ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Energy Load Balancing and Scaling in Cloud System

Akshatha.B
B.E Student,
Dept. of Computer Science,
Ballari Institute of Technology & Management.

Arun Kumar.R
B.E Student,
Dept. of Computer Science,
Ballari Institute of Technology & Management.

Dhanesh.C
B.E Student,
Dept. of Computer Science,
Ballari Institute of Technology & Management.

Kiran Kumar.J
B.E Student,
Dept. of Computer Science,
Ballari Institute of Technology & Management.

Dr Rajashree V Biradar
Professor,
Dept. of Computer Science,
Ballari Institute of Technology & Management.

Abstract

Cloud computing has emerged as one of the most influential paradigms in the IT industry in recent years.

The energy consumption of computer and communication system does not scale linearly with the workload. A system uses a significant amount of energy even when idle or lightly loaded.

The basic approach of solution to resource management in large data centers is to distribute the load on a subset of servers and, whenever possible switch the rest of the servers to one of the possible sleep states. In this implemented the concept of load balancing aiming to optimize the energy consumption of a large-scale system by distributing the workload evenly to the smallest set of servers operating at an optimal energy level.

Keywords— ***load balancing, application scaling, idle servers, server consolidation, energy proportional systems***

INTRODUCTION

In the last few years packaging computing cycles and storeing them as a metered service became a reality. Large farms of computing and storage platforms have been assembled and a fair number of Cloud Service Providers (CSPs) computing services based on three cloud delivery models SaaS (Software as a Service), PaaS (Platform as a Service), and IaaS (Infrastructure as a Service). Warehouse-scale computers (WSCs) are the building blocks of a cloud infrastructure. A hierarchy of networks connect 50 to 100 servers in a WSC. The servers are housed in racks; typically, the 48 servers in a rack are connected by a 48-port Gigabit Ethernet switch.

The switch has two to eight up-links which go to the higher level switches in the network hierarchy. Cloud elasticity, the ability to use as many resources as needed at any given time, and low cost, a user is charged only for the resources it consumes, represents solid incentives for many organizations to migrate their computational activities to a public cloud. The number of CSPs of spectrum of services, and the number of cloud users have increased dramatically during the last few years.

For example, in 2007 the EC2 (Elastic Computing Cloud) was the 1st service provided by AWS (Amazon Web Services); few years later, in 2012, AWS was used by businesses in 200 countries. Amazon's S3 (Simple Storage Service) has surpassed two trillion objects and routinely runs more than 1.1 million peak requests per second. Elastic MapReduce has launched 5:5 million clusters since May 2010 when the service started. The rapid expansion of the cloud computing has a significant impact on the energy consumption in US and the world. The costs for energy and for cooling large-scale data centers are significant and are expected to increase in the future. In 2006, the 6000 data centers in the U.S.

reportedly consumed 61 to 109 kWh of energy, 1:5% of all electricity consumption in the country, at a cost of $4:5 billion. The energy consumption of data centers and of the network infrastructure is predicted to reach 10; 300 TWh/year (1 TWh = 109 kWh) in 2030, based on 2010 levels. These increases are expected in spite of the extraordinary reduction in energy requirements for computing activities. Idle and under-utilized servers contribute significantly to wasted energy, see Section 2. A 2010 survey reports that idle servers contribute 11 million tons of unnecessary CO2 emissions each year and that the total yearly costs for idle servers is $19 billion. Recently, Gartner Research reported that the average server utilization in large data-centers is 18%, while the utilization of x86 servers is even lower, 12%. These results show earlier estimations that the average server utilization is in the 10 to 30% range.

The concept of load balancing was introduced when the distributed computing systems were implemented. It implies that to evenly distribute the workload to a set of servers to maximize the throughput, minimize the response time, and increase the system faults by avoiding and overloading the systems. An important strategy for energy reduction is concentrating the load on a subset of servers and, whenever possible, switching the rest of them to a state with a low energy consumption. This observation implies that the traditional concept of load balancing in a large-scale system could be reformulated as follows: distribute evenly the workload to the smallest set of servers operating at optimal or near-optimal energy levels, while observing the Service Level Agreement (SLA) between the CSP and a cloud user.

An optimal energy level is one when the performance per Watt of power is maximized. Scaling is the process of allocating additional resources to a cloud application in response to a request consistent with the SLA. We distinguish two scaling modes, horizontal and vertical scaling. Horizontal scaling is the most common mode of scaling on a cloud; it is done by increasing the number of Virtual Machines (VMs) when the load of applications increases and reducing this number when the load decreases. Load balancing is critical for this mode of operation. Vertical scaling keeps the number of VMs of an application constant, but increases the amount of resources allocated to each one of them. This can be done either by migrating the VMs to more powerful servers or by keeping the VMs on the same servers, but increasing their share of the server capacity. The alternative involves additional overhead; the VM is stopped, a snapshot is taken, the idle is migrated to a more powerful server, and the VM is restarted at the new site. The alternative to the wasteful resource management policy when the servers are always on, regardless of their load, is to develop energy-aware load balancing and scaling policies. Such policies combine dynamic power management with load balancing and attempt to identify servers operating outside their optimal energy regime and decide if and when they should be switched to a sleep state or what other actions should be taken to optimize the energy consumption.

The vast literature on energy-aware resource management concepts and ideas.

Some of the questions posed by energy-aware load balancing and application scaling are: (a) Under what conditions should a server be switched to a sleep state? (b) What sleep state should the server be switched to? (c) How much energy is necessary to switch a server to a sleep state and then switch it back to an active state? (d) How much time it takes to switch a server to a running state from a sleep state? (e) How much energy is necessary for migrating a VM running on a server to another one? (f) How much energy is necessary for starting the VM on the target server? (g) How to choose the target where the VM should migrate to? (h) How much time does it takes to migrate a VM? The answers to some of these questions depend on the server's hardware and software, including the virtual machine monitor and the operating systems, and change as the technology evolves and energy awareness becomes increasingly more important. In this paper we are concerned with high level policies which, to some extent are independent of the specific attributes of the server's hardware and, due to space limitation, we only discuss (a), (b), and (g). We assume that the workload is predictable, has no spikes, and that the demand of an application for additional computing power during an evaluation cycle is limited. We also assume a clustered organization, typical for existing cloud infrastructure.

There are three primary contributions of this paper: (1) a new model of cloud servers that is based on different operating regimes with various degrees of energy efficiency (processing power versus energy consumption); (2) a novel algorithm that performs load balancing and application scaling to maximize the number of servers operating in the energy-optimal regime; and (3) analysis and comparison of techniques for load balancing and application scaling using three differently-sized clusters and two different average load problems. Models for energy-aware resource management and application placement policies and the mechanisms to enforce these policies such as the ones introduced in this paper can be evaluated, based on published data, or through a combination of these techniques. Analytical models can be used to derive high-level insight on the behavior of the system in a very short time but the biggest challenge is in determining the values of the parameters; while the results from an analytical model can give a good approximation of the relative trends to expect, there may be significant errors in the absolute predictions.

Experimental Data is collected on small-scale systems; such experiments provide useful performance data for individual system components but no insights on the interaction between the system and applications and the scalability of the policies. Trace-based workload analysis and are very useful though they provide information for a particular experimental set-up, hardware configuration, and applications. Typically trace based simulation need more time to produce results. Traces can also be very large and it is hard to generate representative traces from one class of machines that will be valid for all the classes of simulated machines. To evaluate the energy aware load balancing and application scaling policies and mechanisms introduced in this paper we chose simulation using data published in the literature. Operating efficiency of a system and server consolidation are discussed later, respectively. The model described in later section introduces the operating regimes of a processors and the conditions when to switch a server to a sleep state. Load balancing and scaling algorithms suitable for a clustered cloud organization based on the model are presented ; these algorithms aim to optimize the energy efficiency and to balance the load. Simulations experiments and conclusions are covered.

LITERATURE SURVEY

Energy Efficient Resource Management in Virtualized Cloud Data Centers

Rapid growth of the demand for computational power by scientific, business and web-applications has led to the creation of large-scale data centers consuming enormous amounts of electrical power. We propose an energy efficient resource management system for virtualized Cloud data centers that reduces operational costs and provides required Quality of Service (QoS). Energy savings are achieved by continuous consolidation of VMs according to current utilization of resources, virtual network topologies established between VMs and thermal state of computing nodes. We present first results of simulation-driven evaluation of heuristics for dynamic reallocation of VMs using live migration according to current requirements for CPU performance.

The results show that the proposed technique brings substantial energy savings, while ensuring reliable QoS. This justifies further investigation and development of the proposed resource management system.

Managing Overloaded Hosts for Dynamic Consolidation of Virtual Machines in Cloud Data Centers under Quality of Service Constraints

Dynamic consolidation of virtual machines (VMs) is an effective way to improve the utilization of resources and energy efficiency in cloud data centers. Determining when it is best to reallocate VMs from an overloaded host is an aspect of dynamic VM consolidation that directly influences the resource utilization and quality of service (QoS) delivered by the system. The influence on the QoS is explained by the fact that server overloads cause resource shortages and performance degradation of applications. Current solutions to the problem of host overload detection are generally heuristic based, or rely on statistical analysis of historical data. The limitations of these approaches are that they lead to suboptimal results and do not allow explicit specification of a QoS goal. We propose a novel approach that for any known stationary workload and a given state configuration optimally solves the problem of host overload detection by maximizing the mean intermigration time under the specified QoS goal based on a Markov chain model. We heuristically adapt the algorithm to handle unknown non stationary workloads using the Multi size Sliding Window workload estimation technique.

Through simulations with workload traces from more than a thousand Planet Lab VMs, we show that our approach outperforms the best benchmark algorithm and provides approximately 88 percent of the performance of the optimal offline algorithm.

Energy-Efficient Cooperative Protocol for Wireless Networks

In cooperative networks, transmitting and receiving nodes recruit neighbouring nodes to help in communication. We model a cooperative transmission link in wireless networks and then it forms a cluster at transmitter end and a receiver end. In this paper, we propose a cooperative communication protocol for formation of these clusters for cooperative transmission of data. This paper proposes a new reliable and energy efficient cooperative protocol, and we analysed the robustness of the protocol to data packet loss, along with the trade off between energy consumption and error rate.

The analysis results are used to compare the energy savings and the end-to-end robustness of our protocol with two non-cooperative schemes, such as one non-cooperative named as disjoint-paths and one another cooperative scheme named as Cooperation along Non-cooperative path (CAN). The reduction in error rate and the energy savings translate into increased lifetime of cooperative sensor networks.

EXISTING SYSTEM

An important strategy for energy reduction is concentrating the load on a subset of servers and, whenever possible, switching the rest of them to a state with low energy consumption. This observation implies that the traditional concept of load balancing in a large-scale system could be reformulated as follows: distribute evenly the workload to the smallest set of servers operating at optimal or near-optimal energy levels, while observing the Service Level Agreement (SLA) between the CSP and a cloud user. An optimal energy level is one when the performance per Watt of power is maximized.

In order to integrate business requirements and application level needs, in terms of Quality of Service (QoS), cloud service provisioning is regulated by Service Level Agreements (SLAs): contracts between clients and providers that express the price for a service, the QoS levels required during the service provisioning, and the penalties associated with the SLA violations.

In such a context, performance evaluation plays a key role allowing system managers to evaluate the effects of different resource management strategies on the data center functioning and to predict the corresponding costs/benefits.

DISADVANTAGES OF EXISTING SYSTEM:

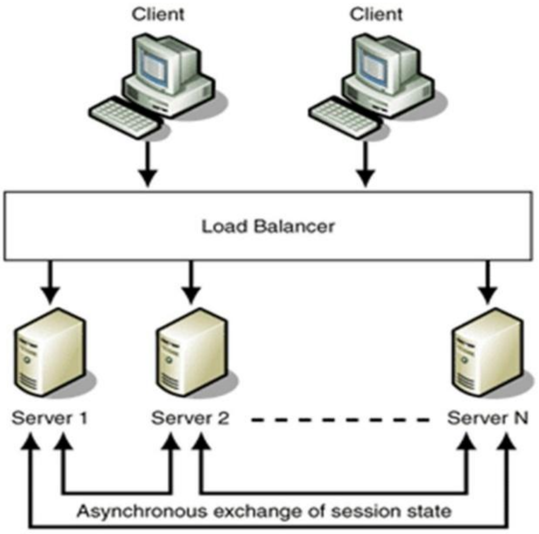
- On-the-field experiments are mainly focused on the offered QoS, they are based on a black box approach that makes difficult to correlate obtained data to the internal resource management strategies implemented by the system provider.
- Simulation does not allow to conduct comprehensive analyses of the system performance due to the great number of parameters that have to be investigated.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

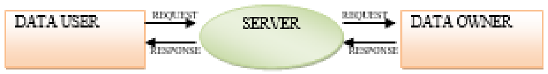
PROPOSED SYSTEM:

- A new model of cloud servers that is based on different operating regimes with various degrees of \energy efficiency" (processing power versus energy consumption);
- A novel algorithm that performs load balancing and application scaling to maximize the number of servers operating in the energy-optimal regime; and analysis and comparison of techniques for load balancing and application scaling using three differently-sized clusters and two different average load profiles.
- The objective of the algorithms is to ensure that the largest possible number of active servers operate within the boundaries of their respective optimal operating regime. The actions implementing this policy are: (a) migrate VMs from a server operating in the undesirable-low regime and then switch the server to a sleep state; (b) switch an idle server to a sleep state and reactivate servers in a sleep state when the cluster load increases; (c) migrate the VMs from an overloaded server, a server operating in the undesirable-high regime with applications predicted to increase their demands for computing in the next reallocation cycles.

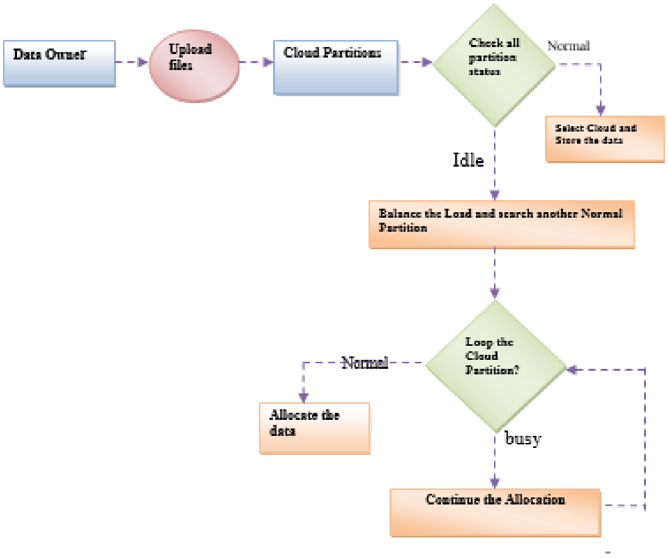
System architecture



Data flow diagram level 0



Data flow diagram level 1


Data Owner
Upload files
Cloud Partitions
Check all partition status
Normal
Select Cloud and Store the data
Idle
Balance the Load and search another Normal Partition
Loop the Cloud Partition?
Normal
Allocate the data
busy
Continue the Allocation

MODULES DESCRIPTION

Load balancing

The concept of load balancing" dates back to the time when the first distributed computing systems were implemented. It means exactly what the name implies, to evenly distribute the workload to a set of servers to maximize the throughput, minimize the response time, and increase the system resilience to faults by avoiding overloading the systems.

Owner and User Registration

The Owner has to register his /her name, password, email, phone no, address. If the owner registration is successful then owner will receive an acknowledgement .After the registration, Owner has to login with his user name and password .if the owner login is successful then he can upload a file with .Java extension according to the selected cloud server. User also has to follow the above procedure.

Display Module

The owner uploaded file will be downloaded by the end user. So they can able to see transaction details of the uploaded file at a specific time and he can also view the attackers like who has attacked the file .User can also view the bar graph of load balancing.

Idle servers

Idle and under-utilized servers contribute significantly to wasted energy, see Section survey reports that idle servers contribute 11 million tons of unnecessary CO2 emissions each year and that the total yearly costs for idle servers is billion. An energy-proportional system consumes no energy when idle, very little energy under a light load, and gradually, more energy as the load increases.

Server consolidation:

The term server consolidation is used to describe: switching idle and lightly loaded systems to a sleep state; workload migration to prevent overloading of systems any optimization of cloud performance and energy efficiency by redistributing the workload discussed in Section For example, when deciding to migrate some of the VMs running on a server or to switch a server to a sleep state, we can adopt a conservative policy similar to the one advocated by auto scaling to save energy. Predictive policies will be used to allow a server to operate in a suboptimal regime when historical data regarding its workload indicates that it is likely to return to the optimal regime.

Energy proportional systems:

The energy efficiency of a system is captured by the ratio performance per Watt of power." During the last two decades the performance of computing systems has increased much faster than their energy efficiency Energy proportional systems. In an ideal world, the energy consumed by an idle system should be near zero and grow linearly with the system load. In real life, even systems whose energy requirements scale linearly, when idle, use more than half the energy they use at full load. Data collected over a long period of time shows that the typical operating regime for data center servers is far from an optimal energy consumption regime.

The dynamic range I s the difference between the upper and the lower limits of the energy consumption of a system as a function of the load placed on the system. A large dynamic range means that a system is able to operate at a lower fraction of its peak energy when its load is low

CONCLUSION

The whole idea is to implement the optimal system of Energy Load Balancing to work efficiently from the current available systems. We conclude optimal usage of load balancers in the cloud systems are deployed, which can route the energy leads in the partitions in a selective manner which can make the download process of multiple connections easy and by having many partitions the efficiency will be decreased, Hence we would deploy only three partitions for optimal efficiency**.**

REFERENCES

[1] D. Ardagna, B. Panicucci, M. Trubian, and L. Zhang. \Energy-aware autonomic resource allocation in multitier virtualized environments." IEEE Trans. on Services Computing, 5(1):2{19, 2012.

[2] J. Baliga, R.W.A. Ayre, K. Hinton, and R.S. Tucker. \Green cloud computing: balancing energy in processing, storage, and transport." Proc. IEEE, 99(1):149-167, 2011.

[3] L. A. Barroso and U. H□ozle. \The case for energyproportional computing." IEEE Computer, 40(12):33{ 37, 2007.

[4] L. A. Barossso, J. Clidaras, and U.H□ozle. The Data-center as a Computer; an Introduction to the Design of Warehouse-Scale Machines. (Second Edition). Morgan & Claypool, 2013.

[5] A. Beloglazov, R. Buyya \Energy e_cient resource management in virtualized cloud data centers."

Proceedings of the 2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Comp., 2010.

[6] A. Beloglazov, J. Abawajy, R. Buyya. \Energy-aware resource allocation heuristics for e_cient management of data centers for Cloud computing." Future Generation Computer Systems, 28(5):755-768, 2012.

[7] A. Beloglazov and R. Buyya. \Managing overloaded hosts for dynamic consolidation on virtual machines in cloud centers under quality of service constraints." IEEE Trans. on Parallel and Distributed Systems, 24(7):1366-1379, 2013.

[8] M. Blackburn and A. Hawkins. \Unused server surveyresultsanalysis."www.thegreengrid.org/media/White Papers/Unused%20Server%20Study WP 101910 v1. ashx?lang=en (Accessed on December 6, 2013).

[9] M. Elhawary and Z. J. Haas. \Energy-e_cient protocol for cooperative networks." IEEE/ACM Trans. on Net- working, 19(2):561{574, 2011.

[10] A. Gandhi, M. Harchol-Balter, R. Raghunathan, and M.Kozuch. \AutoScale: dynamic, robust capacity management for multi-tier data centers." ACM Trans. On Computer Systems, 30(4):1{26, 2012.

[11] A. Gandhi, M. Harchol-Balter, R. Raghunathan, and M.Kozuch. \Are sleep states e_ective in data centers?" Proc. Int. Conf. on Green Comp., pp. 1{10, 2012.

[12] D. Gmach, J. Rolia, L. Cherkasova, G. Belrose, T. Tucricchi, and A. Kemper. \An integrated approach to resource pool management: policies, e_ciency, and quality metrics." Proc. Int. Conf. on Dependable Systems and Networks, pp. 326{335, 2008.

[13]Google.\Google'sgreencomputing:e_ciencyascale."untrusted /www.google.com/en/us/green/pdfs/google - green-computing.pdf (Accessed on August 29, 013).

[14] V. Gupta and M. Harchol-Balter. \Self-adaptive admission control policies for resource-sharing systems." Proc. 11th Int. Joint Conf. Measurement and Modeling Com- puter Systems (SIGMETRICS'09), pp. 311{322, 2009.

[15] K. Hasebe, T. Niwa, A. Sugiki, and K. Kato. \Powersaving in large-scale storage systems with data migration." Proc IEEE 2nd Int. Conf. on Cloud Comp. Tech- nology and Science, pp. 266{273, 2010.

[16] B. Heller, S. Seetharaman, P. Mahadevan, Y. Yiakoumis, P. Sharma, S. Banerjee, and N. McKeown. \Elastic- Tree: saving energy in data center networks." Proc. 7th USENIX Conf. on Networked Systems Design and Implementation, pp. 17{17, 2011.

[17] J. L. Hennessy and D. A. Patterson. Computer Archi- tecture; A Quantitative Approach, 5th Edition. Morgan Kaufmann, 2012.13 2168-7161 (c) 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.Seehttp://www.ieee.org/publications_standards/publications/rights/index.html for more information.