

A REVIEW ON ENERGY EFFICIENT LOAD BALANCING TECHNIQUES FOR SECURE AND RELIBLE CLOUD ECOSYSTEM

A.S. THORAT¹, Prof. S.K.SONKAR²

¹ME Student, Department of Computer Engineering, Amrutvahini COE, Sangamner, Maharashtra, India

²Assistant Professor, Department of Computer Engineering, Amrutvahini COE, Sangamner, Maharashtra, India

ABSTRACT

Cloud computing is a growing paradigm and promising technology that leading towards improving business performance. The rapidly growing rate of the usage of big-scale data centers on cloud has demand for computational power. Cloud computing having datacenters that hosting cloud applications consumes large amount of energy. Therefore minimization of energy consumption and balance the resources are essential in cloud computing. Load balancing is the primary consideration in the cloud computing environment. Its main aim is to distribute workloads across various computing resources and optimize the usage of resources, increase efficiency. Load balancing provides use satisfaction and also the ratio of resource utilization after ensuring the allocation and efficiency of every resource being computed. This paper presents a survey of load balancing mechanism in order to provide the efficient and optimized utilization of resources and overall cost minimization.

Keyword: - Cloud computing, Load balancing, Consolidation, Energy-aware scheduling, Energy proportional systems

1. Introduction

Cloud computing is evaluation of technology and having lots of virtual resources that are useful and accessible and can be used as resources on demand basis with or without nominal charges. The energy-aware scheduling operation model used for application scaling and load balancing on a cloud application server and also that gives some of the most important feature of server consolidation mechanism. The concept of load balancing is come to existence when the first distributed computing systems were implemented. It should have the means exactly what the name distribute the workload to a set of servers to maximize the throughput, minimization of the response time, and increase the system flexibility to faults by avoiding overloading the systems. Load balancing key feature and is one of the central issues in cloud computing. It is a technique in which distribution of the dynamic local workload equally across the different clusters in cloud in order to avoid the situation where few nodes are overloaded while few are remains idle. Cloud computing is a client-server architecture composed by large and power-consuming data centers designed to support the efficient and scalable output required by consumers. The cloud computing having a on demand service, broad network access, resource pooling, rapid elasticity, measured service are the essential characteristics. The dynamic workload results some systems overload and some systems remains unused, therefore it is necessary to distribute this load efficiency for preventing such odd resource utilization. Hence the concept load balance comes into existence that can distribute load across different computer clusters, network links, disk driver and some other resources to improve the throughput and optimal resource utilization, avoid overloading and minimizing response time. Scaling is the method that allocates additional resources to a cloud application in response to a request consistent with the SLA. We can distinguish two scaling methods, i.e. Horizontal and Vertical scaling. Horizontal scaling is the most common method of scaling on a cloud system; it can increase the number of Virtual Machines (VMs) when the load of nodes increases and reduce the number when the load decreases.

2. Literature Survey

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.

J. Baliga, R.W.A. Ayre, K. Hinton, and R.S. Tucker. [2]. The availability of the high-speed internet network and IP connections is provide the delivery of the latest network based services. So the network based computing becomes more widespread and rapidly expanding the energy consumption of the network and the network resources are also rapidly grown .This is done when there is increasing more attention to manage energy consumption across the information and communication technology sectors. while energy uses by data centers having received much attention, but there has been less attention given to the energy consumption of the switching the networks and transmission for connecting users to the cloud. This paper describe the analysis of energy consumption in cloud computing that consider the public and private classes in that we can include the energy consumption in transmission and switching as well as data storage and data processing. This paper also describes the energy consumption is transport and switching having adequate percentage of total energy consumption in the cloud computing.

A. Beloglazov, R. Buyya [3]. This paper describes that in business, scientific and different –applications requires the large computation power so the rapidly increasing the demand of such a resources to the electrical power required by the large scale data centers also increases. This paper also defines for reducing operational costs and also provide quality of services (Qos) using energy efficient resource management system for virtualized data centres with consolidation of VMS are achieved the energy saving according to resource utilization using the live migration results are presented of simulation driven evaluation for VMS dynamic reallocation according to CPU performance equipment . This can results the substantial energy saving and also ensures the reliable Qos.

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.

M. Elhawary and Z. J. Haas. [5]. In this paper author model a cooperative network that having transmission link in the wireless network having transmitter cluster and receiver cluster. Author also propose a protocol for cooperative communication for transmission of data in cooperative network and analyze the robustness of the protocol along with the energy consumption and error rate trade off that shows the result that error rate is reduced and energy saving so lifetime increase of cooperative sensor network.

Gandhi, M. Harchol-Balter, R. Raghunathan, and M.Kozuch [6], In this paper author introduces a Auto scale method that can greatly reduce the number of servers that are needed in data centers for dynamic capacity management. Auto scale techniques can scale the capacity of data centers, added or removed server when needed. Auto Scale can maintain the capacity of the data centers to handle the burst in the request rate and make server efficiency and request size robust. The authors also demonstrate that auto scale technique rapidly improves over the existing dynamic capacity management policies with respect to robustness and meeting SLAs.

Gandhi, M. Harchol-Balter, R. Raghunathan, and M.Kozuch [7]. In this paper author can investigate that sleep state is effective or not in data centers for that they can consider the advantage of sleep states across orthogonal dimensions i.e. type of dynamic power management policy employed, size of data centers and variability in the workload tracing. The result after many traces it shows that sleep state greatly increases dynamic power management and it also suggested that sleep state is more beneficial for the larger data centers.

D. Gmach, J. Rolia, L. Cherkasova, G. Belrose, T. Tucicchi, and A. Kemper [8]. In this paper author can state that integrated approach to the resource pool management. It can use trace based approach to evaluation of which workload consolidated on which server that are used to determine optimal workload that provides the quality of services. The trace based technique repeatedly applies for improving efficiency and quality of service. By interaction of the workload placement controller and reactive controller they can observe the current behaviors of overloaded server for migration of workload and lightly loaded server. They are also studied trade-offs between the power usage and required capacity, resource access quality of service for memory resource and CPU, and the number of migration.

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 and 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.

B. Heller, S. Seetharaman, P. Mahadevan, Y. Yakoumis, P. Sharma, S. Banerjee, and N. McKeown [10]. In this paper author presents Elastic Tree, i.e. network power manager that can dynamically adjust the set of elements in active network switches and links that is for satisfy the traffic roads in data centers they are implement analyze the Elastic tree on a testbed prototype and examine the trade-off between performance, energy efficiency and robustness. The result shows that Elastic tree saves the half of network energy for data center workload. It also shows the configuration of Elastic tree and that can minimize network power bill.

E. Le Sueur and G. Heiser [11]. In this paper, Dynamic voltage and frequency scaling (DVFS) is a technique used for power management to reduce the supply voltage. We have to decrease the clock frequency of the processor as a result the power consumption also reduces that is used for memory-bound workload. They also analyze the clock frequency of processor, power consumption, and smaller dynamic power range by examining the DVFS potentials with the AMD processors. They founded that DVFS is effective over older platforms, that increases energy usage for high memory bound workload.

D. C. Marinescu, A. Paya, and J.P. Morrison [12]. For hierarchically organization of cloud, the author can propose a two stage protocol for resource management. In first stage it can state the locality of coalitions of supply agents formation, and in second stage the proxy base and clock base algorithm are used for combinatorial action. Price discovery for clock phase and multiple rounds for auction are conducted by proxy. These two protocols are used for balance between cloud client and service providers for low cost service and decent profit respectively.

Paya and D. C. Marinescu [13]. The energy consumption of the system have workload scalability problem. The lightly loaded server also requires the more energy so author can propose the concept load balancing to optimize the energy consumption for large scale system that can be distribute the workload among different set of servers that can observe the response time and operate on optimal energy level.

B. Uргаonkar and C. Chandra. [14]. In this paper, author can proves novel dynamic capacity technique for the multitier internet application that employs the flexible queuing model is used for determining that how much resources allocated to the each tier and predictive and reactive methods combination that used to determine when to provision these resources the experiments demonstrate the techniques for having dynamic workload. This technique doubles the application capacity in five minutes that maintains the response time.

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 for 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 server 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.

3. CONCLUSIONS

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 a technique that is able to overcome some of these demerits and that can improve the utilization of resources energy efficiently.

5. ACKNOWLEDGEMENT

I am very much thankful of Prof. S.K. Sonkar for their guidance and consistent encouragement in this paper work.

6. REFERENCES

- [1]. Shunmei Meng, Wanchun Dou, Xuyun Zhang, Jinjun Chen, "KASR: A Keyword Aware Service Recommendation Method on MapReduce for Big Data Applications," IEEE Transactions On Parallel and distributed system, TPDS-12-1141-2013.
- [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]. 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.
- [4]. 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.
- [5]. M. Elhawary and Z. J. Haas. "Energy-efficient protocol for cooperative networks." IEEE/ACM Trans. on Networking, 19(2):561-574, 2011.
- [6]. 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.
- [7]. 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.
- [8]. 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.
- [9]. 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.
- [10]. B. Heller, S. Seetharaman, P. Mahadevan, Y. Yakoumis, 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.
- [11]. E. Le Sueur and G. Heiser. "Dynamic voltage and frequency scaling: the laws of diminishing returns." Proc. Workshop on Power Aware Computing and Systems, HotPower'10, pp. 2-5, 2010.
- [12]. D. C. Marinescu, A. Paya, and J.P. Morrison. "Coalition formation and combinatorial auctions; applications to self-organization and self-management in utility computing." <http://arxiv.org/pdf/1406.7487v1.pdf>, 2014.
- [13]. A. Paya and D. C. Marinescu. "Energy-aware load balancing policies for the cloud ecosystem." <http://arxiv.org/pdf/1307.3306v1.pdf>, December 2013.
- [14]. B. Urgaonkar and C. Chandra. "Dynamic provisioning of multi-tier Internet applications." Proc. 2nd Int. Conf. on Automatic Comp., pp. 217-228, 2005.
- [15]. H. N. Van, F. D. Tran, and J.-M. Menaud. "Performance and power management for cloud infrastructures." Proc. IEEE 3rd Int. Conf. on Cloud Comp., pp. 329-336, 2010.
- [16]. S. V. Vrbsky, M. Lei, K. Smith, and J. Byrd. "Data replication and power consumption in data grids." Proc IEEE 2nd Int. Conf. on Cloud Computing Technology and Science, pp. 288-295, 2010.

BIOGRAPHIES

	<p>Mr. A. S. Thorat is Pursuing Master in Engineering from Amrutvahini College of Engineering Sangamner. Received BE degree from University of Pune. His interested Areas are Cloud Computing, Data mining, Computer Network.</p>
	<p>Prof. S. K. Sonkar is Assistant Professor in Amrutvahini College of Engineering, Sangamner. He is PhD Pursuing from university of Pune. His Research interests includes Cloud Computing, Network security, Data Mining.</p>

