Scientific Journal Impact Factor (SJIF): 1.711

e-ISSN: 2349-9745
p-ISSN: 2393-8161

International Journal of Modern Trends in Engineering and Research
www.ijmter.com

Performance Enhancement of Cloud Computing using Clustering

Neha Pramod Patil, Prof. Prajkta Chapke
Computer science & engineering
H.V.P.M College of engineering
Amravati, India

Abstract-Cloud computing is an emerging infrastructure paradigm that allows efficient maintenance of cloud with efficient uses of servers. Virtualization is a key element in cloud environment as it provides distribution of computing resources. This distribution results in cost and energy reduction, thus making efficient utilization of physical resources. Thus resource sharing and use of virtualization allows improved performance for demanding scientific computing workloads. Number of data centers and physical servers are underutilized so they are used inefficiently. So performance evaluation and its enhancement in virtualized environment like public and private cloud are the challenging issues. Performance of cloud environment is dependent on CPU & memory utilization, Network and I/O disk operations. In order to improve the performance of the virtualization with cloud computing, one of the solutions is to allow highly available data in the cluster form. Thus replicas are available at each data centers and are highly available. In the proposed work, the I/O parameters are chosen for increasing the performance in this domain. This enhancement can be achieved through the clustering and caching technologies. The use of technology for data centers clustering is proposed in this paper. Thus performance and scalability can be improved by reducing the number of hits to the cloud database.

Keywords: Virtualization, cloud computing, clustering, performance enhancement, caching.

I.INTRODUCTION

For many computational applications, large numbers of resources are needed and this utilization of resources is for long period of time. Cloud computing, which is an emerging technology, provides the proper hosting of resources by leasing them from huge data centers only when they are needed. Cloud computing is replacing all existing technologies by offering their customer to pay only what they use. For example, an organization can buy any software or service for required period of time on the cloud rather than to purchase a machine for that purpose. It offers infrastructure, platform, software and data as services and these are subscription based services means pay-as-you-go model. These services are known as Infrastructure as a service (IaaS), platform as a service (PaaS), Software as a service (SaaS) and data as a service (DaaS) respectively. Infrastructure as a service ensures processing, storage, network and other fundamental computing resources to the users. Examples of IaaS based services are Amazon EC2, IBM`s Blue cloud, Eucalyptus, Rackspace Cloud etc. the platform as a service gives a high level integrated environment to build, test, deploy and host customer created applications. Examples of PaaS based services are Google App Engine, Engine Yard, Heroku etc. Software as a service is a software delivery model in that the applications are accessed by simple interface like web browser over Internet. Examples of SaaS based services are Web Mail, Google Docs, Facebook etc (G.Malathy et al). Data as a service provides an infrastructure for web scale data mining and knowledge discovery in order to empower the applications and services with intelligence.

Cloud computing models such as public, private,community and hybrid models can be implemented by using virtualization. Virtualization is the virtual evaluation of computing elements like hardware, software, memory, storage, network and so on (C. Pelletingeas, 2010). It allows the sharing of physical resources and higher utilization rate with optimal storage. It also reduces the power consumption and hardware investment and improves the system management without extra cost. Thus cloud is a package of services that offers infrastructure, platform, software and data as services. So many researches are being made for improving these flavors of services. But the dark side of using this virtualization is degradation of performance due to extra overhead. CPU usage, memory, storage and network are the performance factors for cloud computing. Since fast accessing of data and resources is highly demanded in cloud environment. Any organization adopting cloud computing certainly expect the kind of enhanced performance. But this performance is degraded due to limited bandwidth, high response time, inefficient CPU & memory utilization, scalability bottleneck and unnecessary use of data centers.

I/O virtualization poses a more difficult problem because I/O devices are shared among all virtual machines. It requires a privilege domain from guest VMs to access I/O. This intervention leads to longer I/O latency and higher CPU overhead due to context switches between the guest VMS and VMM (Virtual Machine Monitor). Performance of cloud computing is also dependent on the underlying cloud infrastructure. This work is aimed to address different issues that are responsible for improving the performance of cloud computing.

II. PROBLEMS PERCEIVED IN THIS AREA

It is difficult to match a good definition of cloud computing due to lack of standardization and economical impact. Standardizations of reliability and security are main concern now a days. Because these issues are facing daily new challenges. As cloud computing is an on demand service that shares a pool of resources over the network. Thus cloud security and reliability to its users are the major issues to be researched in this area and both of them make it hard to understand. In other hand, economical impact of cloud allows that resources could be used in more efficient and intelligent ways by reducing the cost. This can be achieved by virtualization because it requires less storage space of the servers and also reduction of power consumption. Latency and interoperability are also major issues to be solved in cloud computing because their causes are engineered into the cloud platforms themselves. One of the most important issues in cloud computing is the performance overhead. Since fast accessing of data and resources is highly demanded in cloud environment. Any organization adopting cloud computing certainly expect the kind of enhanced performance. But this performance is degraded due to limited bandwidth, high response time, inefficient CPU & memory utilization, scalability bottleneck and unnecessary use of data centers.

I/O virtualization poses a more difficult problem because I/O devices are shared among all virtual. It requires a privilege domain from guest VMs to access I/O. This intervention leads to longer I/O latency and higher CPU overhead due to context switches between the guest VMS and VMM (Virtual Machine Monitor). Performance of cloud computing is also dependent on the underlying cloud infrastructure.

III. AUTHORS PROPOSED METHODOLOGY

Performance is the major concern in the field of cloud computing. People are running into the scalability. But increasing number of Virtual machine and CPU is not the key solution for scalability, because cost is another issue to be researched then. So the area which is identified for improving the performance is to avoid the unnecessary use of databases. Data centers are highly loaded for accessing I/O requests. Although operations on huge amount of data in cloud computing are quite complex and lead to be performance overheads. There may be a case of server failure. So from the development, maintenance and performance perspective, all these can be serious issues to be refined. One of the solutions is to allow highly available data in the cluster form. Thus replicas are available at each data centers and are highly available. In our work, the I/O parameters are chosen for increasing the performance in this domain. This enhancement can be achieved through the clustering and caching technologies. Cluster based data centers are highly efficient for performing I/O operations. Clusters can be defined as collection of virtual data servers and are treated as a single machine. Clustering avoids uninterrupted access to data and also helps when network or storage connectivity is lost. Caching, in other hand, prevents over hits to the databases.

Network caching and VM image caching are the two aspects that are proposed for this purpose. Thus the idea behind both these technologies is to minimize the information that move among the different cloud components. Although clustering can be done at different level such as OS level, Application level, web server level and database level. Clustering among the data centers that are located throughout the world, allow highly available data for customers without any delay. Thus if one data center is goes down, everything in the second data center is clustered with the first, so there is no problem for the time being. And you still have database/web/app server in the second data center. In the favour of our work several issues are identified for better performance of cloud services. In some previous studies it is established that clustering can be a key contributing factor to improve the performance in cloud computing. G.Malathy et al proposed the Reservation Cluster approach for performance enhancement in cloud computing. The concept of reservation cluster is to schedule the tasks. Unscheduled tasks are sent to the reservation cluster and in this cluster all the tasks are scheduled simultaneously without any iteration. It reduces the amount of computation time and resource usage and allows better performance.

Clustering is one of the well-known Data mining techniques to find useful pattern from a data in a large database. These patterns are very useful for the knowledge workers such as financial analyst, Manger to take right managerial decisions. K-Means clustering is one of the most famous clustering algorithms applied in different types of domains such as Biology and Zoology, Medicine and Psychiatry, Sociology and Criminology, Geology, Geography and Remote sensing, Pattern recognition and Market research, and Education (Julie, 1982) to find the useful patterns. Today's business world is fast and dynamic in nature. It involves lot of data gathered from different sources. These data are stored in Data Warehouses. The most challenging task of the business people is to transform these data into useful information called knowledge. Data mining techniques are used to achieve this task. Cloud Computing offers several benefits to the business organization to cut the initial investments to establish infrastructure for storage and compute. Many business organizations have already started migration of their business data into cloud data centers. Most often they need to mine useful information from the data stored in the cloud data centers with regards to business decisions. So the main objective of this work is to incorporate and implement K-Means Data mining technique into Cloud environment.K-means algorithm: K-Means algorithm follows the partitional or nonhierarchical partitioning the given data set into specific number groups called Clusters.Each cluster is associated with a enter point called centroid. Each point is assigned to a cluster with the closest centroid. The main drawback of K-Means is the number of clusters must be known in advance, which is defined by K.

IV. ALGORITHM AND THEORY

Both our algorithm and that of are based on the online facility location algorithm of . For the facility location problem, the number of clusters is not part of the input (as it is for k-means), but rather a *facility cost* is given; an algorithm to solve this problem may have as many clusters as it desires in its output, simply by denoting some point as a facility. The solution cost is then the sum of the resulting k-means cost ("service cost") and the total paid for facilities. Our algorithm runs the online facility location algorithm of [24] with a small facility cost until we have more than ~ E e*(k* log *n)* facilities. It then increases the facility cost, re-evaluates the current facilities, and continues with the stream. This repeats until the entire stream is read. The details of the algorithm are given as Algorithm 1. The major differences between our algorithm and that of [9] are as follows. We ignore the overall service cost in determining when to end a phase and raise our facility cost *f.* Further, the number of facilities which must open to end a phase can be any ~ E e*(k* log *n),* the constants do not depend directly on the competitive ratio of online facility location . Finally, we omit the somewhat complicated end-of-phase analysis of, which used matching to guarantee that the' number of facilities decreased substantially with each phase and allowed bounding the number of phases by kl~gn' We observe that our number of phases will be bounded by logj3 *OPT;* while this is not technically bounded in terms of *n,* in practice this term should be smaller than the linear number Algorithm 1 Fast streaming k-means (data stream, *k,* ~, (3)

```
Initialize f = Ij(k(l +logn)) and an empty set K
while some portion of the stream remains unread do
while IKI :::; ~ = 8(k log n) and some portion of the stream is unread do
Read the next point x from the stream
Measure 6 = minYEK d(x, y)2
if probability fJ j f event occurs then
setK r- KU{x}
else
assign x to its closest facility in K
if stream not exhausted then
while IKI > ~ do
Set f r- {31
Move each x E K to the center-of-mass of its points
Let W x be the number of points assigned to x E K
Initialize K containing the first facility from K
for each x E K do
Measure fJ = minyEK d(x, y)2
if probability wx 6j f event occurs then
setKr-KU{x}
else
assign" x to its closest facility in K
SetK r- K
else
Run batch k-means algorithm on weighted points K
Perform ball k-means (as per [9]) on the resulting set of clusters.
```

We will give a theoretical analysis of our modified algorithm to obtain a constant approximation bound. Our constant is substantially smaller than those implicit in [9], with most of the loss occurring in the final non-streaming k-means algorithm to consolidate ~ means down to *k.* This requires as many as ~ distance computations; there are a number of results enabling fast computation of approximate nearest neighbors and applying these results will improve our running time. If we can assume that errors in nearest neighbor computation are independent from one point to the next (and that the expected result is good), our analysis from the previous section applies. Unfortunately, many of the algorithms construct a random data structure to store the facilities, then use this structure to resolve all queries; this type of approach implies that errors are *not* independent from one query to the next. Nonetheless we can obtain a constant approximation for sufficiently large choices of (3. For our empirical result, we will use a very simple approximate nearest-neighbor algorithm based on random projection. This has reasonable performance in expectation, but is not independent from one step to the next. While the theoretical results from this particular approach are not very strong, it works very well in our experiments. The Initial centroids will be chosen randomly. The centroid is nothing but the mean of the points in the cluster. Euclidean distance is used to measure the closeness. K-Means generates different clusters in different runs (Murat 2011).

The work was focused to implement and deploy K-Means algorithm in Google Cloud using Google App Engine with cloud SQL. Thus cloud computing allows mining of large databases and storing them with less cost. Results show that clustering works well in the cloud. Since cloud is having the large data sets and these data sets are regularly accessed as per user requirements. Michael Shindler have proposed fast and accurate k means algorithm as a solution where the data is too large and must be accessed sequentially. BigCross dataset and census 1990 dataset were considered as large data sets for applying the proposed clustering algorithm. Experiments show the significant results in terms of cost and response time. P. Ashok has proposed renovated k-means algorithm for Iris and Wine datasets. Proposed algorithm is compared with K-means, static weighted K-means and Dynamic weighted K-means on three different distance function. Less execution time and minimum iteration count of proposed k- means help to improve the performance. Karedla, R. has proposed the caching strategies to increase the system response time and data throughput of the disk. Results show that caching offers the twise performance of its size. Luo has proposed the Active Query Caching for improving scalability of database web servers. It allows the load sharing of database in order to reduce the network traffic. This caching is applied at query level for simplifying query containment checking and query evaluation at the proxy. This query level caching allows fast response to the user queries and hence improve the web server performance. Proposed architecture stack using clustering is shown in figure.

Proposed clustering and caching technologies are integrated at the physical level where the data servers are located. Thus clusters of data centers are formed at this level.

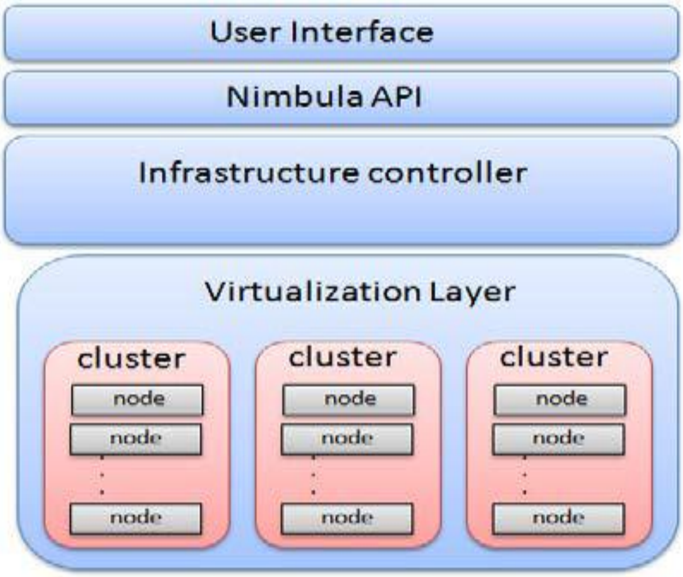

Fig1: Architecture for cloud computing using Clustering

And abstract level is responsible for load sharing among the datacenters. In order to scale the performance of data centers, a pool of data centers tied together to act as a single unit called cluster. These clustering technologies are transparent to client applications. Clustering and caching techniques can be seemed a better solutions for fast accessing of data and I/O operations. In this way, proposed methods for reducing data traffic and also minimizing the database hits can have better performance than cloud environments that are not using clustering of data centres.

V. CONCLUSION

Performance of virtualization with cloud computing is a major issue to be researched. Poor performance can lack the interest of customers. Clustering and caching are the proposed methodologies for improving the performance in this work. Both these technologies have their own significant. Clustering of data centers, network caching and VM image caching are the key points that are used as performance parameters. Expected results can have better performance than existing one.

REFERENCES

[1] A. Mahendiran, N. Saravanan, N. Venkata Subramanian and N. Sairam, "Implementation of K-Means Clustering in Cloud Computing Environment" , Research Journal of Applied Sciences, Engineering and Technology 4(10): 1391-1394, 2012 ISSN: 2040-7467.

[2] Wei Huang , Jiuxing Liu, Bulent Abali and Dhabaleswar K. Panda, "A Case for High Performance Computing with Virtual Machines", ICS '06 Proceedings of the 20th annual international conference on supercomputing, pages 125-134.

[3] C. Pelletingeas, "Performance Evaluation of Virtualization with Cloud Computing", MSc Advanced Networking, 2010,

[4] Alexandru Iosup, Simon Ostermann, M. Nezih Yigitbasi, "Performance Analysis of Cloud Computing Services for Many-Tasks Scientific Computing", IEEE Transaction on parallel and distributed system, VOL. 22, NO. 6, JUNE 2011.

[5] Mattias Sunding, Maximizing Virtual Machine Performance, vkernel corporation, A Quest Software Company, http://www.vkernel.com

[6] Devarshi Ghosal, R.Shane Canon and Lavanya Ramakrishnan,"I/O Performance of Virtualized Cloud Enviornment",

[7] Joyent White paper, Performance and scale in cloud computing,

[8] Nikolaus Huber, Marcel von Quast, Micahal Hauck and Samuel Konev, "Evaluating and Modeling Virtualization Performance Overhead for Cloud Enviornments"

[9] P. Ashok, Dr. G.M Kadhar Nawaz, E.Elayaraja and V. Vadivel, "Improved Performance of Unsupervised Method by Renovated K-Means "