A Survey Paper on Load balancing Model Based on Cloud Partition Techniques for Cloud Environment

Akshada Deshmukh[1], Prof.R.L.Paikrao[2]

[1] *ME II Year, Computer Department, AVCOE, Sangamner, India*
[2] *Head, Computer Department, AVCOE, Sangamner, India*

ABSTRACT

Load balancing in the cloud computing environment has an important impact on the performance. Good load balancing makes cloud computing more efficient and improves user satisfaction. This article introduces a better load balance model for the public cloud based on the cloud partitioning concept with a switch mechanism to choose different strategies for different situations. The algorithm applies the game theory to the load balancing strategy to improve the efficiency in the public cloud environment.

Keyword -*load balancing model, public cloud, cloud partition, game theory.*

INTRODUCTION

The load balancing model given in this article is aimed at the public cloud which has numerous nodes with distributed computing resources in many different geographic locations. Thus, this model divides the public cloud into several cloud partitions. When the environment is profoundly and astronomically immense and intricate, these divisions simplify the load balancing. The cloud has a main controller that culls the congruous partitions for arriving jobs while the balancer for each cloud partition culls the best load balancing strategy.

Cloud computing is a magnetizing technology in the field of computer science. In Gartner's report, it verbally expresses that the cloud will bring changes to the IT industry. The cloud is transmuting our life by providing users with incipient types of accommodations. Users get accommodation from a cloud without fixating on the details. NIST gave a definition of cloud computing as a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and accommodations) that can be rapidly provisioned and relinquished with minimal management effort or accommodation provider interaction. More and more people fixate on cloud computing. Cloud computing is efficient and scalable but maintaining the stability of processing so many jobs in the cloud computing environment is a very involute quandary with load balancing receiving much attention for researchers.

Since the job advent pattern is not prognostic able and the capacities of each node in the cloud differ, for load balancing quandary, workload control is crucial to amend system performance and maintain stability. Load balancing schemes depending on whether the system dynamics are consequential can be both static and dynamic. Static schemes do not utilize the system information and are less involute while dynamic schemes will bring supplemental costs for the system but can transmute as the system status changes. A dynamic scheme is utilized here for its flexibility. The model has a main controller and balancers to accumulate and analyze the information. Thus, the dynamic control has little influence on the other working nodes. The system status then provides a substructure for culling the right load balancing strategy.

LITERATURE SURVEY

- **Energy-aware Load Balancing and Application Scaling for the Cloud Ecosystem :-**

Ashkan Paya and Dan C. Marinescu, in this paper they 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 discussed in the literature.

- **Cloud computing: Distributed internet computing for IT and scientific research :-**

M. D. Dikaiakos, D. Katsaros, P. Mehra, G. Pallis,and A. Vakali, One vision of 21st century computing is that users will access Internet services over lightweight -portable devices rather than through some descendant of the traditional desktop PC. Because users won't have (or be interested in) powerful machines, who will supply the computing power? The answer to this question lies with cloud computing. Cloud computing is a recent trend in IT that moves computing and data away from desktop and portable PCs into large data centers. It refers to applications delivered as services over the Internet as well as to the actual cloud infrastructure — namely, the hardware and systems software in data centers that provide these services. The key driving forces behind cloud computing are the ubiquity of broadband and wireless networking, falling storage costs, and progressive improvements in Internet computing software. Cloud-service clients will be able to add more capacity at peak demand, reduce costs, experiment with new services, and remove unneeded capacity, whereas service providers will increase utilization via multiplexing, and allow for larger investments in software and hardware.

- **The NIST definition of cloud computing :-**

P. Mell and T. Grance, Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model promotes availability and is composed of five essential characteristics, three service models, and four deployment models.

- **Load distributing for locally distributed systems :-**

Deepak Gupta and Pradip Bepari, A distributed system consists of, possibly heterogeneous, computing nodes connected by a communication network. Such a system can be used effectively only if the software presents a single system image of this physically distributed system to its users. Thus all resources of any node should be easily and transparently accessible from any other node. One of the most important of such resources is the CPU. The CPUs of all the nodes in the system can be made transparently available to all nodes if the nodes share their computing load. Thus the system should decide the best node to execute any job regardless of where the job originated and may even migrate some jobs during their execution. This calls for transparent process migration among the nodes. In this paper, we brie y discuss the issues in such load sharing. There are two orthogonal kinds of issues to be addressed here. The first kind relate to the policies for migration. For example, where a new job should be executed, when and where an executing process should be migrated etc. The second kind of issues relate to the mechanisms for migration. For example, how to checkpoint and transfer the state of a running process, can a process be migrated to a machine with a different architecture, can old programs be migrated without modifications etc. We first brief survey the current state of the art in this area. Then, we talk about the work being done at IIT Kanpur in this area.

- **Load balancing in the cloud: Tools, tips and techniques :-**

B. Adler, Load balancing is a method to distribute workload across one or more servers, network interfaces, hard drives, or other computing resources. Typical datacenter implementations rely on large, powerful (and expensive) computing hardware and network infrastructure, which are subject to the usual risks associated with any physical device, including hardware failure, power and/or network interruptions, and resource limitations in times of high.

- **Availability and load balancing in cloud computing :-**

Z. Chaczko, V. Mahadevan, S. Aslanzadeh, and C. Mcdermid,"Cloud computing" is a term, which involves virtualization, distributed computing, networking, software and web services. A cloud consists of several elements

such as clients, datacenter and distributed servers. It includes fault tolerance, high availability, scalability, flexibility, reduced overhead for users, reduced cost of ownership, on demand services etc. Central to these issues lies the establishment of an effective load balancing algorithm. The load can be CPU load, memory capacity, delay or network load. Load balancing is the process of distributing the load among various nodes of a distributed system to improve both resource utilization and job response time while also avoiding a situation where some of the nodes are heavily loaded while other nodes are idle or doing very little work. Load balancing ensures that all the processor in the system or every node in the network does approximately the equal amount of work at any instant of time. This technique can be sender initiated, receiver initiated or symmetric type (combination of sender initiated and receiver initiated types). Our objective is to develop an effective load balancing algorithm using Divisible load scheduling theorem to maximize or minimize different performance parameters (throughput, latency for example) for the clouds of different sizes (virtual topology depending on the application requirement).

- **Load balancing of nodes in cloud using ant colony optimization :-**

K. Nishant, P. Sharma, V. Krishna, C. Gupta, K. P. Singh, N. Nitin, and R. Rastogi, In this paper, we proposed an algorithm for load distribution of workloads among nodes of a cloud by the use of Ant Colony Optimization (ACO). This is a modified approach of ant colony optimization that has been applied from the perspective of cloud or grid network systems with the main aim of load balancing of nodes. This modified algorithm has an edge over the original approach in which each ant build their own individual result set and it is later on built into a complete solution. However, in our approach the ants continuously update a single result set rather than updating their own result set. Further, as we know that a cloud is the collection of many nodes, which can support various types of application that is used by the clients on a basis of pay per use. Therefore, the system, which is incurring a cost for the user should function smoothly and should have algorithms that can continue the proper system functioning even at peak usage hours.

- **A comparative study into distributed load balancing algorithms for cloud computing :-**

M. Randles, D. Lamb, and A. Taleb-Bendiab, The anticipated uptake of Cloud computing, built on well-established research in Web Services, networks, utility computing, distributed computing and virtualization, will bring many advantages in cost, flexibility and availability for service users. These benefits are expected to further drive the demand for Cloud services, increasing both the Cloud's customer base and the scale of Cloud installations. This has implications for many technical issues in Service Oriented Architectures and Internet of Services (IoS)-type applications; including fault tolerance, high availability and scalability. Central to these issues is the establishment of effective load balancing techniques. It is clear the scale and complexity of these systems makes centralized assignment of jobs to specific servers infeasible; requiring an effective distributed solution. This paper investigates three possible distributed solutions proposed for load balancing; approaches inspired by Honeybee Foraging Behavior, Biased Random Sampling and Active Clustering.

EXISTING SYSTEM

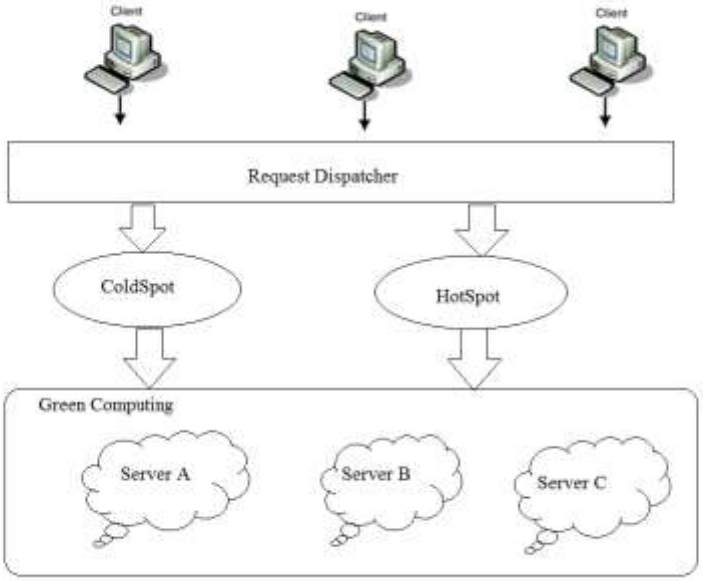
- Applications running in today’s data centers show high workload variability. While seasonal patterns, trends and expected events may help building proactive resource allocation policies, this approach has to be complemented with adaptive strategies which should address unexpected events such as flash crowds and volume spikes.
- Additionally, the limitations of current I/O infrastructures in the face of dramatic increase of data generation require, the ability to build novel abstractions and models for robust decision making regarding data layout and data locality.
- In single storage cloud system each cloud customer's data is stored on single higher con figuration server. Even if that server has huge amount resources such as RAM, Hard disk, processing power, it has certain limit. If it crosses that limit then particular resource performance slows down. The capacities of each node in the cloud differ, for load balancing problem, workload control is crucial to improve system performance and maintain stability. Load balancing schemes depending on whether the system dynamics are important can be either static or dynamic. Static schemes do not use the system information and are less complex

while dynamic schemes will bring additional costs for the system but can change as the system status changes.

DISADVANTAGES

- Since the main controller deals with information for each partition, smaller data set will lead to the higher processing rates. The balancers in each partition gather the status information from every node and then choose the right strategy to distribute the jobs.
- Since there is more than one index that can be utilized, the selection usually differ from one algorithm to another.

PROPOSED SYSTEM DESIGN



- The load balancing model is aimed at the public cloud which has numerous nodes with distributed computing resources in many different geographic locations. Thus, this model divides the public cloud into several cloud partitions. When the environment is very large and complex, these divisions simplify the load balancing. The cloud has a main controller that chooses the suitable partitions for arriving jobs while the balancer for each cloud partition chooses the best load balancing strategy. There are several cloud computing categories with this work focused on a public cloud.
- A public cloud is based on the standard cloud computing model, with service provided by a service provider. A large public cloud will include many nodes and the nodes.
  In different geographical locations. Cloud partitioning is used to manage this large cloud.
  A cloud partition is a subarea of the public cloud with divisions based on the geographic
  Locations.
- The load balancing strategy is based on the cloud partitioning concept. After creating the cloud partitions, the load balancing then starts, when a job arrives at the system, with the main controller deciding which cloud partition should receive the job. The partition load balancer then decides how to assign the jobs to the nodes. When the load status of a cloud partition is normal, this partitioning can be accomplished locally.

CONCLUSION

All of the existing System having how to store record as well as some important think related load balancing System while proposed system contains more advanced scenario which is not contain in existing system.

REFERENCES

[1] R. Hunter, The why of cloud, http://www.gartner.com/DisplayDocument?doc cd=226469&ref=gnoreg,2012.
[2] M. D. Dikaiakos, D. Katsaros, P. Mehra, G. Pallis, and A. Vakali, Cloud computing: Distributed internet computing for IT and scientific research, *Internet Computing*, vol.13, no.5, pp.10-13,Sept.-Oct.2009.
[3] P. Mell and T. Grance, The NIST definition of cloud computing, http://csrc.nist.gov/ publications/nistpubs/800-145/SP800-145.pdf,2012.
[4] Microsoft Academic Research, Cloud computing, http://libra.msra.cn/Keyword/6051/cloud-computing?query=cloud%20computing,2012.
[5]GooglTrends,Cloudcomputing, http://www.google.com/trends/explore#q=cloud%20computing,2012.
[6] N. G. Shivaratri, P. Krueger, and M. Singhal, Load distributing for locally distributed systems,*Computer*,vol.25,no.12,pp.33-44,Dec.1992.
[7] B. Adler, Load balancing in the cloud: Tools, tips and techniques, http://www.rightscale. com/info center/whitepapers/Load-Balancing-in-the-Cloud.pdf, 2012.
[8] Z. Chaczko, V. Mahadevan, S. Aslanzadeh, and C. Mcdermid, Availability and load balancing in cloud computing, presented at the 2011 International Conference on Computer and SoftwareModeling,Singapore,2011.
[9] K. Nishant, P. Sharma, V. Krishna, C. Gupta, K. P. Singh, N. Nitin, and R. Rastogi, Load balancing of nodes in cloud using ant colony optimization, in *Proc. 14th International Conference on Computer Modelling and Simulation (UKSim)*, Cambridgeshire, United Kingdom, Mar. 2012, pp. 28-30.
[10] M. Randles, D. Lamb, and A. Taleb-Bendiab, A comparative study into distributed load balancing algorithms for cloud computing, in *Proc. IEEE 24th International Conference on Advanced Information Networking and Applications*, Perth, Australia, 2010, pp. 551-556.