A Survey On 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— Cloud Computing is an emerging computing paradigm. It aims to share data, calculations, and service transparently over a scalable network of nodes. Since Cloud computing stores the data and disseminated resources in the open environment.So, the amount of data storage increases quickly. In the cloud storage, load balancing is a key issue. It would consume a lot of cost to maintain load information, since the system is too huge to timely disperse load.

Load balancing is one of the main challenges in cloud computing which is required to distribute the dynamic workload across multiple nodes to ensure that no single node is overwhelmed. It helps in optimal utilization of resources and hence in enhancing the performance of the system. A few existing scheduling algorithms can maintain load balancing and provide better strategies through efficient job scheduling and resource allocation techniques as well. In order to gain maximum profits with optimized load balancing algorithms, it is necessary to utilize resources efficiently. This paper discusses some of the existing load balancing algorithms in cloud computing and also their challenges.

Key Words: Load Balancing in Cloud Computing.

I.INTRODUCTION

Cloud computing is the use of computing resources (hardware and software) that are delivered as a service over a network (typically the Internet). The name comes from the common use of a cloud-shaped symbol as an abstraction for the complex infrastructure it contains in system diagrams. Cloud computing entrusts remote services with a user's data, software and computation. Cloud computing consists of hardware and software resources made available on the Internet as managed third-party services.

The term cloud computing encompasses many different types of services. Therefore, evaluating business needs carefully before choosing a cloud vendor is imperative. Software-, Platform-, and Infrastructure-as-a-service vendors differ not only in the type of products they offer, but also in their cloud architectural infrastructure. The broad architectural differences in cloud computing products, the drawbacks to more generic approaches in cloud delivery, and the best practices philosophy of constructing cloud computing infrastructures are examined. Ultimately, the cloud architecture is described from server and operating system through data center and software development platform.

Not All Clouds Deliver the Same Service: Cloud computing be as simple as a web-based application available over the Internet or as complex as an array of high-powered, massively provisioned data centers capable of processing thousands of online transactions per second for millions of worldwide customers. The only constant is the cloud centrally located computing resources available over the Internet or wide area network.1 Reviewing the broad types of cloud infrastructures and their primary IT and business functions helps businesses more fully understand this bewildering span of cloud computing products and their capabilities.

Different types of services as shown in the below figure 1.1 are as follows

a.SAAS

b.PAAS

c.IAAS

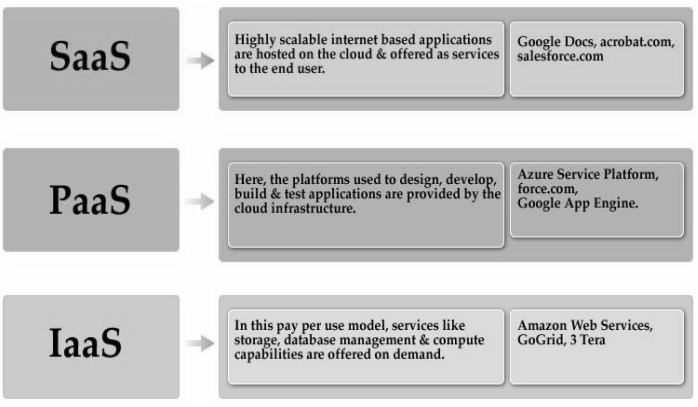

Figure.1 Service Models

a. Software as a Service

- Among products now defined as cloud computing, Software as a Service (SaaS) is the most targeted in terms of functionality.
- With respect to the above figure 1.1 Companies rent time and space from online applications ranging from sales lead tracking software to Web-based email.
- Enterprise level SaaS providers deliver a wide variety of sophisticated applications such as product lifecycle management, supply chain management, and many other vertical applications. If a business needs these specific applications, the SaaS model can save them the expense of buying hardware, software, and long-term maintenance.
- In many cases, drawbacks include the inability to fully customize the solution for individual business requirements, complexities in integrating the SaaS application with the existing business IT infrastructure, and difficulties in predicting and budgeting pay-as-you-go pricing.
- For many companies, however, these targeted applications work well to solve specific business problems and are not meant to replace or significantly augment corporate computing resources.

b. Platform as a Service

- When businesses need an as-needed application development environment, Platform as a Service (PaaS) is an effective and efficient type of cloud computing resource to consider.
- With respect above figure 1.1 Most importantly, companies can produce new applications more quickly and with a greater degree of flexibility than with older development platforms.
- Programmers and development managers especially appreciate that the Service Provider handles all the care and maintenance of the underlying operating system,Server, storage, and application containers.
- PaaS is especially useful when development teams are widespread geographically or when partner companies or divisions share development efforts.
- Engineers can more easily share and back up a central repository of application data as well as implement tighter version control and environment variables.
- PaaS cloud segments for development that painstakingly mirror the actual deployment environment of the application ensure a more stable and polished finished product. Companies can build up and tear down PaaS environments on demand with no capital expenditures or long-term investment.
- One significant drawback with many PaaS implementations, however, is the tie-in to one vendor's platform and infrastructure.
- Customers need to ensure that the platform allows for maximum portability of applications and data. Overall, companies can save considerable capital and operating expenses using a PaaS solution.

C. Infrastructure as a Service

- The most foundational use of cloud computing is Infrastructure as a Service (IaaS): the rental of a complete computing platform for running applications, hosting a company's entire computing environment.
- The latter use is not as common, and the majority of companies deploying IaaS now use it as a means to expand their computing capabilities in targeted IT areas without drastically increasing capital expenditures on new hardware and software. In fact, to better serve this market, most IaaS vendors emphasize their server and storage space as opposed to providing complete data center or application services. This server rack and disk space rental model is a natural outgrowth of the industry.
- Some of the most prominent IaaS suppliers are e-commerce and Internet information businesses that first entered the cloud computing market as a means to recoup revenue from excess, unused hardware in their data centers. Server rack space is a useful commodity for businesses that require more computing power but want to avoid long-term capital outlays.
- However, many companies have realized that rack space and virtual server rentals do not provide a complete solution, and reliance on ever-increasing cloud-based hardware resources produces many of the same IT management issues as outright server and network hardware ownership.

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.Essential Characteristics.

Cloud computing is an attracting technology in the field of computer science. In Gartner's report, 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. 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.

Load balancing schemes depending on whether the system dynamics are important can be either static and 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. A dynamic scheme is used here for its flexibility. The model has a main controller and balancers to gather and analyze the information. Thus, the dynamic control has little influence on the other working nodes. The system status then provides a basis for choosing the right load balancing strategy.

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.

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 visionof 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 ondemand, 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 centres. 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 centres 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 distrubuted system .They used Cooperative game to model the load balancing problem and their solution was based on Nash Equlibrium 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 timecompared 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 distrubuted systems like:1) Global approach 2)Cooperative approach 3)Nonoperative approach.

They used noncooperative 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 soved 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 pa-pers 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. LOAD BALANCING

Load Balancing is a computer networking method to distribute workload across multiple computers or a computer cluster, network links, central processing units, disk drives, or other resources, to achieve optimal resource utilization, maximize throughput, minimize response time, and avoid overload. Using multiple components with load balancing, instead of a single component, may increase reliability through redundancy. The load balancing service is usually provided by dedicated software or hardware, such as a multilayer switch or a Domain Name System server.Load balancing is one of the central issues in cloud computing . It is a mechanism that distributes the dynamic local workload evenly across all the nodes in the whole cloud to avoid a situation where some nodes are heavily loaded while others are idle or doing little work. It helps to achieve a high user satisfaction and resource utilization ratio, hence improving the overall performance and resource utility of the system. It also ensures that every computing resource is distributed efficiently and fairly . It further prevents bottlenecks of the system which may occur due to load imbalance. When one or more components of any service fail, load balancing helps in continuation of the service by implementing fair-over, i.e. in provisioning and de-provisioning of instances of applications without fail.The goal of load balancing is improving the performance by balancing the load among these various resources (network links, central processing units, disk drives .) to achieve optimal resource utilization, maximum throu ghput, maximum response time, and avoiding overload. To distribute load on different systems, different load balancing algorithms are used.

IV. METRICS FOR LOAD BALANCING IN CLOUD

Various metrics considered in existing load balancing techniques in cloud computing are discussed below- Scalability is the ability of an algorithm to perform load balancing for a system with any finite number of nodes. This metric should be improved.

- Resource Utilization is used to check the utilization of re-sources. It should be optimized for an efficient load balancing.
- Performance is used to check the efficiency of the system. This has to be improved at a reasonable cost, e.g., reduce task response time while keeping acceptable delays.
- Response Time is the amount of time taken to respond by a particular load bal ancing algorithm in a distributed system. This parameter should be minimized.
- Overhead Associated determines the amount of overhead involved while implementing a load-balancing
  algorithm. It is composed of overhead due to movement of tasks, inter-processor and interprocess communication. This should be minimized so that a load balancing technique can work efficiently.

V. CLOUD PARTITIONING LOAD BALANCING STRATERGY 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 players 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 & can not receive jobs until it return to the normal status

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.

VII 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/cloud-computing?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,N. Nitin, and R. Rastogi, Load balancing of nodes in cloud using ant colony optimization, in Proc. 14th International Conference on Computer Modelling andSimulation (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.
[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.