ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

Survey on Load Rebalancing For Distributed File System in Cloud with Security

Jayesh D. Kamble, Prof. Y.B.Gurav

IInd Year ME, Department of Computer Engineering, PVPIT, Pune, India

Associate Professor& HOD of Department of Computer Engineering, PVPIT, Pune, India

ABSTRACT: In a Cloud computing, distributed file system is used as a key building block by using map reduce paradigm. In such systems, the node performs different operations like computing as well as storage. In distributed file systems, those different operations are performed on different nodes parallel by partitioning a large file into small chunks.In a cloud computing setting, failure is that the norm and nodes is also upgraded, replaced, and accessorial within the system. Files may get created, deleted, and appended dynamically. This will affect as load imbalance in a distributed file system; it means the file chunks are not distributed as uniformly as possible among the nodes. In cloud, if number of storage nodes, number of files and accesses to that file increases then the central node (master in MapReduce) becomes bottleneck. The load rebalancing task is used to eliminate the load on central node. Using load rebalancing algorithm the load of nodes is balanced as well as the movement cost is reduced. In this survey paper the problem of load imbalancing is overcome. And we are going to consider the security while rebalancing the load of distributed file system as well.

KEYWORDS: Cloud, Distributed File System, HDFS, Hadoop, Load Rebalancing, Security

I. INTRODUCTION

As we know Cloud computing is emerging as a new archetype of large scale distributed computing. Cloud computing is responsible for moving computation and data storage away from desktop to portable PCs into large data center, which became part of computer science now. It has the capability to harness the power of Internet and wide area network to resources that are available remotely, thereby, providing cost effective solution to the most of the real life requirement. It provides the scalable IT resources such as applications and service, as well as infrastructure on which they operate, over the Internet, as pay-as-per-use basis to adjust the capacity quickly and easily. It helps to accommodate changes in demand.

A cloud computing system is very user friendly, as it not require any expertise to use. It is sold on demand, typically by the minute or the hour. A cloud can be private or public. Private or public, the main objective of cloud computing is to provide easy, scalable access to computing resources and IT services. Cloud computing provide whole things as a service to their users, like as: storage of data as a service, application software as a service, computing platform as a service and computing infrastructure as a service etc.

In Cloud computing, the number of computer systems that are connected using communication network. There are various characteristics of cloud i.e. Scalable, on demand service, User centric, Powerful, Versatile, Platform independent etc. In cloud three technologies are included the MapReduce programming, Virtualization and distributed file systems for the data storage purpose.

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

Distributed file system is effective model of file systems that is used in the form of chunks for cloud computing. Map reduce programming of distributed file system is used in cloud computing applications. Map reduce is nothing but the master-slave architecture in hadoop. Master act like Namenode and Slave act like Datanode. Master takes large problem, divides it into sub problem and assigns it to worker node i.e. to multiple slaves to solve problem individually. In distributed file system, a large file is divided into number of chunks and allocates each chunk to separate node to perform MapReduce function parallel over each node. In a distributed file system, the load of a node is directly proportional to the number of file chunks the node possesses. There are chances of files in a cloud of getting created, deleted, and appended, and nodes may be upgraded, replaced and added arbitrarily in the file system, distribution of the file chunks are not uniformly distributed to among the nodes. Among storage nodes load balance is a critical function in clouds. The distributed file systems in clouds rely on central nodes to manage the metadata information of the file systems and to balance the loads of storage nodes based on that metadata.

Now a days the increase in storage and network, load balancing is the main factor in the large scale distributed systems. Load should be balance over multiple nodes to improve system performance, resource utilization, response time and stability. Load balancing is divided into two categories: static and dynamic. In static load balancing algorithm, it does not consider the previous behavior of a node while distribute the load. But in case of dynamic load balancing algorithm, it checks the previous behavior of node while distribute the load. In cloud, if number of storage nodes, number of files and assesses to that file increases then the central node (master in MapReduce) becomes bottleneck. The load rebalancing task is used to eliminate the load on central node.

In distributed file systems, we will see the different approaches of how to reduce network traffic (or *movement cost*) caused by rebalancing the loads of nodes as much as possible to increase the network bandwidth to cloud applications. For security in cloud computing, we can maintain these files in encrypted format using cryptographic algorithms.

II. BACKGROUND

A. Virtualization:

Virtualization is the most intellectual change that PCs and servers have experienced, said Simon Crosby, chieftechnology officer for Citrix Systems' Data Center and Cloud Division [11]. "IT departments have long been at themercy of the technical demands of legacy applications", explained Chris Van Dyke, [12] Microsoft's chieftechnology strategist for the oil and gas industry. "Now, instead of having to maintain older operating systemsbecause of the needs of a legacy application, IT departments can take advantage of the performance and securitygains in a new OS (in one virtual machine) while supporting legacy applications in another. Also, the process ofdeploying applications becomes simpler, because applications can be virtualized and deployed as a single virtualmachine". [14]Single PC or server parallel can run multiple operating systems or multiple sessions of a single OS by using Virtualization technology. This lets users put numerous applications even those that run on different operatingsystems on a single PC or server instead of having to host them on separate machines as in the past. The approach isthus becoming a common way for businesses and individuals to optimize their hardware usage by maximizing thenumber and kinds of jobs a single CPU can handle.[13].

B. Hyperviser:

It is nothing but multiple operating systems (or multiple instances of the same operating system) on a single computer system. The hypervisor manages the system's processor, memory, and other resources to allocate what each operating system requires. .[15]

C. I-A-A-S

The capability provided to the consumer is to provision processing, storage, networks, and otherfundamental computing resources where the consumer is able to deploy and run arbitrary software, which caninclude operating systems and applications[18]. The consumer does not manage or control the underlying cloudphysical infrastructure but has control over operating systems, storage, deployed applications, and possibly limitedcontrol of select networking components[17].

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

D. Private Cloud

The cloud infrastructure is operated solely for an organization.[19] It may be managed by the organizationor a third party and may exist on premise or off premise.

E. Parallel Data Processing

Particular tasks of processing a job can be assigned to different types of virtual machines which areautomatically instantiated and terminated during the job execution, parallel.[16].

F. Distributed File System

Files are stored on different storage resources, but appear to users as they are put on a single location. Adistributed file system should be transparent, fault-tolerant and scalable.[20]

III. LITERATURE SURVEY

- ***MapReduce: Simplified Data Processing On Large Clusters [5]***

In distributed systems, in implementation for processing and generating large scale datasets, MapReduceprogramming model used. It is used at Google for many different purposes. Map and reduce are the functions which is used over here. Map functiongenerate set of intermediate key pairs and reduce function merges all intermediate key values associated with sameintermediate key. The map and reduce function allows to perform parallelize operation easily and re-execute themechanism for fault tolerance. At the run-time, system takes care of detail information of partitioning the input data,schedule the program execution across number of available machines, handling failures and managing intercommunicationbetween machines.In distributed file system nodes simultaneously perform computing and storage operations. The large file inpartitioned into number of chunks and allocate it to distinct nodes to perform MapReduce task parallel over nodes.Typically, MapReduce task processes on many terabytes of data on thousands of machines. This model is easy to use; ithides the details of parallelization, optimization, fault-tolerance and load balancing. MapReduce is used for Google'sproduction Web search service, machine learning, data mining, etc. Using this programming model, redundant executionused to reduce the impact of slow machines, handle machine failure as well as data loss.

- ***Load Balancing Algorithm for DHT based structured Peer to Peer System [6]***

In distributed environment, Peer to peer system has an emerging application. As compared to client-serverarchitecture, peer to peer system improved resource utilization by making use of unused resources over network. Peer topeer system uses Distributed Hash Table (DHTs) as an allocation mechanism. It perform join, leave and updateoperations. Here load balancing algorithm uses the concept of virtual server to temporary storage of data. Using theheterogeneous indexing, peers balanced their loads proportional to their capacities. In this, decentralized load balancealgorithm construct network to manipulate global information and organized in tree shape fashion. Each peer canindependently compute probability distribution capacities of participating peers and reallocate their load in parallel.

- ***Optimized Cloud Partitioning Technique to Simplify Load Balancing [7]***

Cloud computing has some issue regarding resource management and load balancing. In this paper, Cloudenvironment is divided into number of parts by making use of cloud cluster technique which helps for the process of loadbalancing. The cloud consists of number of nodes and it partitioned into n cluster based on cloud cluster technique. Inthis, it consists of main controller which maintains all information regarding all load balancer in cluster and index table.Initial step is to choose the correct cluster and it follows algorithm as

1) In cloud environment, the nodes connected to central controller are initialized as 0 in index table.
2) When controller receives new request, it queries the load balancer of each cluster for job allocation.
3) Then controller pass index table to find next available node having less weight. If found then continue theprocessing otherwise index table reinitialized to 0 and in an increment manner then again controller passes tableto find next available node.
4) After completing the process the load balancer update the status in allocation table.

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

Cloud partitioning method consist 2 steps:-
1) Visit each node randomly to match it with neighbor node. If it having same characteristics and shares similardata with minimal cost then two nodes are combined into new node with share same details. Repeat until there isno neighbor node having similar characteristics. Subsequently update the cost between neighbor two nodes andcurrent neighbor node.
2) After joining two nodes into new node having similar characteristics visited node send the information to newnode instead of sending it twice.It gives the high performance, stability, minimum response time and optimal resource utilization.

- ***Histogram-Based Global Load Balancing in Structured Peer to Peer System [8]***

Peer to peer system having solution for sharing and locating resources over internet. In this paper there are twokey components. First is histogram manager that maintain histogram that reflect global view of distribution of load.Histogram stores statistical information about average load of no overlapping groups of nodes. It is used to checkwhether node is normally loaded, Light or heavily loaded. Second component is load balance manager that take theaction of load redistribution if node becomes light or heavy. Load balancing manager balance the load statically whennew node joins and dynamically when existing node become light or heavily loaded. The cost of constructing histogramand maintaining it may be expensive in dynamic system. To reduce the maintaining cost two techniques are used.Constructing and maintaining histogram is expensive if node join and leave system frequently. Every new node in peer topeer system find its neighbor node and these neighbor nodes need to share its information with new node to setupconnection. Now the cost of histogram is totally based on histogram update message caused by changing the load ofnodes in the system .To reduce the cost approximate value of histogram is taken.

IV. LOAD BALANCING

In distributed file system, to improve the resource utilization and job response time the process load balancing on nodes widely used. In Load balancing, distribution of the load among various nodes of a distributed system as well as avoiding a situation where some of the nodes are heavily loaded while other nodes are idle or doing very little work. Load balancing assures 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. The main goal is to develop an effective load balancing algorithm using divisible load scheduling theorem to maximize or minimize different performance parameters for the clouds of different sizes. With the help of this, it ensures process of reassigning the total load to the individual nodes of the collective system to make resource utilization effective and to improve the response time of the job, simultaneously removing a condition in which some of the nodes are over loaded while some others are under loaded. A load balancing algorithm, it depends on the present behavior of the system which is dynamic in nature does not consider the previous state or behavior of the system. The important things to consider while developing such algorithm are : estimation of load, comparison of load, stability of different system, performance of system, interaction between the nodes, nature of work to be transferred, selecting of nodes and many other ones. Elimination the dependence on central nodes is known as load rebalancing. The storage nodes are structured as a network based on distributed hash tables (DHT). DHTs enable nodes to self-organize and repair while constantly offering lookup functionality in node dynamism, simplifying the system provision and management. Specifically, in this study, suggest offloading the load rebalancing task to storage nodes by having the storage nodes balance their loads spontaneously. This eliminates the dependence on central nodes. Storage nodes are structured as a network based on distributed hash tables discovering a file chunk can simply refer to rapid key lookup in DHTs, and given that a unique handle is assigned to each file chunk. DHTs enable nodes to self-organize and repair while constantly offering lookup functionality in node dynamism, simplifying the system provision and management.

V. SECURITY WHILE LOAD REBALANCING

Our objective is to allocate the chunks of files as uniformly as possible among the nodes such that no node manages an excessive number of chunks. And the most important part is security that we can provide while rebalancing of load in distributed file system. The security features in CDH4 enable Hadoop to prevent malicious user impersonation. The Hadoop daemons leverage Kerberos to perform user authentication on all remote procedure calls

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

(RPCs). Group resolution is performed on the Hadoop master nodes, NameNode, JobTracker and ResourceManager to guarantee that group membership cannot be manipulated by users. The user who submitted job, Map tasks are run under these user accounts, ensuring isolation there. In addition to these features, new authorization mechanisms have been introduced to HDFS and MapReduce to enable more control over user access to data. The security features in CDH4 meet the needs of most Hadoop customers because typically the cluster is accessible only to trusted personnel. In particular, Hadoop's current threat model assumes that users cannot:

1. Have root access to cluster machines.
2. Have root access to shared client machines.
3. Read or modify packets on the network of the cluster.

VI. CONCLUSION

Balancing the overloads &underload of nodes and reducing the demanded movement cost as much as possible, while taking advantage of physical network locality and node heterogeneity can be done by this approaches. And can improve performance while rebalancing of nodes. If security is provided while rebalancing, it will be more beneficial. Load imbalance factor, movement cost, and algorithmic overhead can be handled efficiently. To securing the data, implemented the RSA algorithm. Load Rebalancing for Distributed File Systems in Clouds can discard the issue like high delays, handle heterogeneous resources, efficiently adjust to dynamic operational conditions, offer efficient task distribution, and so it can provide minimum node idle time.

REFERENCES

1. Hung-Chang Hsiao, Hsueh-Yi Chung, HaiyingShen and Yu-Chang Chao, “Load rebalancing for distributed file systems in cloud” IEEE Trans. On parallel and distributed systems, vol. 24, no. 5, pp.951-962, May 2013
2. Hadoop Distributed File System, http://hadoop.apache.org/hdfs/, 2012.
3. HDFS Federation, http://hadoop.apache.org/common/docs/r0.23.0/hadoop-yarn/hadoop-yarn-site/Federation.html, 2012.
4. U.Karthik Kumar,” A Dynamic LoadBalancing Algorithm in Computational GridUsing Fair Scheduling” International Journal of Computer Science Issues, Vol. 8, Issue 5, No 1, September 2011.
5. J. Dean and S. Ghemawat, “MapReduce: Simplified Data Processing on Large Clusters,” Proc. Sixth Symp.Operating System Design and Implementation (OSDI ’04), pp. 137-150, Dec. 2004.
6. ChahitaTanak, Rajesh Bharati “Load Balancing Algorithm for DHT Based Structured Peer to PeerSystem”International Journal of Emerging Technology and Advanced Engineering (ISSN 2250-2459, ISO9001:2008 Certified Journal, Volume 3, Issue 1, January 2013)
7. P.Jamuna and R.Anand Kumar “Optimized Cloud Computing Technique To Simplify LoadBalancing”International Journal of Advanced Research in Computer Science and Software Engineering,Volume 3, Issue 11,November 2013.
8. QuangHieu Vu, Member, IEEE, Beng Chin Ooi, Martin Rinard, and Kian-Lee Tan “Histogram-Based GlobalLoad Balancing in Structured Peer-to-Peer Systems” IEEE transaction on knowledge and data engineering,vol.21, no. 4, April2009.
9. GaochaoXu, Junjie Pang, and XiaodongFu“A Load Balancing Model Based on Cloud Partitioning for the Public Cloud”, IEEETRANSACTIONS ON CLOUD COMPUTING YEAR 2013.
10. Lee, R. and B. Jeng, ”Load-balancing tactics in cloud,” in proc. International Conference on Cyber-Enabled Distributed Com-puting and Knowledge Discovery (CyberC), IEEE, pp:447-454, October 2011.
11. K. McKusick and S. Quinlan, “GFS: Evolution on Fast-Forward,” Comm. ACM, vol. 53, no. 3, pp. 42-49, Jan. 2010.
12. Vishnu S. Pendyala, Synopsys Simon S.Y. Shim, San Jose State University, ”The Web as the ubiquitous computer ,September 2009 ,Web Technologies, journal of IEEE Computer Society”P 90-92.
13. HDFS Federation, http://hadoop.apache.org/common/docs/ r0.23.0 / hadoop-yarn/hadoop-yarn-site/Federation.html, 2012.
14. A. Stoica, R. Morris, D. Liben-Nowell, D.R. Karger, M.F. Kaashoek,F. Dabek, and H. Balakrishnan, “Chord: A Scalable Peer-to-PeerLookup Protocol for Internet Applications,” IEEE/ACM Trans.Networking, vol. 11, no. 1, pp. 17-21, Feb. 2003.
15. Benjamin Depardon,CyrilS_eguin,Gael Le Mahec “Analysis of Six Distributed File Systems”,hal-00789086, version 1 - 15 Feb 2013.
16. M. Jelasity, S. Voulgaris, R. Guerraoui, A.-M. Kermarrec, and M.V. Steen, “Gossip-Based Peer Sampling,” ACM Trans. ComputerSystems, vol. 25, no. 3, Aug. 2007.
17. Sagan, Space-Filling Curves, first ed. Springer, 1994.
18. C. Guo, G. Lu, D. Li, H. Wu, X. Zhang, Y. Shi, C. Tian, Y. Zhang, and S. Lu, “BCube: A High Performance, Server-Centric NetworkArchitecture for Modular Data Centers,” Proc. ACM SIGCOMM ’09, pp. 63-74, Aug. 2009.
19. Abu-Libdeh, P. Costa, A. Rowstron, G. O’Shea, and A.Donnelly, “Symbiotic Routing in Future Data Centers,” Proc.ACMSIGCOMM ’10, pp. 51-62, Aug. 2010.
20. George Lawton., “Moving the OS to the Web,”, March 2008, journal of ieee computer society, P 16-19.

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

BIOGRAPHY

Mr. JayeshKamble is a student of Masters in Engineering, Computer Department, PVPIT ,Pune University. He received Bachelors of Engineering in 2013 from Pune University. His research interests are Computer Networks (Software Defined Networks), Network Security, Distributed Systems etc.

Prof.Y.B.Gurav is working as Associate Professor and Head of Department of Computer Engineering in PVPIT, Pune University with 16 years of teaching experience. He Completed his master in engineering (CSE) And nowhe is perusing his Ph.D. His research interests are Network Security, Distributed Systems, HCI, Compiler systems etc.