Volume 6, Issue 12, December 2016 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Efficient Load Balancing Scheme in Cloud Using Resource Allocation Algorithm

B. Vijaya Bhaskar Reddy
Associate Professor, Department of CSE,
Shri Shirdi Sai Institute of Science & Engineering,
Anantapur, Andhra Pradesh, India

B. Bala Madan Mohan Reddy
M.tech, Department of CSE,
Shri Shirdi Sai Institute of Science & Engineering,
Anantapur, Andhra Pradesh, India

Abstract: Our system is based on an economic approach to managing shared server resources, in which services bid for resources as a function of delivered performance. The design and implementation of resource management in a hosting center operating system, with an emphasis on energy as a driving resource management issue for large server clusters. A cloud application in response to a request consistent with the SLA. Scaling is the process of allocating additional resources to a request consistent with the SLA. So this project proposing, an energy-aware operation model used for load balancing and application scales on a cloud. Our approach is designing an energy-optimal operation regime and attempting to maximize the number of servers operating in this regime. The goals are to provision server resources for services in a way that automatically adapts to offered load, improve the energy efficiency of server by dynamically resizing the active server set, and respond to power supply disruptions or thermal events by degrading service.

Keywords: CSP, FCFS, SLA

I. INTRODUCTION

The concept of “load balancing” dates back to the time the first distributed computing systems were implemented in the late 1970s and early 1980s. 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 one or more systems in the distributed environment. Distributed systems became popular after communication networks allowed multiple computing engines to effectively communicate with one another and the networking software became an integral component of an operating system. Once processes were able to easily communicate with one another using sockets 1 , the client-server paradigm became the preferred method to develop distributed applications; it enforces modularity, provides a complete isolation of clients from the servers, and enables the development of stateless servers.

The client-server model proved to be not only enduring, but also increasingly successful; three decades later, it is at the heart of utility computing. In the last few years packaging computing cycles and storage and offering 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) offer computing and storage services based on three different delivery models SaaS (Software as a Service), PaaS (Platform as a Service), and IaaS (Infrastructure as a Service). Reduction of energy consumption thus, of the carbon footprint of cloud related activities, is increasingly more important for the society.

Indeed, as more and more applications run on clouds, more energy is required to support cloud computing than the energy required for many other humanrelated activities. While most of the energy used by data centers is directly related to cloud computing, a significant fraction is also used by the networking infrastructure used to access the cloud. This fraction is increasing, as wireless access becomes more popular and wireless communication is energy intensive. In this paper we are only concerned with a single aspect of energy optimization, minimizing the energy used by cloud servers. Unfortunately, computer and communication systems are not energy proportional systems, in other words, their energy consumption does not scale linearly with the workload; an idle system consumes a rather significant fraction, often as much as 50%, of the energy used to deliver peak performance.

Cloud elasticity, one of the main attractions for cloud users, comes at a stiff price as the cloud resource management is based on over-provisioning. This means that a cloud service provider has to invest in a larger infrastructure than a “typical” or average cloud load warrants. At the same time, cloud elasticity implies that most of the time cloud servers operate with a low load, but still use a large fraction of the energy necessary to deliver peak performance. The low average cloud server utilization affects negatively the common measure of energy efficiency, the performance per Watt of power and amplifies the ecological impact of cloud computing.

The strategy for resource management in a computing cloud we discuss is to concentrate the load on a subset of servers and, whenever possible, switch the rest of the servers to a sleep state. In a sleep state the energy consumption is very low. This observation implies that the traditional concept of load balancing could be reformulated to optimize the energy consumption of a large-scale system as follows: distribute evenly the workload to the smallest set of servers operating at an optimal energy level, while observing QoS constraints, such as the response time. An optimal energy level is one when the normalized system performance, defined as the ratio of the current performance to the maximum performance, is delivered with the minimum normalized energy consumption, defined as the ratio of the current energy consumption to the maximal one.

II. RELATED WORK

Shunmei Meng, Wanchun Dou, Xuyun Zhang, Jinjun Chen[1] Live Migration of Virtual Machines:- Migrating operating system instances across distinct physical hosts is a useful tool for administrators of data centers and clusters: It allows a clean separation between hardware and software, and facilitates fault management, load balancing, and low-level system maintenance. By carrying out the majority of migration while OSes continue to run, we achieve impressive performance with minimal service downtimes; we demonstrate the migration of entire OS instances on a commodity cluster, recording service downtimes as low as 60ms. We show that that our performance is sufficient to make live migration a practical tool even for servers running interactive loads.

A. Beloglazov and R. Buyya. [4]. The most effective way to improve the resources utilization and energy efficiency in cloud data centers is dynamic consolidation of virtual machines so it can directly affect the resource utilization and quality of service (Qos) when determine the reallocation of VM‟s from overload . The Qos is influenced because of the server get overloaded that causes resource shortage and performance degradation problem of applications .The heuristic based solutions of this problem of detection overloaded host. This paper gives a novel approach that can solve the host overload detection problem that can maximize the mean time of migration using the Markov chain model the multi size sliding window workloads estimation technique use to handle the workload.

V. Gupta and M. Harchol-Balter [9]. In this paper author can takes admission control problem in resource sharing problem in resource sharing system i.e. transaction processing system and web servers. Authors can abstract the Processor sharing (PS) server with adequate server rate and First Come First Serve (FCFS) queue an d analyze the performance model. It also shows that by minimizing the mean response time the peak energy is not always optimal. They show that the dynamic policies are more robust for unknown traffic intensities.

H. N. Van, F. D. Tran, and J.-M. Menaud. [15]. The main aim for data centers in cloud computing is to improve the profit and minimizing the power consumption and maintains SLAs. In this paper, author can describes a framework foe resource management that combines a dynamic virtual machine placement manager and dynamic VM provisioning manager. It can take several experiments that how system can be controlled to make trade-offs between energy consumption and application performance.

S. V. Vrbsky, M. Lei, K. Smith, and J. Byrd. [16]. The energy cost of data centers are rapidly growing now a days, so we use serer consolidation for reduce the energy cost. In this paper, author analyze the workload of servers by observing potentials for power saving. It also investigates the low risk consolidation. From analysis two new methods are designed that can achieved the power saving.

III. SYSTEM ANALYSIS

A. Existing System

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. Idle and under-utilized servers contribute significantly to wasted energy.

B. Disadvantages

- Less feasibility
- Storage management is less
- High computational services

C. Proposed System

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. Scaling is the process of allocating additional resources to a cloud application in response to a request consistent with the SLA. In this project, proposing an energy-aware operation model used for load balancing and application scaling on a cloud. Our approach is designing an energy-optimal operation regime and attempting to maximize the number of servers operating in this regime. This mainly focus on

(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 profiles.

D. Advantages:

1. Good in Storage management
2. Energy consumption is less

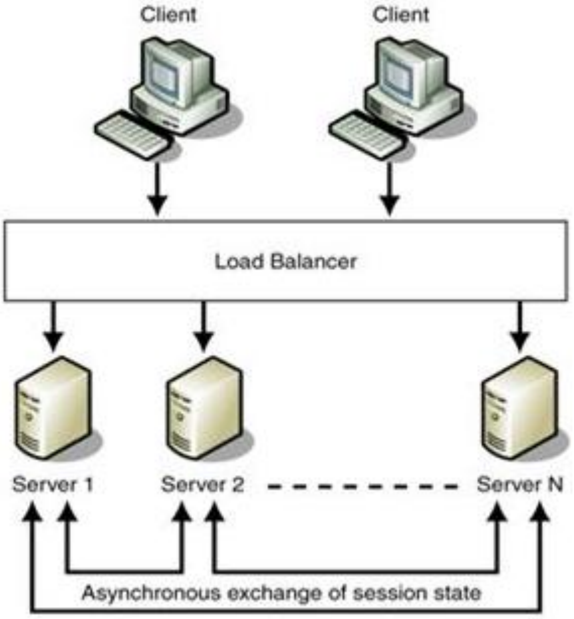
3. It is possible to evaluate energy performance after completing work.
4. Automatic scaling function for feasibility.

IV. SYSTEM DESIGN

Systems design is the process of defining the architecture, components, modules, interfaces, and data for a system to satisfy specified requirements. Systems design could be seen as the application of systems theory to product development.

A. System Architecture:

System architecture is the conceptual model that defines the structure, behaviour, and more views of a system. An architecture description is a formal description and representation of a system, organized in a way that supports reasoning about the structures and behaviours of the system.



V. CONCLUSION

In this review paper we did the study of existing load balancing and workload migration techniques. Previous existed system having problem such as larger energy consumption, more computational time. Low average server utilization and its impact on the environment make it imperative to devise new energy-aware policies. A quantitative evaluation of an optimization algorithm or an architectural enhancement is a rather intricate and time consuming process; several benchmarks and system configurations. In this paper some energy aware load balancing techniques are discussed. These techniques are aimed to allocate the resources to the VM requests for reducing the energy consumption. Each of these techniques has some merits and demerits. In future, we will try to design an technique that is able to overcome some of these demerits and that can improves the utilization of resources energy efficiently.

REFERENCE

[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 Datacenter 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., 2

[6] A. Beloglazov, J. Abawajy, R. Buyya. “Energy-aware resource allocation heuristics for efficient 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 survey results analysis.” 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-efficient protocol for cooperative networks.” IEEE/ACM Trans. on Networking, 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 effective 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, efficiency, and quality metrics.” Proc. Int. Conf. on Dependable Systems and Networks, pp. 326–335, 2008.

[13] Google. “Google’s green computing: efficiency at scale.” http://static.googleusercontent.com/external content/ untrusted dlcp/www.google.com/en/us/green/pdfs/google -green-computing.pdf (Accessed on August 29, 2013).

[14] V. Gupta and M. Harchol-Balter. “Self-adaptive admission control policies for resource-sharing systems.” Proc. 11th Int. Joint Conf. Measurement and Modeling Computer 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 computing.