International Research Journal of Advanced Engineering and Science
ISSN: 2455-9024

Data Management and Virtulization: BigData

Kavitha. C[1], Kulkarni Varsha[2]
[1, 2]Department of Computer Science and Engineering, Sri Venkateshwara College of Engineering, Bengaluru, Karnataka, India

Abstract**—** *The objective of this paper is to handle complex data which is usually difficult to maintain in large scale industries. As all know data is an important term, which plays a major in all the fields without this data nothing can be performed. Data is a technical word used to refer information about a particular object based on which various operations are carried out. Recently, the data is increasing in large amounts that made difficult to analyse, manage and store. This huge information or data processing creates heavy burden on Data centers in terms of computation, storage and communication. Hence there is increase in the cost in many aspects like operations, services and electricity etc. to data center providers. Therefore, minimization of cost has become a major issue in these upcoming big data era. The objective of this project is to optimize these factors Big data placement, Task assignment and routing to minimize the overall computation and communication cost. In geo-distributed data centers joint optimization has been done on these three factors for big data services.*

Keywords**—** *Complex data, Big data managenent, big data placement, task assignment routing*.

I. INTRODUCTION

As of now the computation cost of the data centers has been increased due to increase in the services requested by the users. Via task placement the computation cost of the number of activated servers can be reduced by Data center Resizing (DCR) [3]. Based on DCR, research has been done and explored that the electricity cost was heterogeneously high at data centers located in different areas , Data Center Resizing was used to lower this cost [4]–[6].

Frameworks of Big data services [7] consists of a file system which is distributed at various data centers among which data chunks and their replicas are distributed and also placed at different places for good performance to access data in parallel and also for fine grained load balancing. To avoid remote load balancing and also to reduce the communication cost, the data locality should be improved where the inputs reside on the servers by placing different work on it [7], [8]. Due to some of the weakness mentioned below, even though the above solutions have some good results they are lacking in obtaining cost efficient Big data process.

First, data sites at different geographical area cause wastage of resources. For example, a server consists of large amount of computation resources among which few stay idle because of their less popularity. Even though data resources are not being used much but still the servers are active which increases the cost of operation.

Second, based on the transmission rates and unique features of particular links in a network, the cost varies [9] for example, among all data centers optical fiber will make communication easier. However, the different data link in a network may fail to make use of existing routing plan among data centers. Since the data of tasks reside on the servers where the task is performed, so many tasks cannot be assigned to the same server due to storage and computation capacity constraints. Downloading some data from remote servers cannot be avoided. Here the routing strategy plays important role on the cost of transmission. The transmission cost, a network uses many links and is proportional to the energy. If the usage of number of links is increased the transmission cost also increases. Therefore it is necessary to reduce the number of links usage and also should consider the transmission requirements.

Third, the existing solution of big data has not considered about the Quality of Service (QoS). Even though there are cloud services which have Service-Level-Agreement (SLA), the Big Data application also provide this SLA and is done between the service accepter and one who is providing service.

II. RELATED WORK

"*Power Struggles: Coordinated Multi-level Power Management for the Data Center*"

At Data center environments- power supply, high consumption electricity and management of heat are the challenges which should be taken care. This problem is evaluated based on different techniques for different factors at local as well as global levels and also on different software and hardware. There is no co-ordination among these solutions which leads interference with one another in an unpredictable dangerous ways. In this paper, the problem is addressed. Here two key contributions are made. First, power management solution is validated based on different approaches. This is illustrated using simulation based on nine different real world enterprises and 180 server traces and the advantages of correctness, stability and efficiency are explained. Second, in this paper by considering a unified architecture as a reference and analyzing quantitatively, conclusions are made on the impact of system design choices, architecture, implementations and workloads.

"*Minimizing Electricity Cost: Optimization of Distributed Internet Data Centers in a Multi- Electricity-Market Environment*"

The Cyber-Physical System (CPS) is combination of computation, networking and physical processes is an agile trend of research. Example of CPS is Internet Data Center (IDC) is one of the emerging systems. The increase in internet users has increased in demand of internet services thus the power usage also increases. Many researches are took place for the reduction of the power consumption of IDC and also in turn to reduce electricity cost of IDC. The price of electricity may depend on the time and location diversities and this will be a problem to service providers. In this paper the problem of electricity cost has been analyzed at different electricity

Kavitha. C and Kulkarni Varsha, "Data management and virtulization: BigData," *International Research Journal of Advanced Engineering and Science*, Volume 1, Issue 4, pp. 29-31, 2016.

International Research Journal of Advanced Engineering and Science
ISSN: 2455-9024

market environment by considering the Quality of Service with location and time diversity. Based on real life electricity price of data the internet locations may illustrate efficiency and the ability to produce the desired result.

III. Existing System

In the present systems, many researches are made to reduce the communication cost and computation cost. By Data center resizing (DCR) [3] and adjusting the number of activated servers the computation cost can be reduced. The computation capacity constraints are considered mainly while ignoring the transmission constraints by the cloud computing tasks generally. The main disadvantage here is computation capacity constraints are focused where as the transmission rate is ignored.

IV. Proposed Solution

The Data centers for different servers are located at different places across world. Here two data centers are considered and among those four servers are divided respectively. Now a file is uploaded using user interface this will be placed at any one of the data center generally. Here instead of that the file is divided in to chunks and those are equally distributed among all these four servers randomly.
Once all the chunks are assigned to particular servers that are nothing but data placement, the next is to assign tasks to the system with different factors. The results are obtained and to do further task the data virtualization is done. The time consumption for the computation and communication is reduced along with cost also minimized.

Advantages

The communication and operational cost has been reduced and stored data is managed and data is obtained as when is necessary for the task performance in the system. The computation speed is increased.

A. Architecture

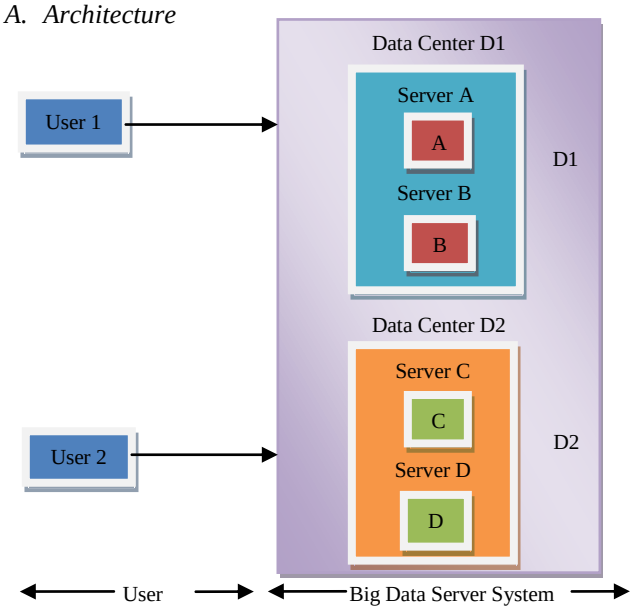

Fig. 1. User accessing data centers.

Here, n number of users can access the Servers which are placed at different data centers as shown in figure above. Big Data Server System consists of two Data centers D1 and D2 which in turn each data center have two servers A, B, C and D. This is the architecture of the system.

C. Algorithm: Two-Dimentional Markov Chain

Inputs: *Memory (mi), Data centre (D1, D2), Resource file (FR), Server S (S_1, S_2, S_3, S_4).*
Output: *Cost Minimization (Cj), Data Processing (Di).*
Procedure: *Create _Data center ( )*

{
D1 - A
D1 - B
D2 - C
D2 - D
C1 → D1, D2
}

Procedure: *Task assignment ( )*

{
Register_ User
Login _User
Upload f1 → D1, D2
Whether chunk K is placed on Server j
Yj.k = {1 if chunk k is placed on server j
0 Otherwise
Distribute chunks j€J
∑ Yj.k = P, (P1 .....Pn)
Xi = {1 if Server is activated
0 Otherwise
Condition servers ON status
}

Procedure: *Task Placement ( ) {*
Chunks ch1, ch2 ch3..... chn
Placed on different servers
Two-dimensional Markov Chain:
D1 → A
D1 → B
D2 → C
D2 → D
}

Procedure: *Data processing ( ) {*
Virtualize the Data Using Object
V 1 → {A, B, C, D}
User U1 → Request f1
Ui = f_1 ∩ {A, B, C, D}
User U1 ← Access the Data
}

END

D. User Module Description

1) First user makes registration by providing his details.
2) Users login by using username and password

Kavitha. C and Kulkarni Varsha, "Data management and virtulization: BigData," *International Research Journal of Advanced Engineering and Science*, Volume 1, Issue 4, pp. 29-31, 2016.

International Research Journal of Advanced Engineering and Science
ISSN: 2455-9024

3) After login he chooses the file and split that file in to number of chunks.
4) After splitting the files into chunks he view the chunks and details of the chunks
5) Next distribute the chunks to the servers
6) After distribution he assign the task details to be performed on chunks to the servers
7) After assigning the task he gets the response (result) from the servers.

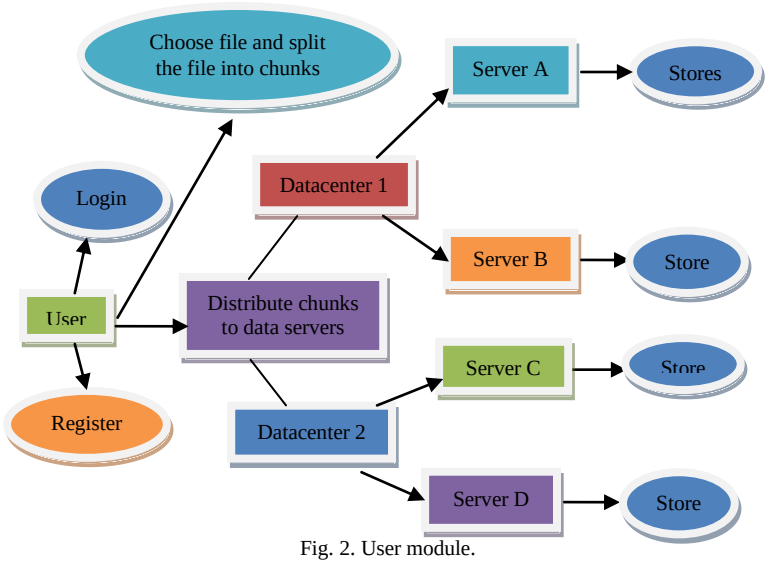

Fig. 2. User module.

E. Server Module Description

1) First server makes registration by providing his details.
2) Receive the chunks (i.e..what user has distributed the splitted chunks) from the user and store the chunks
3) Servers receive the task details from user and give response for the particular task what user has assigned.

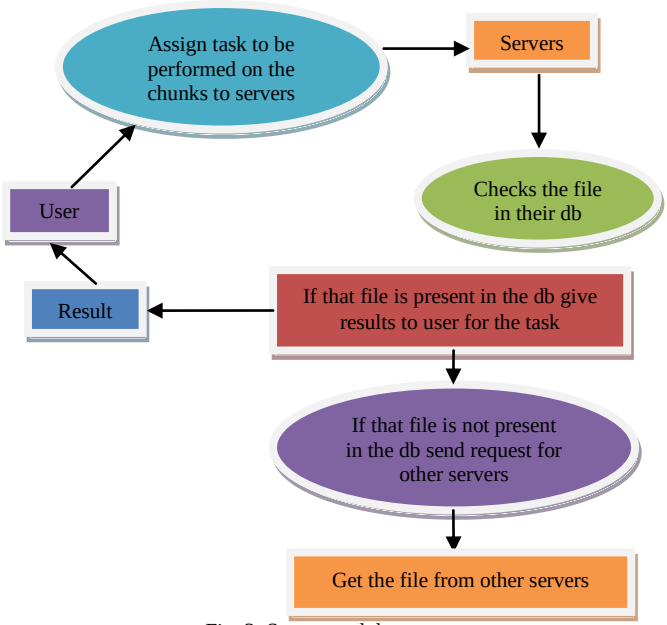

Fig. 3. Server module.

V. CONCLUSION

By studying the data placement, task assignment, data center resizing routing these factors play very important role in computation cost minimization. Two Dimensional Markov chain is used in this paper to reduce operational cost; thus increasing the performance, QoS and minimizing cost of operation.

REFERENCES

[1] "Data Center Locations," http://www.google.com/ about/data centers/inside/locations/index.html.
[2] R. Raghavendra, P. Ranganathan, V. Talwar, Z. Wang, and X. Zhu, "No "Power" struggles: Coordinated multi-level power management for the data center," in *Proceedings of the 13th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, ACM, pp. 48–59, 2008.
[3] L. Rao, X. Liu, L. Xie, and W. Liu, "Minimizing Electricity Cost: Optimization of Distributed Internet Data Centers in a Multi- Electricity-Market Environment," in *Proceedings of the 29th International Conference on Computer Communications (INFOCOM), IEEE*, pp. 1–9, 2010.
[4] Z. Liu, M. Lin, A. Wierman, S. H. Low, and L. L. Andrew, "Greening geographical load balancing," in *Proceedings of International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS)*, ACM, pp. 233–244, 2011.
[5] R. Urgaonkar, B. Urgaonkar, M. J. Neely, and A. Sivasubramaniam, "Optimal power cost management using stored energy in data centers," in *Proceedings of International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS)*, ACM, pp. 221–232, 2011.
[6] B. L. Hong Xu and Chen Feng, "Temperature aware workload management in geo-distributed datacenters," in *Proceedings of International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS)*, ACM, pp. 33–36, 2013.
[7] J. Dean and S. Ghemawat, "Map reduce: simplified data processing on large clusters," *Communications of the ACM*, vol. 51, no. 1, pp. 107–113, 2008.
[8] S. A. Yazd, S. Venkatesan, and N. Mittal, "Boosting energy efficiency with mirrored data block replication policy and energy scheduler," *SIGOPS Oper. Syst. Rev.*, vol. 47, no. 2, pp. 33–40, 2013.
[9] I. Marshall and C. Roadknight, "Linking cache performance to user behaviour," *Computer Networks and ISDN Systems*, vol. 30, no. 223, pp. 2123–2130, 1998.
[10] H. Jin, T. Cheocherngngarn, D. Levy, A. Smith, D. Pan, J. Liu, and N. Pissinou, "Joint host-network optimization for energy- efficient data center networking," in *Proceedings of the 27th International Symposium on Parallel Distributed Processing (IPDPS)*, pp. 623–634, 2013.

Kavitha. C and Kulkarni Varsha, "Data management and virtulization: BigData," *International Research Journal of Advanced Engineering and Science*, Volume 1, Issue 4, pp. 29-31, 2016.