

Energy-Aware Load Balancing and Application Scaling For the Cloud Ecosystem

Mr. Rakesh Prakash Kumawat¹, Prof. Ms. Megha Singh²

¹M. Tech Student, ²Asst. Professor & HOD, Dept. of Computer Sci. and Engineering
Central India Institute of Technology, Indore, MP, India

Abstract: *To introduce an energy-aware operation model used for load balancing and application scaling on a cloud. The basic philosophy of our approach is defining an energy-optimal operation regime and attempting to maximize the number of servers operating in this regime. Idle and lightly-loaded servers are switched to one of the sleep states to save energy. The load balancing and scaling algorithms also exploit some of the most desirable features of server consolidation mechanisms.*

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

I. INTRODUCTION

A. Problem Statement

The Problem as Follows,

Server gets overloaded due to excess of request from different host then it goes into the sleep mode. So by using server consolidation properties transfer the request from sleep mode to running mode on other server.

Use Auto scaling and peak energy level

II. LITERATURE SURVEY

A. Energy-Aware resource allocation heuristics for efficient management of data centers for Cloud computing

Anton Beloglazov a., Jemal Abawajyb, Rajkumar Buyya. Cloud computing offers utility-oriented IT services to users worldwide. Based on a pay-as-you-go model, it enables hosting of pervasive applications from consumer, scientific, and business domains. However, data centers hosting Cloud applications consume huge amounts of electrical energy, contributing to high operational costs and carbon footprints to the environment. Therefore, we need Green Cloud computing solutions that can not only minimize operational costs but also reduce the environmental impact. In this paper, we define an architectural framework and principles for energy-efficient Cloud computing. Based on this architecture, we present our vision, open research challenges, and resource provisioning and allocation algorithms for energy-efficient management of Cloud computing environments.

B. Towards Energy-Proportional Computing for Enterprise-Class Server Workloads

Wu-chun Feng. Massive data centers housing thousands of computing nodes have become commonplace in enterprise computing, and the power consumption of such data centers is growing at an unprecedented rate. Adding to the problem is the inability of the servers to exhibit energy proportionality, i.e., provide energy-efficient execution under all levels of utilization, which diminishes the overall energy efficiency of the data center. It is imperative that we realize effective strategies to control the power consumption of the server and improve the energy efficiency of data centers. With the advent of Intel Sandy Bridge processors, we have the ability to specify a limit on power consumption during runtime, which creates opportunities to design new power-management techniques for enterprise workloads and make the systems that they run on more energy-proportional.

C. Energy-Aware Autonomic Resource Allocation in Multitier Virtualized Environments

Danilo Ardagna, Barbara Panicucci, Marco Trubian, and Li Zhang

With the increase of energy consumption associated with IT infrastructures, energy management is becoming a priority in the design and operation of complex service-based systems. At the same time, service providers need to comply with Service Level Agreement (SLA) contracts which determine the revenues and penalties on the basis of the achieved performance level. This paper focuses on the resource allocation problem in multitier virtualized systems with the goal of maximizing the SLAs revenue while minimizing energy costs. The main novelty of our approach is to address—in a unifying framework—service centers resource management by exploiting as actuation mechanisms allocation of virtual machines (VMs) to servers, load balancing, capacity allocation, server

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

power state tuning, and dynamic voltage/frequency scaling. Resource management is modeled as an NP-hard mixed integer nonlinear programming problem, and solved by a local search procedure. To validate its effectiveness, the proposed model is compared to top-performing state-of-the-art techniques. The evaluation is based on simulation and on real experiments performed in a prototype environment. Synthetic as well as realistic workloads and a number of different scenarios of interest are considered. Results show that we are able to yield significant revenue gains for the provider when compared to alternative methods (up to 45 percent). Moreover, solutions are robust to service time and workload variations.

D. Green Cloud Computing: Balancing Energy in Processing, Storage, and Transport

By Jayant Baliga, Robert W. A. Ayre, Kerry Hinton, and Rodney S. Tucker

Network-based cloud computing is rapidly expanding as an alternative to conventional office-based computing. As cloud computing becomes more widespread, the energy consumption of the network and computing resources that underpin the cloud will grow. This is happening at a time when there is increasing attention being paid to the need to manage energy consumption across the entire information and communications technology (ICT) sector. While data center energy use has received much attention recently, there has been less attention paid to the energy consumption of the transmission and switching networks that are key to connecting users to the cloud. In this paper, we present an analysis of energy consumption in cloud computing. The analysis considers both public and private clouds, and includes energy consumption in switching and transmission as well as data processing and data storage. We show that energy consumption in transport and switching can be a significant percentage of total energy consumption in cloud computing. Cloud computing can enable more energy-efficient use of computing power, especially when the computing tasks are of low intensity or infrequent. However, under some circumstances cloud computing can consume more energy than conventional computing where each user performs all computing on their own personal computer (PC).

E. The Use of Ubiquitous Computing for Business Process Improvement

Alaaeddine Yousfi, Adrian de Freitas, Anind K. Dey and Rajaa Saidi

Due to the cut throat competition among organizations, business process improvement is now an everyday activity. A relentless activity that makes business processes more complex than ever. As they get more complex, the improvement rounds become time-consuming, costly and the quality of each outcome is put into jeopardy, which is somehow paradoxical with the concept of improvement. In this paper, we propose a business process improvement technique based on ubiquitous computing. First, we couple business processes with ubiquitous computing and define a ubiquitous business process. Then, we explain how ubiquitous computing positively impacts the performance metrics of business processes. Afterwards, we set a specification for designing ubiquitous business processes by extending BPMN. Finally, we propose a concrete case study about time-banking to corroborate our theory. A comparative study of the same process, in ubiquitous and non-ubiquitous versions, is established. The results clearly illustrate that ubiquitous computing impacts positively the business process performance metrics. Still, the case study corroborates that ubiquitous computing not only improves a business process but also enables it to get improved with the least of human interventions.

III. 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.

A. 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

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

correlate obtained data to the internal resource management strategies implemented by the system provider. Simulation does not allow conducting comprehensive analyses of the system performance due to the great number of parameters that have to be investigated.

IV. PROPOSED SYSTEM

There are three primary contributions of this paper: 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.

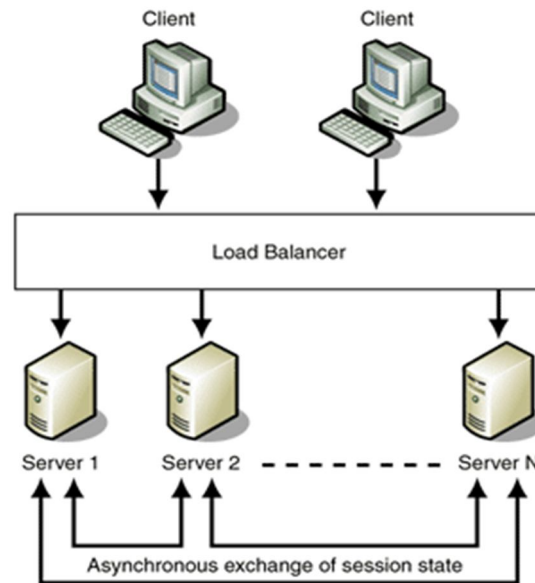
The objective of the algorithms is to ensure that the largest possible number of active servers operate within the boundaries of their respective optimal operating regime. The actions implementing this policy are: (a) migrate VMs from a server operating in the undesirable-low regime and then switch the 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.

A. Advantages Of Proposed System

After load balancing, the number of servers in the optimal regime increases from 0 to about 60% and a fair number of servers are switched to the sleep state.

There is a balance between computational efficiency and SLA violations; the algorithm can be tuned to maximize computational efficiency or to minimize SLA violations according to the type of workload and the system management policies.

B. System architecture



C. Mathematical Module

Input :

$U(Z) = \{ u_1, u_2, u_3, \dots, u_n \}$ user

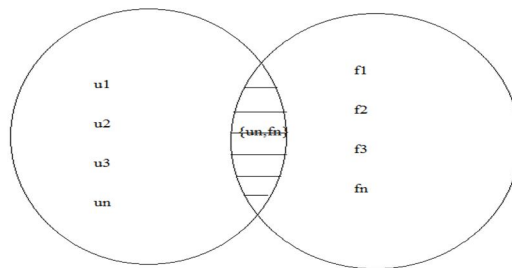
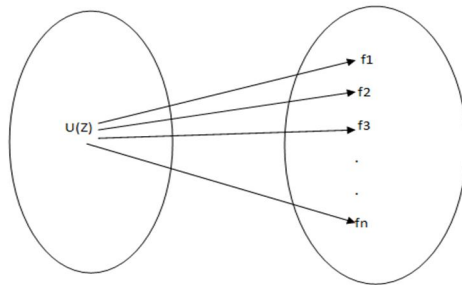
$F(Z) = \{ f_1, f_2, f_3, \dots, f_n \}$ file

$S(Z) = \{ s_1, s_2, s_3 \}$ server

$D(Z) = \{ d_1, d_2, d_3, \dots, d_n \}$ data

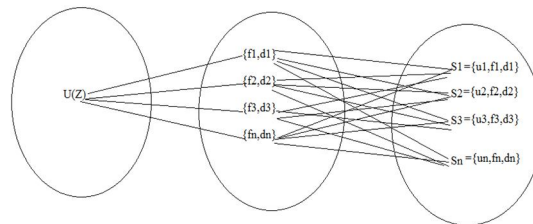
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

$M(Z) = \{m1, m2, m3, \dots, mn\} \dots \dots \dots$ migration
 $U(Z) \rightarrow F(Z) : \{(un, fn)\} \rightarrow$ uploading



$$U(Z) \cup F(Z)$$

$$U(Z) \cup F(Z) \cup D(Z) \cup S(Z) :$$



D. Algorithm

1. Input

Workload (W) $\rightarrow \{w1, w2, w3, \dots\}$

Resource (R) $\rightarrow \{R1, R2, R3, \dots\}$

Output

Migration List (M) $\rightarrow \{m1, m2, m3, \dots\}$

E. Energy efficient algorithm

1. START

2. Extract Total workload list

$W(Z) \rightarrow \{w1, w2, w3, \dots, wn\}$

3. Access total Resource list

$R(Z) \rightarrow \{R1, R2, R3, \dots, Rn\}$

4. User upload file

$U(Z):F(Z) \rightarrow \{(un, fn)\}$

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

5. check first cloud server work load
6. limitation of server depends on energy level
7. if server is go to energy threshold
8. file migrate to another server->HOT SPOT Process.
9. check remaining server workload.
10. find Min(workload Resources)->optimization
11. Manage workload of every server->Green Computing.
12. check energy level
13. end

V. OBJECTIVE

To consider the resources like requests on network, memory of requested data and CPU utilization of server and by checking and analyzing such a resources, we can balance the load by migrating or switching the workload to the different server.

VI. CONCLUSION

Finally Conclude that By considering the resources like requests on network, memory of requested data and CPU utilization of server and by checking and analyzing such a resources, we can balance the load by migrating or switching the workload to the different server and get the optimal efficiency for improving the throughput and minimize the cost of the system using the Green computing that uses the auto scaling by the server.

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 energy proportional computing." IEEE Computer, 40(12):33{ 37, 2007.
- [4] L. A. Barosso, 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 efficient resource management in virtualized cloud data centers." Proceedings of the 2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Comp., 2010.