shared among several systems. In this case virtual machines and virtual disks are in charge of processing. In the other word, applications are independent to machines. There are several levels in modern virtualization. Using datacenters in cloud is a system virtualized which make different operating systems to run concurrently. Xen and VMWare [28], [29] are examples of system virtualized. Virtual Machine Monitor (VMM) manages the hardware resources in different virtual machines. In this article we are going to present a load balancing technique between several hosts and at the same time can reduce power consumption. Load balancing assures that load distribution among hosts is approximately uniform and eliminates the situation where a server bears the majority of processing work[30]. It is important to make sure that datacenters prevent hotspots and balanced the high load in hotspots with transferring some loads to other hosts [31]. The high energy consumption caused by massive growth in demand of high throughput applications costs a lot for cloud service providers. Recent studies [32], [33] showed remarkable and impressive costs for energy consumption and cooling systems. There are several ways to reduce energy consumption in cloud computing centers, this article uses the combination of dynamic voltage and frequency scaling (DVFS).

III. THE PROPOSED METHOD: DYNAMIC LOAD BALANCING

In this article we are going to present a load balancing technique between several hosts and at the same time can reduce power consumption. Load balancing assures that load distribution among hosts is approximately uniform and eliminates the situation where a server bears the majority of processing work[30]. It is important to make sure that datacenters prevent hotspots and balanced the high load in hotspots with transferring some loads to other hosts [31]. The high energy consumption caused by massive growth in demand of high throughput applications costs a lot for cloud service providers. Recent studies [32], [33] showed remarkable and impressive costs for energy consumption and cooling systems. There are several ways to reduce energy consumption in cloud computing centers, this article uses the combination of dynamic voltage and frequency scaling (DVFS) [34].

IV. PROPOSED METHOD

In this section we are going to present load balancing algorithms in two different situations. The first case is a data center with the maximum loads and all the resources are occupied. The next one is a normal condition which the datacenter dose not uses the maximum productivity and there might be some idle resources. Load balancing and Load-Power Aware (LPA) algorithms are examined. For this purpose, we use the proposed architecture in [26] which is inspired by the topology of the data center (Figure 2). Physical infrastructure of datacenters in this architecture is formed by clusters of servers or computing hosts, network and storage arrays and local administration. Local management is a host which controls other computing hosts and sources. The virtual machines on the computing hosts run the user applications. The amount of RAM, CPU power of each host and the network bandwidth are identified. Each user request should specify the number of required VMs, amount of RAM, MIPS and network bandwidth. The local administration
checks the efficiency of available resources on the hosts and directs the user request to perfect host. If the requested VMs are more than one, the local administration can select different host for each VM. The load balancing and power management techniques for reducing the consuming power would be executed by using live migration [35] in local administration. It guaranties the VM transformation between two hosts without any service interruption. With this technology, it is possible to balance the load of high loaded hosts or switching some machines off to manage power consumption. The following paragraphs contain some definitions about power consumption and CPU frequencies.

Table1: Load balancing with DVFS Algorithm.

Check client Rts
Calculation \( LoadFraction_{vm,host} = \frac{p(f)}{P_{max}} \), in PM
Set a Active VM in Servers
Calculation DVFS Calculation Attractiveness
\( Attractiveness_{host} = IBscore_{host} \times LoadFraction_{vm,host} \)
Checked Imbalance score
Check Existing VM Ability to Response Rts
If, Calculation DVFS<VM Ability to Response Then
Check the user and wait for new call in time slots
Else If, Calculation DVFS> VM Ability to Response
Check and Calculation DVFS parameter
Checked Imbalance score
Find the highest VM power With Considering the DVFS table
Immigration apply to the highest PM with highest VM power
End
End
Waiting for new requests.

Modern CPUs can run at different speeds and frequencies through the use of dynamic frequency scaling (DFS), dynamic voltage scaling (DVS), or a combination of dynamic voltage and frequency scaling. This technique leads to a nonlinear relationship between frequency and power. According to [36] at every level of productivity, the relationship between the frequency and power could be well approximated by quadratic model, which means we can write:

\[
 p(f) = P_{min} + \alpha (f - f_{min})^2
\]

(1)

\[
 \alpha = \frac{P_{max} - P_{min}}{(f_{max} - f_{min})^2}
\]

(2)

In equation (1), the power (P) is in watts (W), frequency (f) is in GHz and \( \alpha \) is a coefficient in W/(GHz)^2. We have assumed in our model that the hosts CPU inside the data center follow such nonlinear relationship. CPUs are executed with limited amounts of operating frequencies in the range [fmin, fmax]. It is a tradeoff between performance and costs. In our model, all servers or host, despite different amount of CPU resources follow equation 1. In the case of low load, hosts can switch to passive mode. It is demonstrated in equation 2 than even if there is no load in a host, it consumes \( P_{min} \) to keep the host in the active mode. So switching hosts to passive mode can have a significant impact on power consumption. Imbalance score is defined as follow to calculate the overload state.

\[
 IBscore(u, T_{upper}) = e^{(u-T_{upper})}
\]

(3)

This equation demonstrates the amount of load on every host. \( u \) is an efficiency measure like CPU or RAM efficiency and \( T_{upper} \) is the upper threshold for host efficiency. If \( u \) is less than \( T_{upper} \), \((u-T_{upper})\) would be negative and the value of IBscore decreases.
production between LoadFraction and IBscore computes host attractiveness. Proposed algorithm is described in Table 1.

V. SIMULATIONS

Process simulation will be discussed in next section detail has been paid. DVFS using open source software packages loaded in clouds basic information and then apply the proposed method for driving simulation. In Table 2 the parameters used in the proposed method is discussed. System design for large-scale cloud data center be different experiments run on cloud infrastructure to enable real different algorithms to be evaluated. However, implementation of such experiments on real infrastructure is difficult, especially if you need such experiments be repeated several times. For this reason, the simulation can be a better method to evaluate different policies and algorithms are proposed. We have to evaluate the algorithm, CloudSim simulator has been used [38]. The simulation model based virtualization application enables to manage resources and applications that we have here a short description about it. The main idea of your toolkit Simjava is a java library for the environment and has been developed at the University of Edinburg, has taken (Figure 3).

![Diagram](image)

Figure 3: Compare power consumption in the proposed method (Using DVFS) and without DVFS

It has been used architecture for cloud computing architecture is similar to that in the proposed method. Although if the user has more than one VM is requested, local administration units can be different for each VM hosts are chosen. As mentioned, transfer or migrate live machines is done after selecting machines to transfer, as the hosts algorithm to select the virtual machine. Of the host, the host will be chosen policies the value of the utility function is less or more appropriate. Recent condition prevents compaction is the creation of conditions. After selecting the destination host, the transition begins and after the transfer, implementation of new host continues and resources are open source host. As we said we have used CloudSim and MATLAB simulation to evaluate results. In fact, the core of this simulator for model loads balancing algorithms. As well as the impact of technology on power consumption and computing overhead added DVFS expanded on it. To show the amount of overhead, in each period we calculate the amount SLA data center. Cloud computing simulation has data center and 100 host virtualization capabilities. In fact, assumed that virtualization like Xen installed on them that can share resources. Each of the single-core hosting model, which is equal to 1000, 2000 or 3000 MIPS have the ability to be 1000, 2000 or 3000 million instructions per second do, 8GB of RAM and 1TB of memory per host as well as their storage. The Centre has created 220 virtual machines, each running a single application. Therefore, any application is composed of 1,500,000 to instructions. As before mentioned, the purpose of load balancing algorithm, a balanced distribution of workload among the host of the gathering, which prevents other times of the hotspot is on a host. In 1500 second (Figure 4(A)) the total execution time for transfer or migration of virtual machines and an average of violating the SLA, 66/13% respectively. In case of violation of the SLA and the transfer of VM will have to use this algorithm. Since power consumption enormous cost to impose cloud computing service provider in Standby mode simulation technology as well as taking home DVFS Unemployment has been used. A result of a comparison of Figures (Figure 4(A,B,C)) use of this technology is quite clear. The total amount of power consumption when using DVFS is 31/20 KWh and without this technology is 62/58 KWh. Therefore, the use of this technology can be up to 28.84% resulting in reduced power consumption.
VI. CONCLUSION

Cloud providers to profit from providing services to gain customers and cloud users, such as organizations and enterprises, for the sake of profits due to the reduction or elimination of costs related to infrastructure maintenance services as in-house gain. Virtualization to isolate virtual machines from each other so that each virtual machine can be made possible depending on user needs in a number of hosts to be configured independently. In this paper, a method is presented using DVFS technology. As discussed in the results, using the DVFS method has been effective in reducing power consumption. For future work of this study, one can consider the use of meat-heuristic methods for managing power consumption reduction.

ACKNOWLEDGMENTS

This work is supported in part by the National Natural Science Foundation of China under Grants 61632009 & 61472451, in part by the Guangdong Provincial Natural Science Foundation under Grant 2017A030308006, and High-Level Talents Program of Higher Education in Guangdong Province under Grant 2016ZJ01.

REFERENCES


