IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
PP 21-24

Challenges and performance enhancement associated withvirtualized data:A case study

Mishra Jyoti Prakash[1], Mishra Anil Kumar[2], Mishra Sambit Kumar[3]
(Gandhi Institute for Education and Technology, Baniatangi
(Gandhi Institute for Education and Technology, Baniatangi
(Gandhi Institute for Education and Technology, Baniatangi

Abstract: *In the present days, the technology associated with Internet of Things(IoT) maintains the continuity towards providing internet connections as well as setting link among the physical and cyber space. In addition to increased volume, the IoT generates Big Data characterized by velocity in terms of time and location dependency, with a variety of multiple modalities and varying data quality. In general, IoT may be thought of as a combination of embedded technologies associated with wired and wireless communications, sensor and actuator devices, and the physical objects connected to the Internet. Its primary objective is to simplify and strengthen human activities as well as expertise. It needs data to represent better services to users as well as enhance the related performance. Since it focuses on basic concept associated with smart data, it may be a challenge towards processing data. It may also be considered as a set of connected devices to transfer data among one another in order to optimize the performance automatically. To prepare the processed data during the application as well as during communications may also be a challenge. In this article, it has been proposed to study the challenges associated with virtualized data as well as data in IoT.*
Key words : *Big Data, IoT, Virtualization, Sensor, Actuator, Data Center, Cluster*

I. Introduction

The primary intention of Internet of Things (IoT) is to design an economical compatible associated system. As a whole, it may be visualized as a set of connected devices to transfer data among one another in order to optimize their performance without human interventions.
Basically, it includes four main components, e.g. sensors, processing networks, analyzing data and monitoring the system. Accordingly, the communication protocols associated with this technology may be linked as device to device or device to server or deserver to server.

The required steps associated in this technology are first of all the data may be sent to data centers. After analyzing and processing the data, load balancing along with the time required to transfer the data to desired locations should be determined. So while the volume of data may be high, the CPU utilization in virtualized servers or cloud servers may be considered to achieve high efficiency. Basically, there are different cloud computing approaches, e.g. infrastructure as a service (IaaS), in which various equipments such as hardware, servers , and networks may be owned, platform as a service (PaaS), in which all the equipments may be put for rent on the internet and software as a service(SaaS), in which a distributed software model may be presented. As the sensors generate data on a repeated manner while processing high volume data, challenges may be faced. To overcome these challenges, the relevant data may be divided into packets and each packet may be assigned to different processing elements. Accordingly while linked to cloud environment, some improvements like decrease in network loading, increase in speed while processing data, less CPU usage and processing higher volume of data may occur.

II. Review of Literature

F. Chen et al.[1] in their work have focused on reviewing data mining knowledge and Techniques with suitable applications. They have reviewed some specific data mining functions . Practically they observed that the data generated by data mining applications sometimes may be similar to that of the IoT data.

C.-W. Tsai et al.[2] have concentrated to the challenges in preparing and processing data on the IoT through data mining techniques. Accordingly they have divided their work into three major area. Initially they have analyzed and explained the basic concepts of IoT, the data, and the challenges that exist in the relevant area, like mining algorithms for IoT.

R. Kapoor et al.[3] in their work have focused on packet slicing. It is observed that the increased number of smaller packets has almost no effect on the I/O performance due to the Interrupt Coalescing and Large Receive Offload mechanisms being adopted on Commercial off the shelf switches.

A. Ahmed et al. [4] in their work have focused on Virtual Data Centers with VMs as end-points. They proposed a resource management framework called Greenhead for embedding VDCs across geographically distributed Data Centers. Their approach had two phases. The first phase is to divide VDC requests into partitions. In the second phase, each partition is assigned to a Data Center based on electricity prices, power usage, the availability of renewable resources, and the carbon footprint.

M. Alicherry et al.[5] in their work have proposed a centralized resource allocation scheme for geo-distributed clouds to minimize the service delay among selected servers. The Two phase heuristic algorithm uses a sub-graph selection to divide the requested resources among the chosen servers.

T. I. Goiri et al.[6] have proposed a scheduling policy that models and manages a virtualized Data Center. It focuses on the allocation of VMs in Data Center nodes to optimize the provider's profit. In particular, it considers energy efficiency, virtualization overheads, and penalties for Service Level Agreement violations.

C. Lee et al.[7] in their work have analyzed that splitting the computation tasks reduces the constraints and variables required to allocate resources. It also allows the centralized controller to respond efficiently and promptly to any sudden fluctuations or to further requests, thereby reducing the computational complexity at the central controller. Vertical coordination between these two tasks also significantly improves efficiency, particularly in a large-scale environment.

K. M. Metwally et al.[8] in their work have proposes the procedure to evaluate the scalability with blocking condition. It occurs when IaaS requests are rejected. Thus, blocking ratio may be measured as the ratio between the number of rejected IaaS requests and the total number. They have performed two sets of experiments to evaluate the scalability of the decentralized approaches and their economic benefits against our previous centralized approach.

Xin et al. [9] in their work have proposed a Two-phase IaaS provisioning algorithm that uses minimum k-cut to split a VDC request into partitions before assigning them to different locations in order to balance the load. It relies on a centralized controller, which impacts the efficiency and the scalabilityof the solution.

M. Al-Fares et al.[10] in their work have discussed network topology. Each Data Center usually follows a FatTree network topology with an aggregated value of the Intra-Data Center network bandwidth according to different oversubscription values.

III. Experimental observations

In general case, projecting on the performance related to an application for large number of users involves predicting the maximum throughput achieved. So, factors affecting the maximum throughput may include both hardware and software resources of each of the servers associated with the application. In that case, the number of resources affecting the application performance may be too big. Accordingly, when the resource utilization reaches close to 100%, it minimizes the throughput. Further increase in the number of users beyond a certain point may also result in reduction in throughput.

III a. Assigning sequenced data and identifying similarities among data points

In general, after choosing a specific dataset, database, and variables, queries may be constructed.
For example,

```
query= exec(words,(select * from Dataset,DS1.db1 where datavar>=100));
output=fetch(query);
display data
```

Algorithm-1

Step 1 : The input space is partitioned into different regions. Each region may be associated with each data center.
Step2 : The data points associated with each data centre may be in a position to predict the data points associated with other data centers.
Step 3 : Input: sequenced and trained data set, D = f(xi; yi)
Step 4 : Apply fitness parameters to the nodes and trained data set
Step 5 : Perform the task scheduling

```
if the task is associated with the sequenced data,
  then  node.prediction := matched value with the data points in the relation, R
else
node.prediction := predicted data points associated with other data centers
```

Algorithm-2

Step 1 : Identify the data set S1 probably unlabeled with the related data values
Step 2 : Cluster the data set into different groups and measure the data points within the cluster

Step 3 : Identify each cluster centre point and match with near data points
Step 4 : Identify the similarities among data points within each cluster
Step 5 : If the data point xn is associated or linked or similar to cluster centre, then set binary indicator variable to 1.
Step 6 : Project the data points linear principal subspace
Step 7 : Obtain the desired normal set of projected data points

Initially it may be thought for quantifying the performance of virtualized data associated with virtual machines. The number of virtual machines including resource management factors, parameters associated with CPU scheduling may be considered for scalability. Usually, it increases the number of virtual machines until all available physical resources are used. The number of virtual machines sometimes may be increased beyond the amount of available resources and execution may be focused on different types of workloads. As being observed, the performance overhead for CPU virtualization may be below 5% due to the hardware support. But anyway, memory and network I/O virtualization overhead may be achieved up to 40%.

Generally, the data centers may be directly linked with the virtual machines, and by that achieve the better performance. However, the hardware specifications of the data centers like the memory , CPU speed , the number of cores inside each CPU ,the operating system, the number of virtual machines inside each data centers and the cost that for each virtual machine may be considerably measured to use the memory and CPU cores. As observed, the overall response time for the data centers and the cost for the virtual machines to serve the requests if the associated virtual machine server is set to closest data center. Approximately the response time is faster as compared to the processing time while conducting the simulation test over 100 virtual machines with minimum 1000 requests and 10 user bases.

IV. Challenges associated with virtualized data

It is obvious that, virtualized data associated with the data centers optimizes the performance and reduces operating costs. Also it increases resource availability and provide adequate cloud capacity for applications. But still then allocating resources in the virtualized data center may create load-sharing problems. As an example, the moments the multiple workloads are multiplexed using the VM hypervisor, IO streams at the same time compete for available resources. Again just like big data, virtualization works best when it includes everything and limiting the scope of the virtual infrastructure may include cost and complexity.

V. Discussion and future direction

It has been seen that relocation of virtual machine may be sometimes required for recovery purpose with better geographical coverage. But sometimes, it may be difficult to adopt a virtual machine replacement technique needed for both the cloud user and cloud provider. The methodologies associated with each virtual machine may be suitable under certain specific conditions/objectives. Accordingly parametric measures may be considered to evaluate the performance. Initially, it is required to monitor the amount of energy consumed by data center resources and the number of virtual machine migrations being occurred during the adaptation of the virtual machines.

VI. Conclusion

Virtualization is implemented either using hosted or hypervisor architecture. Basically, there are three primary virtualization technologies used , emulated or full virtualization, paravirtualization, and hardware supported virtualization. Hardware supported virtualization is still in early stage whereas the other technologies are well adopted. Accordingly, the full-virtualization software may be more appropriate for the multi tasking operating systems. Therefore, as the cost and the performance of cloud data centers become a practical concern, it has been observed a better effect towards the virtual machines to manage data centers. But somehow, the data security remains unresolved in cloud infrastructures. As the virtual machines require a secure connection between both source and target servers, the reliable protocols may be defined to establish and manage a protected communication.

References

[1]. F. Chen, P. Deng, J. Wan, D. Zhang, A. V. Vasilakos, X. Rong, Data mining for the internet of things: literature review and challenges, International Journal of Distributed Sensor Networks 2015 (2015) 12.
[2]. C.-W. Tsai, C.-F. Lai, M.-C. Chiang, L. T. Yang, Data mining for internet of things: a survey, IEEE Communications Surveys & Tutorials 16 (1) (2014) 77-97.
[3]. R. Kapoor, A. Snoeren, G. Voelker, and G. Porter. Bullet Trains: A study of NIC burst behavior at microsecond timescales. in Proceedings of CoNEXT, 2013.
[4]. A. Ahmed, Z. M. Faten, L. Rami, B. Raouf, and P. Guy, "Greenhead: Virtual data center embedding across distributed infrastructures," IEEE Transactions on Cloud Computing, vol. 1, no. 1, January-June 2013.

[5]. M. Alicherry and T. Lakshman, "Network aware resource allocation in distributed clouds," in INFOCOM, 2012 Proceedings IEEE, March 2012, pp. 963–971.
[6]. T. I. Goiri, "Energy-efficient and multifaceted resource management for profit-driven virtualized data centers," Future Generation Computer Systems, vol. 8, 2012.
[7]. C. Lee, P. Wang, and D. Niyato, "A real-time group auction system for efficient allocation of cloud internet applications," IEEE Transactions on Services Computing, vol. 8, no. 2, pp. 251–268, March 2015.
[8]. K. M. Metwally, A. Jarray, and A. Karmouch, "A Cost-Efficient QoS-Aware model for cloud IaaS resource allocation in large datacenters," in 4th IEEE International Conference on Cloud Networking, Niagara Falls, Canada, Oct. 2015.
[9]. Y. Xin, I. Baldine, A. Mandal, C. Heermann, J. Chase, and A. Yumerefendi, "Embedding virtual topologies in networked clouds," in Proceedings of the 6th International Conference on Future Internet Technologies, ser. CFI '11. New York, NY, USA: ACM, 2011, pp. 26–29. [Online]. Available: http://doi.acm.org/10.1145/2002396.2002403.
[10]. M. Al-Fares, A. Loukissas, and A. Vahdat, "A scalable, commodity data center network architecture," in Proceedings of the ACM SIGCOMM 2008 Conference on Data Communication, ser. SIGCOMM '08. ACM, 2008, pp. 63–74.