International Journal of Engineering Innovation & Research
Volume 3, Issue 3, ISSN: 2277 – 5668

A Package Complementary Load Balancing Model Based on Cloud Partitioning for the Public Cloud

Ashwini Patil
Student 4th Sem M.Tech.
Computer Science and Engineering
M. S. Engineering College, Bangalore, India
ashwini21patil@gmail.com

Aruna M. G.
Associate Professor
Dept. of Computer Science and Engineering
M. S. Engineering College, Bangalore, 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.**

In cloud computing, load balancing is required to distribute the dynamic local workload evenly across all the nodes. It helps to achieve a high user satisfaction and resource utilization ratio by ensuring an efficient and fair allocation of every computing g resource. Proper load balancing aids in minimizing resource consumption, implementing fail-over, enabling scalability, avoiding bottlenecks and over-provisioning etc.

The Amazon EC2 is used to launch the instance and deploy our project.

***Keywords* – AWS, Cloud Computing, EC2, Game Theory, Load Balancing, Round Robin.**

I. 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 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.

Cloud computing is an attracting technology in the field of computer science. In Gartner's report [1], it says that the cloud will bring changes to the IT industry. The cloud is changing our life by providing users with new types of services. Users get service from a cloud without paying attention to the details [2]. 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 services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. More and more people pay attention to 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 complex problem with load balancing receiving much attention for researchers. Since the job arrival pattern is not predictable and the capacities of each node in the cloud differ, for load balancing problem, workload control is crucial to improve system performance and maintain stability.

II. RELATED WORK

One vision of 21st century computing is that users will access Internet services over lightweight portable devices (PDA's, Tablets) 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. As of now, computer networks are still in their infancy, but as they grow up and become sophisticated, we will probably see the spread of 'computer utilities' which, like present electric and telephone utilities, will service individual homes and offices across the country. This vision of the computing utility based on the service provisioning model anticipates the massive transformation of the entire computing industry in the 21st century whereby computing services will be readily available on demand, like other utility services available in today's society. Similarly, computing service users (consumers) need to pay providers only when they access computing services.

In addition, consumers no longer need to invest heavily or encounter difficulties in building and maintaining complex IT infrastructure. Over the years, new computing paradigms have been proposed and adopted, with the emergence of technological advances such as multi-core processors and networked computing environments, to edge closer toward achieving this grand vision. These new computing paradigms include cluster computing, Grid computing, P2P computing, service computing, market-oriented computing, and most recently 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 is 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 3demand, 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.

Papers Cited:

S. Penmatsa and A. T. Chronopoulos, Game-theoretic static load balancing for distributed systems, Journal of Parallel and Distributed Computing- In this paper they solved a static load balancing problem for single class and multiclass jobs in a distributed system .They used Cooperative game to model the load balancing problem and their solution was based on Nash Equilibrium Bargaining which provides a Pareto optimal solution for the distributed system and is also a fair solution. The objective of their approach was to provide fairness to all the jobs (in a single-class system) and the users of the jobs (in a multi-user system). To provide fairness to all the jobs in the system, they used a cooperative game to model the load balancing problem. Their solution was based on the Nash Bargaining Solution (NBS) which provides a Pareto optimal solution for the distributed system and is also a fair solution.

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 this paper they, solved the working efficiency of the cloud as some nodes which are overloaded will have a higher task completion time compared to the corresponding time taken on an under loaded node in the same cloud. They used ACO for load balancing. Their approach aims at efficiently distribution of the load among the nodes and such that the ants never encounter a dead end for movements to nodes for building an optimum solution set.

S. Aote and M. U. Kharat, A game-theoretic model for dynamic load balancing in distributed systems- In this paper they solved a load balancing problems in distributed systems like:1) Global approach 2)Cooperative approach 3)Cooperative approach.

They used non- cooperative load balancing game, and considered the structure of the Nash equilibrium. Based on this structure they derived a new distributed load balancing algorithm. Their main focus was to define the load balancing problem and the scheme to overcome it, by using new area called game theory.

Load Balancing Techniques in Cloud Computing: Systematic Re-View- In this paper they solved the load balancing problem to distribute the dynamic local workload evenly across all the nodes. It helps to achieve a high user satisfaction and resource utilization ratio by ensuring an efficient and fair allocation of every computing resource. Proper load balancing aids in minimizing resource consumption, implementing fail-over, enabling scalability, avoiding bottlenecks and over-provisioning etc. In this paper, a systematic review of existing load balancing techniques is presented. Out of 3,494 papers analyzed, 15 papers are identified reporting on 17 load balancing techniques in cloud computing. Their study concludes that all the existing techniques mainly focus on reducing associated overhead, service response time and improving performance etc. Various parameters are also identified, and these are used to compare the existing techniques.

III. System Model

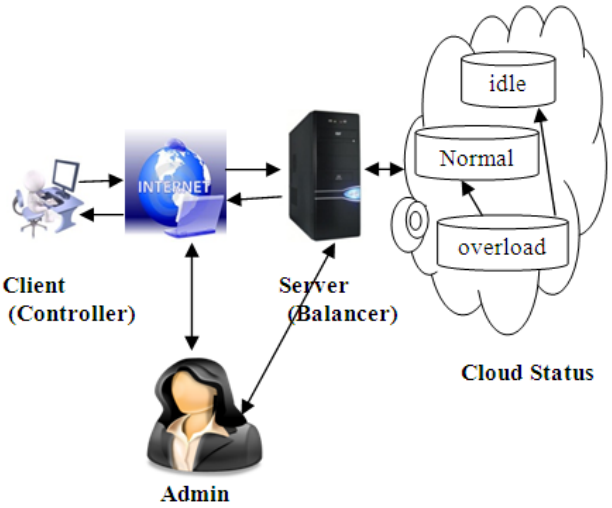


Fig.1. Typical Cloud Partition

Figure 1 shows the main architecture for typical cloud partitioning .The load balance solution is done by the main controller (admin) and the balancers (server). The main controller first assigns jobs to the suitable cloud partition and then communicates with the balancers in each partition to refresh this status information. Since the main controller deals with information for each partition, smaller data sets 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. The relationship between the balancers and the main controller is shown in the above figure 1 if there is no process for online shopping then the cloud partitioning is idle if it processes 2 times it's in normal state and if more than 3 times it is in overloaded state.

III. Proposed Model

A. Main Controller and Balancer

The load balance solution is done by the main controller and the balancers. The main controller first assigns jobs to the suitable cloud partition and then communicates with the balancers in each partition to refresh this status information. Since the main controller deals with information for each partition, smaller data sets 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.

B. Assigning jobs to the cloud partition

When a job arrives at the public cloud, the first step is to choose the right partition. The cloud partition status can be divided into three types:

(1) Idle: When the percentage of idle nodes exceeds, change to idle status.
Assigning jobs to the nodes in the cloud partition
(2) Normal: When the percentage of the normal nodes exceeds, change to normal load status.
(3) Overload: When the percentage of the overloaded nodes exceeds, change to overloaded status.

The main controller has to communicate with the balancers frequently to refresh the status information. The main controller then dispatches the jobs using the following strategy: When job i arrives at the system, the main controller queries the cloud partition where job is located. If this location's status is idle or normal, the job is handled locally. If not, another cloud partition is found that is not overloaded.

C. Assigning jobs to the nodes in the cloud partition

The cloud partition balancer gathers load information from every node to evaluate the cloud partition status. This evaluation of each node's load status is very important. The first task is to define the load degree of each nodes. The node load degree is related to various static parameters and dynamic parameters. The static parameters include the number of CPU's, the CPU processing speeds, the memory size, etc. Dynamic parameters are the memory utilization ratio, the CPU utilization ratio, the network bandwidth, etc.

D. Load balancing evaluation

In this module, we develop a load balancing evaluation model using developing a Cloud system for Online Shopping and deploying it. As it's a Online Shopping, many users from many locations can use the System. So there may occur the problem of Load. To overcome this problem using our model, we evaluate it and show the dynamic performance of our system by maintaining the job states based on their load condition and prove our system.

IV. CLOUD PARTITIONING LOAD BALANCING STRATEGY AND ALGORITHM DESCRIPTION

IDLE: Round Robin algorithm based on the load degree evaluation

When the cloud partition is idle, many computing resources are available and relatively few jobs are arriving. In this situation, this cloud partition has the ability to process jobs as quickly as possible so a simple load balancing method can be used. There are many simple load balance algorithm methods such as the Random algorithm, the Weight Round Robin, and the Dynamic Round Robin. The Round Robin algorithm is used here for its simplicity.

The Round Robin algorithm is one of the simplest load balancing algorithms, which passes each new request to the next server in the queue. The algorithm does not record the status of each connection so it has no status information. In the regular Round Robin algorithm, every node has an equal opportunity to be chosen. However, in a public cloud, the configuration and the performance of each node will be not the same; thus, this method may overload some nodes. Thus, an improved Round Robin algorithm is used, which called "Round Robin based on the load degree evaluation". The algorithm is still fairly simple. Before the Round Robin step, the nodes in the load balancing table are ordered based on the load degree from the lowest to the highest. The system builds a circular queue and walks through the queue again and again. Jobs will then be assigned to nodes with low load degrees. The node order will be changed when the balancer refreshes the Load Status Table. at the refresh period T . When the balance table is refreshed, at this moment, if a job arrives at the cloud partition, it will bring the inconsistent problem. The system status will have changed but the information will still be old. This may lead to an erroneous load strategy choice and an erroneous nodes order. A flag is also assigned to each table to indicate Read or Write. When the flag = "Read", then the Round Robin based on the load degree evaluation algorithm is using this table. When the flag = "Write", the table is being refreshed, new information is written into this table. Thus, at each moment, one table gives the correct node locations in the queue for the improved Round Robin algorithm, while the other is being prepared with the updated information. Once the data is refreshed, the table flag is changed to "Read" and the other table's flag is changed to "Write".

NORMAL: Game theory algorithm (non-cooperative games) Load balancing strategy for the normal status.

When the cloud partition is normal, jobs are arriving much faster than in the idle state and the situation is far more complex, so a different strategy is used for the load balancing. Each user wants his jobs completed in the shortest time, so the public cloud needs a method that can complete the jobs of all users with reasonable response time. Penmatsa and Chronopoulos [13] proposed a static load balancing strategy based on game theory for distributed systems. And this work provides us with a new review of the load balance problem in the cloud environment. As an implementation of distributed system, the load balancing in the cloud computing environment can be viewed as a game. Game theory has non-cooperative games and cooperative games. In cooperative games, the decision makers eventually come to an agreement which is called a binding agreement. Each decision maker decides by comparing notes with each others. In non-cooperative games, each decision maker makes decisions only for his own benefit. The system then reaches the Nash equilibrium, where each decision maker makes the optimized decision (perfect). The Nash equilibrium is when each player in the game has chosen a strategy and no player can benefit by changing his or her strategy while the other player's strategies remain unchanged. Nash equilibrium to minimize the response time of each job. The load balancing strategy for a cloud partition in the normal load status can be viewed as a non cooperative game.

Game Theory algorithm (cooperative games)

In cooperative games, the decision makers eventually come to an agreement which is called a binding agreement

Conclusion: when all the server is overloaded then decision makers i.e service provider will shift the server from overloaded server to normal

Load Balancing Strategy for idle Status

When the cloud partition is idle, many computing resources are available and relatively few jobs are arriving. In this situation, this cloud partition has the ability to process jobs as quickly as possible so a simple load balancing method can be used.

- Idle when
  Load Degree (N)=0
  N=Process

Load Balancing Strategy for Normal Status

When the cloud partition is normal, jobs are arriving much faster than in the idle state and the situation is far more complex, so a different strategy is used for the load balancing. Each user wants his jobs completed in the shortest time, so the public cloud needs a method that can complete the jobs of all users with reasonable response time.

- Normal when
  0< Load Degree (N)<=$Load\_Degree_{high}$
  N=Process
  $Load\_Degree_{high}$=3 processes
  The node is normal & it can process other jobs

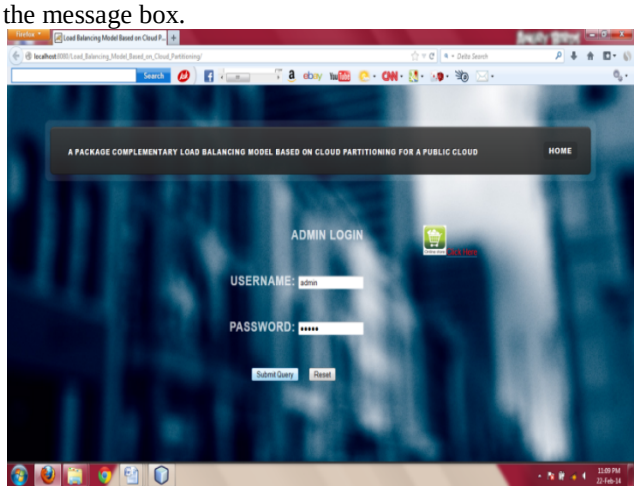
Load Balancing Strategy for Overloaded Status

- Overloaded When $Load\_Degree_{high}$<=Load Degree(N)
  The node is not available & cannot receive jobs until it return to the normal status

V. SNAPSHOTS

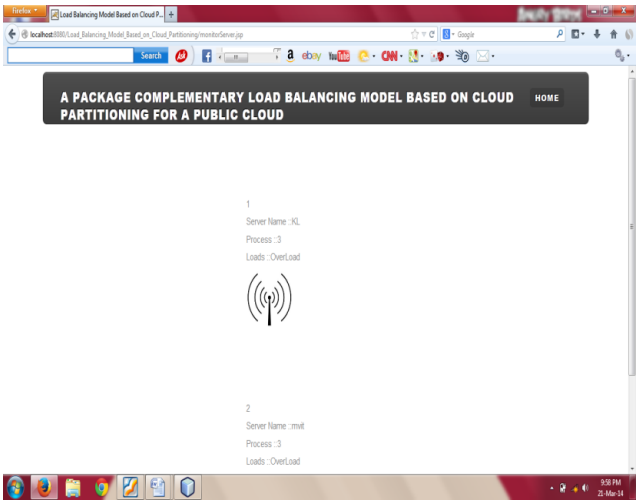
1. Admin Login Page

Here we need to login for main controller in order to know the status of the cloud partitioning with same username, password and image. Here we have written the test cases for all the three conditions and error is alerted in the message box.



2. Cloud Status

Here we get the cloud status for added server by processing the website of online shopping.



VI. CONCLUSION

Resource Management is an important issue in cloud environment. Cloud computing is the delivery of computing and storage capacity as a service to a community of end-recipients. The name comes from the use of a cloud-shaped symbol as an abstraction for the complex infrastructure it contains in system diagrams. Cloud computing entrusts services with a user's data, software and computation over a network. We shown 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.

Amazon EC2 is used to launch the instance in Oregon region and deploy the application.

FUTURE WORK

Setting up a refresh period: For load balancing the main controller and balancers should have the most recent status information. So, the statistical information should be refreshed at fixed time intervals. The interval should not be too short or too long. Tests and statistical tools are needed to set the refresh interval.

Better method for load status calculation: Good load balancing can be achieved by using a better algorithm. It will help to set accurate load degree parameters. The balancing mechanism should be more comprehensive

ACKNOWLEDGEMENTS

We would like to thank the editors and anonymous Reviewers for their valuable comments and helpful Suggestions.

REFERENCES

[1] R. Hunter, The why of cloud, http://www.gartner.com/ DisplayDocument? doc cd=226469&ref= g noreg, 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/cloudcomputing?query=cloud%20computing, 2012.
[5] GoogleTrends,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/white-papers/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 Software Modeling, Singapore, 2011.
[9] K. Nishant, P. Sharma, V. Krishna, C. Gupta, K. P. Singh. Nitin, and R. Rastogi, Load balancing of nodes in cloud using ant colony optimization, in Proc. 14th International Conference on Computer Modeling and Simulation (UKSim), Cambridge shire, 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.
[11] A.Rouse, Publiccloud, http://searchcloudcomputing.techtarget.com/definition/public-cloud, 2012.
[12] D. MacVittie, Intro to load balancing for developers—The algorithms, https://devcentral.f5.com/blogs/us/intro-to-load-balancing-for-developers-ndash-the-algorithms,2012.
[13] S. Penmatsa and A. T. Chronopoulos, Game-theoretic static load balancing for distributed systems, Journal of Parallel and Distributed Computing, vol. 71, no. 4,pp. 537-555, Apr. 2011.
[14] D. Grosu, A. T. Chronopoulos, and M. Y. Leung, Load balancing in distributed systems: An approach using cooperative games, in Proc. 16th IEEE Intl.Parallel and Distributed Processing Symp., Florida, USA, Apr. 2002,pp. 52-61.
[15] S. Aote and M. U. Kharat, A game-theoretic model for dynamic load balancing in distributed systems, in Proc. the International Conference on Advances in Computing, Communication and Control (ICAC3 '09), New York, USA, 2009, pp. 235-238.Gaochao Xu.

AUTHOR'S PROFILE

Ashwini Patil

is student of M.Tech. in Department of Computer Science and Engineering pursuing in
M S Engineering College Bangalore affiliated to VTU Belgaum. Has done my BE in Information Science from SIET. Area of interest are Software Engineering, Cloud Computing.

Mrs. Aruna M. G.

has done her BE in CSE from Bangalore University in 2001.M.Tech in CSE from Dr. MGR university in 2006 pursuing her PhD in Computer Networks from VTU. Research interests are Cloud Computing, Computer Networks, Information Security, and Cryptography. Her areas of interest are Database, Grid Computing.