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. 3, Issue 6, June 2015

An Energy Efficient QoS Based Replication Strategy

Reena S. More[1], Prof.Nilesh V. Alone[2]

PG Student, Department of Computer Engineering, R H. Sapat College of Engineering, Nashik, M.S. India[1]

Head, Department of Computer Engineering, R H. Sapat College of Engineering, Nashik, M.S. India[2]

ABSTRACT: Cloud computing is a model that relies on sharing computing resources rather than having local servers or personal devices to handle applications.Data-intensiveapplications draw more and more attentions in the last few years.Cloud computing provides the tools and technologies to build data intensive parallel applications with much more affordable prices. Data Corruption occurrence on these data intensive applications does not meet the QoS Requirements. Different applications are having different quality of service (QoS). Every data is not being replicated on the high performance node. Data Intensive Application requires high cost to replicate data and as cloud computing includes number of different nodes it consumes large amount of energy. Huge amounts of energy is consumed by cloud which results in high operating costs and carbon dioxide emissions to the environment. So, to avoid this we propose a system that will reduce the cost for replication and also reduce the large amount of energy consumption of heterogeneous nodes.

KEYWORDS: QoS, Data Intensive Application.

I. INTRODUCTION

Large scale distributed systems are becoming more and more popular as data is being shared in data-intensive scientific applications. Cloud computing is evolved by addressing the Quality of Service and reliability problems. Compared to traditional parallel computing techniques Cloud computing provides the tools and technologies to build data intensive parallel applications with much more affordable prices. Applications that explore, query, analyse, visualise, and, in general, process very large scale data sets are known as Data Intensive Applications. In this information age, digital data in the scale of hundreds of petabyte are being generated every day. This includes electronic transactions, publications and media managed by government establishments and commercial organisations, electronic records from numerous sensors, social contents created by casual internet users, results from huge scientific experiments and many more. At such an astonishing information growth rate, it is not surprising that many organisations recognised the critical needs for data-intensive applications to process and analyse these ever-growing data in order to gain competitive advantages.Cloud computing provides scalable computing and storage resources to applications via the Internet. With the flexible and transparent features in the resource allocation various types of data-intensive applications are developed in the cloud computing environment.

As data intensive application require large storage cost. QoS requirement must be satisfied for large data intensive application [1].Cloud infrastructure includes heterogeneous node characteristics. Cloud computing includes node heterogeneity, every data is not being replicated on the high performance node. Few data may get replicated on the low performance node. Later if there is data corruption on the node, running high QoS application, data is retrieved from the low performance node. High QoS application may be violated as the data is being retrieved from low performance node. Low performance node are having slow disk access latencies and slow communication. Quality of Service requirement is defined from the aspect of the request information. Having large number of nodes in the cloud computing, it is very difficult to ask all nodes with the same performance and capacity in their CPUs, memory, disks. Cloud computing includes many storage nodes, and consume lots of energy. Our aim is to minimize the cost and consume less energy as possible. This is challenging as large data is included.

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. 3, Issue 6, June 2015

Cloud computing includes many storage nodes. Mainly Power consumption of the computing nodes is mostly determined by the CPU, disk storage, memory, network interfaces. The goal is to minimize the energy consumption inthe storage nodes.Energy is saved by processing the large amount of data on the high performance node. By this approach it is easy to replicate the data in less access time.

The paper is organized as follows: Section II overviews the related work carried out different methods for data replication, QoS requirement and energy consumption, while Section III Problem Statement. Section IV Consists of System Architecture. Section V Mathematical model. Section VI Implementation Section VII Experimental results regarding efficiency parameters of the proposed algorithm.

II. RELATED WORK

There are many survey related to different research topics, like Lazy Update, replica for content distribution.High availability and service level agreement must be satisfied for the application and deployment in the cloud.Huge storage cost is required for Data Intensive application to store data in the cloud.

Gao and L. Diao [2] use lazy update approach to improve the throughput of data access and to reduce the response time to separate the process of data replication and data access. Replica is placed in the appropriate server of the system.

W. Li, Y. Yang et. Al. [3] used a cost effective mechanism namely PRCR for the effective data management. For maintaining the data reliability this mechanism checks the availability of replicas. The disadvantage of the mechanism is it is not considering the performance of data access. X. Tang and J. Xu [4] described two classes of service models is being used: replica-aware services and replica-blind services. By knowing the location information, servers are capable of optimizing request to improve responsiveness. Hsiangkai Wang et al. [5] determines the positions of replicas to satisfy the quality requirements imposed by data requests. For heterogeneous environment Quality assurance is important. Data replication is an important technique to speed up data access in Data Grid.Anton Beloglazov et al. [6] says that cloud data centers consume enormous amounts of electrical energy resulting in high operating costs and carbon dioxide emissions. The goal is to improve the utilization of computing resources and reduce energy consumption under workload independent quality of service constraints

III. PROBLEM STATEMENT

Due to a large number of nodes in the cloud computing system, the probability of hardware failures is more. Some hardware failures will damage the disk data of nodes. As a result, the running data-intensive applications may not read data from disks successfully. To tolerate the data corruption, the data replication technique is extensively adopted in the cloud computing system to provide high data availability. However, the QoS requirement of an application is not taken into account in the data replication. When data corruption occurs, the QoS requirement of the high-QoS application cannot be supported continuously.Heterogeneous nodes are included and mainly Power consumption of the computing nodes is mostly determined by the CPU, disk storage, memory, network interfaces. The goal is to minimize the energy consumption in the storage nodes. Energy is saved by processing the large amount of data on the high performance node. By this approach it is easy to replicate the data in less access time.

IV. SYSTEM ARCHITECTURE

The main goal of the system is to consider the data intensive application for replication and to reduce the energy consumption.We refer to the architecture of the Hadoop distributed file system (HDFS) to design our replication algorithms as shown in Figure 1.In the proposed system Input data is divided into 64 MB Block and then it is replicated on datanodes. Data should not be replicated on the same rack. If data is placed on the same rack and if the rack failure occurs then both copies of data will be lost. First step is to check the health of cluster i.e. how many datanodes are in working. Second step Namenode takes the input form the client and divides it into 64 MB Block and will replicate on datanodes. By this time we can check the availability, bandwidth and memory of every datanode present in the cluster. After replication every time the available space and used space get updated. If data corruption occurs of any block Hadoop internally recovers from the different data nodes and will return the input data.Cloud infrastructure has different kind of nodes having different performance, and these different nodes consumes large amount of energy. The

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. 3, Issue 6, June 2015

main aim is to improve resource utilization as efficiently as possible.To handle the high energy use problem, it is required to eliminate inefficiencies of the resources that are utilized to serve application workloads. In the proposed system we achieve this goal by dynamically processing the large amount of data on high performance node.We have used hive to show the difference between the partition and nonpartition input data. The large amount of data is being processed on the high performance node. As data is not processed on low performance node their energy will be saved.

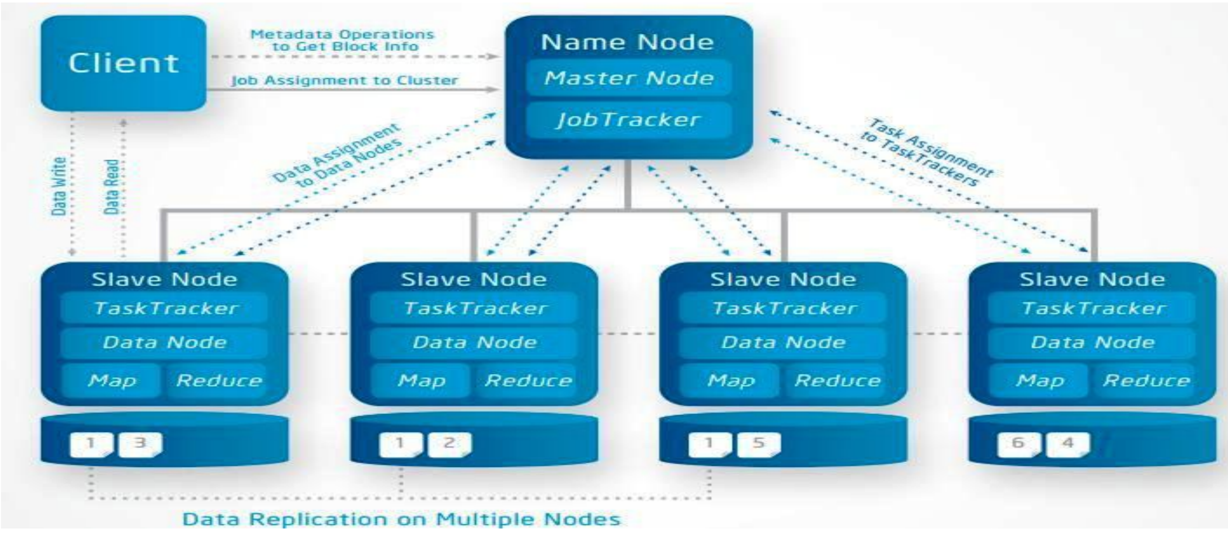

Figure 1. System architecture

V. MATHEMATICAL MODEL

Let S be a technique for Data Replication
Such That S= {s, e, X, Y, F} Where,

s represents Start state - Gets the data
e represents End state - View the results
X represents the set of inputs:
X= {D}
i. D= Data
Y is the set of outputs:
Y = {C}
i. C= Data Replication
F is the set of function
F= {F1, F2}

ii. F1= Divides the data into blocks of 64MB
iii. F2= Process the blocks into <key, value >pair using Map-Reduce.

Custom Partitioning

We have used Custom Partitioning algorithm to partition the data.The mechanism sending specific key-value pairs to specific reducers is called partitionin**g**.A Partitioner in Map Reduce world partitions the key space. The partitioner is used to derive the partition to which a key-value pair belongs. It is responsible for bring records with same key to same partition so that they can be processed together by a reducer. The partitioning phase takes place after the map phase and before the reduce phase. The number of partitions is equal to the number of reducers. The data gets partitioned across

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. 3, Issue 6, June 2015

the reducers according to the partitioning function. The basic idea is to partition a large problem into smaller subproblems. To the extent that the sub-problems are independentMap-Reduceprovides a very effective tool for tackling large-data problems. MapReduce is a programming model and an associated implementation for processing and generating large data sets.Users specify a *map* function that processes a key/value pair to generate a set of intermediate key/value pairs, and a *reduce* function that merges all intermediate values associated with the same intermediate key. Figure 2 shows how the Map-Reduce execute the kay-value pair.

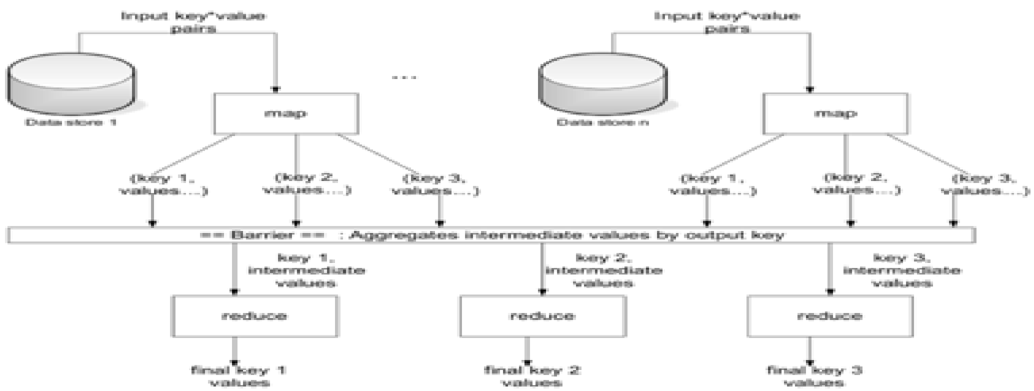

Figure 2. Map-Reduce

VI. IMPLEMENTATION

The system is implanted in Hadoop Framework. We have created Hadoop cluster. HDFS Components are used to form the cluster.Java is used for partitioning the data into data block. We have also used Hive to Show the comparative results between Partition and Non-partition data. Hadoop is installed on CentOS 6.4 Linux Operating System.

VII.SIMULATION RESULTS

The simulation studies involve the deterministic small cluster as shown in Figure 1. Hadoop is used to evaluate the performance of the proposed replication algorithms in a large-scale cloud computingsystem. Proposed algorithm is experimented under this cloud with 5 nodes. The proposed system is evaluated byconsidering the parameters such as replication cost, execution time. Itis the time taken to complete the replication process which involves disk I/O to process large volume of data,replicationidentification and allocation. While data center energy is widely being used and less attention is paid to the energyconsumption of the switching and transmission networks that are key to connecting users to the cloud. Cloud computing can enable more energy-efficient use of computing power, mainly when the computing tasks are having low intensity or infrequent or easily accessible.

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. 3, Issue 6, June 2015

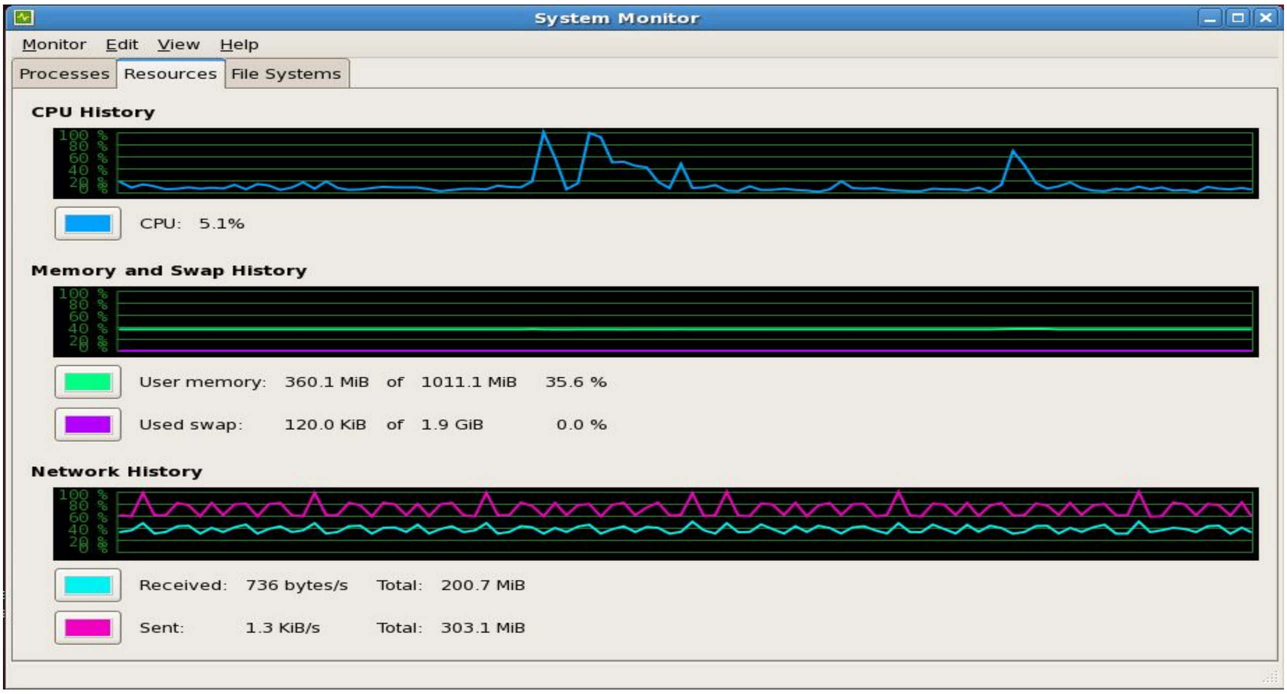

Figure 3-Health of current cluster

Figure 1. Shows the health of the resources. Out of Total Memory how much memory is in used and the available memory is shown. CPU use in this cluster and network history.

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
_SUCCESS	file	0 KB	3	64 MB	2015-05-24 07:16	rw-r--r--	hdfs	supergroup
_logs	dir				2015-05-24 07:15	rwxr-xr-x	hdfs	supergroup
part-r-00000	file	0.02 KB	3	64 MB	2015-05-24 07:16	rw-r--r--	hdfs	supergroup
part-r-00001	file	0.01 KB	3	64 MB	2015-05-24 07:16	rw-r--r--	hdfs	supergroup
part-r-00002	file	2.72 KB	3	64 MB	2015-05-24 07:16	rw-r--r--	hdfs	supergroup

Figure 4- Custom Partitioning result

Figure 4 shows the result of Partitioning. Input data is first accepted by Namenode.Name node will execute custom partitioning algorithm using Map-Reduce. In Custom Partitioning we are partioning datainto three partitions and replication factor is also of 3. As we want to replicate the large data on High performance node hence the size of third partition is set to large size.

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. 3, Issue 6, June 2015

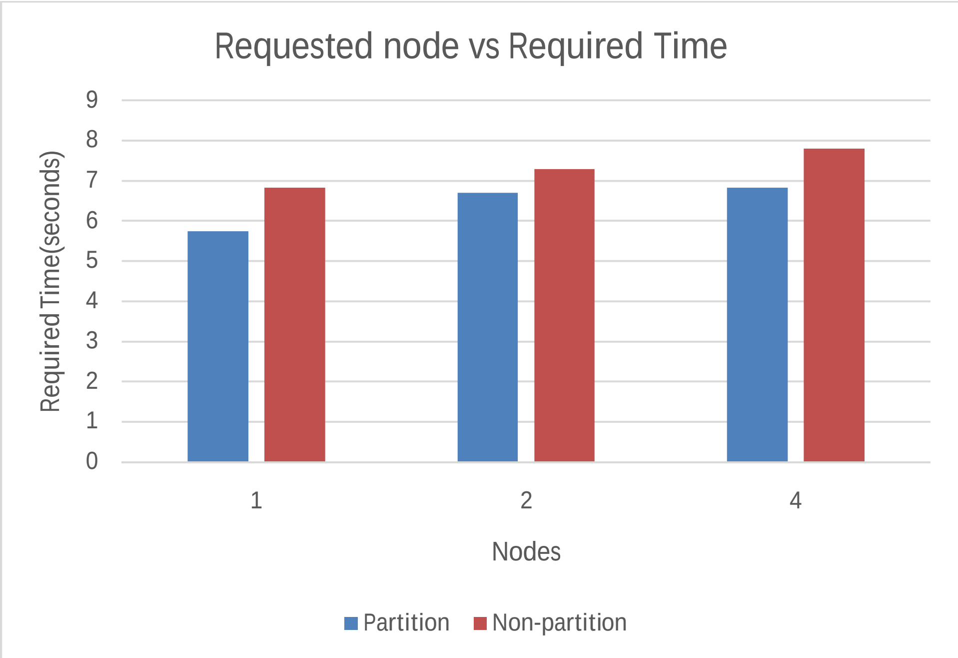

Figure 5.Partitioning and Non-Partitioning Requested node vs RequiredTime

Figure 5 shows the difference between partition and non-partitioning data. From this figure we conclude that partitioning the data and replicating is much better that replicating the data without partition.

VII. CONCLUSION

The simulation results showed that the proposed algorithm performs better with the total transmission energy metric. In this work, result of how the energy is being consumed in cloud infrastructure is tried to predict. The proposed system replicate the data on the high performance node and reduce the energy by processing the data on high performance node and also reduce the replication cost.So in this replication process we replicate the data on the highly qualified node and keep only the required nodes in the activate nodes. This save the energy of the computing nodes in the cloud computing environment.

ACKNOWLEDGMENT

R. S. More thanks Prof. N.V.Alone,Head of Comp. Engg Department, GES's R. H. Sapat College of engineering for his valuable suggestions and comments to improve the quality of this paper and also thanks toMrs. Archana Vaidya, PG Co-ordinator,GES's R. H. Sapat College of Engineering, Nashik.With a great sense of attitude the authors acknowledge Dr.Prafulla C. Kulkarni, Principal,GES's R. H. Sapat College of Engineering,Nashik for his kind support and blessings. We are also very much thankful to all those who indirectly helped in making this paper.

REFERENCES

[1] Jenn-Wei Lin,Chien-Hung Chen, and J. MorrisChang "QoS Aware Data Replication for Data Intensive Applications in Cloud Computing Systems", IEEE Transaction, Year 2013.

[2] A.Gao and L.Diao, "Lazy Update Propagation for Data Replication in Cloud Computing," in Proc. 2010 5th Int Conf. Pervasive Computing and Applications (ICPCA), Dec. 2010.

[3] W. Li, Y. Yang, J. Chen, and D. Yuan, "A Cost-Effective Mechanism for Cloud Data Reliability Management Based on Proactive Replica Checking," in Proc. 2012 12th IEEE/ACM Int. Symp. Cluster, Cloud and Grid Computing (CCGrid), May 2012.

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. 3, Issue 6, June 2015

[4] X. Tang and J. Xu, "QoS-Aware Replica Placementfor Content Distribution," IEEE Trans. Parallel and Distributed System, vol. 16, no. 10, pp. 921–932, Oct. 2005.
[5] H. Wang, P. Liu, and J.-J. Wu "A QoS-Aware Heuristic Algorithm for Replica Placement," in Proc. 7th, IEEE/ACM Int. Conf. Grid Computing, Sep. 2006, pp. 96–103.
[6] Anton Beloglazov,Jemal Abawajyb, Rajkumar Buyyaa "Energy- aware resource allocation heuristics for efficient management of data centers for Cloud computing". Year 2013.
[7] Eduardo Pinheiro, Wolf-Dietrich Weber and Luiz Andr´e Barroso "Failure trends in a large disk drive population" Appears in the Proceedings of the 5th USENIX Conference on File and Storage Technologies (FAST'07), February 2007.
[8] Feng Wang, Jie Qiu, Jie Yang, Bo Dong, Xinhui Li, Ying Li"Hadoop High Availability through Metadata Replication" *CloudDB'09*, November 2, 2009, Hong Kong, China.
[9] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The Hadoop Distributed File System," in Proc. IEEE 26th Symp. Mass Storage Systems and Technologies (MSST), Jun. 2010, pp. 1–10.
[10] Huangke Chen, Laurence T. Yang Ji Wang, Shu Yin, and Xiaocheng Liu "Real-Time Tasks Oriented Energy-Aware Scheduling in Virtualized Clouds", IEEE Transactions on Cloud Computing, vol. 2, no. 2, april-june 2014.
[11] Apache Hadoop Project. [Online]. Available: http://hadoop. apache.org.
[12] A. M. Soosai, A. Abdullah, M. Othman, R. Latip, M. N. Sulaiman, and H. Ibrahim, "Dynamic Replica Replacement Strategy in Data Grid," in Proc. 2012 8th Int. Conf. Computing Technology.

BIOGRAPHY

Reena S. More received her BE degree from Pune University. She is currently pursuing her master degree in computer science.

Nilesh V Alone is currently working as Assistant Professor in Department of Computer Engineering, R. H. Sapat College of Engineering, Management Studies & Research, and Nashik.He is having 13 years of experience in teaching. His area of interest is in Cloud Computing andMachine Learning.

Copyright to IJIRCCE