Volume 4, Issue 11, November 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Survey on Big Data Processing in Geo Distributed Data Centers

Sarannia, N. Padmapriya, Asst prof
Department of Computer Science & Engg.,
IFET College of Engineering, Villupuram, India

Abstract—Big Data contains large-volume, complex and growing data sets with multiple, autonomous sources. Big data processing is the explosive growth of demands on computation, storage, and communication in data centers, which hence incurs considerable operational expenditure to data center providers. Therefore, to minimize the cost is one of the issue for the upcoming big data era. Using these three factors, i.e., task assignment, data placement and data routing, deeply influenced by the operational expenditure of geo distributed data centers. In this paper, we are ambitious to study the cost minimization problem via a joint optimization of these three factors for big data processing in geo-distributed data centers. Proposed using n-dimensional markov chain and procure average task completion time.

Keywords--big data , geo distributed data center, minimize cost

I. INTRODUCTION

Big data is an one of the emerging hot research topic because its mostly used in data center application in human society, such as government, climate, finance, and science. Currently, most research work on big data falls in data mining, machine learning, and data analysis[7].The name itself contains the meaning of data will be “so big” in large volume of both structured and unstructured data present. The challenges include capture, curation, storage, search, sharing, transfer, analysis and visualization. The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to "spot business trends, prevent diseases, combat crime and so on.

Big data is difficult to work with using most relational database management systems and desktop statistics and visualization packages, requiring instead "massively parallel software running on tens, hundreds, or even thousands of servers". What is considered "big data" varies depending on the capabilities of the organization managing the set, and on the capabilities of the applications that are traditionally used to process and analyze the data set in its domain. Big Data is a moving target; what is considered to be "Big" today will not be so years ahead. "For some organizations, facing hundreds of gigabytes of data for the first time may trigger a need to reconsider data management options. For others, it may take tens or hundreds of terabytes before data size becomes a significant consideration.fig number 1.1 shows the definition of the big data and the important characteristics of big data[21].

Fig 1 Big Data Definition

Every day, 2.5 quintillion bytes of data are created and 90 percent of the data in the world today were produced within the past two years [4].Another example is our capability for data generation has never been so powerful and enormous ever since the invention of the information technology in the early 19th century. Example for big data, on 4 October 2012, the first presidential debate between President Barack Obama and Governor Mitt Romney triggered more than 10 million tweets within 2 hours [5].

Among all these tweets, the specific moments that generated the most discussions actually revealed the public interests, such as the discussions about Medicare and vouchers. Such online discussions provide a new means to sense the public interests and generate feedback in real-time, and are mostly appealing compared to generic media, such as radio or TV broadcasting. The most fundamental challenge for Big Data applications is to explore the large volumes of

data and extract useful information or knowledge for future actions [8]. Currently, Big Data processing mainly depends on parallel programming models like MapReduce, as well as providing a cloud computing platform of Big Data services for the public. MapReduce is a batch-oriented parallel computing model.

There is still a certain gap in performance with relational databases. Improving the performance of MapReduce and enhancing the real-time nature of large-scale data processing have received a significant amount of attention, with MapReduce parallel programming being applied to many machine learning and data mining algorithms. Data mining algorithms usually need to scan through the training data for obtaining the statistics to solve or optimize model parameters.It calls for intensive computing to access the large-scale data frequently. To improve the efficiency of algorithms[20].Geo-distributed data centers are operated by many organizations such as Google and Amazon are the powerhouses behind many Internet-scale services.They are deployed across the Internet to provide better latency and redundancy.

These data centers run hundreds of or thousands of servers, so it consumes megawatts of power with massive carbon footprint, and also incur electricity bills of millions of dollars [3]. Data explosion is one of the rising demand for big data processing in recent years and modern data centers that are usually distributed at different geographic regions, e.g., in worldwide Google's 13 data centers over 8 countries in 4 continents [1].Gartner predicts that by in 2015, 71% of worldwide data centers will be spending hardware from the big data processing, which will surpass $126.2 billion. These are the reasons behind for we study the cost minimization problem for big data processing in geo distributed data centers.Cost of Big data processing have following issues. First, data locality may result in a waste of resources. For example, most computation resource of a server with less popular data may stay idle. Second, data center resizing.The low resource utility further causes more servers to be activated and hence higher operating cost.

II. LITRATURE SERVEY

Raghavendra.P,Z.Wang[15]Proposed the key challenges in data center environments are Power delivery, electricity consumption, and heat management. Propose using different power management strategy such as virtual machine controller and efficiency controller. Using these strategy to validate the power in data centers.***A.Sivasubramanian,B.Urgaonkar et al[12]***Proposed the Datacenter power consumption has one of the a significant impact on both its recurring electricity bill (Op-ex) and one-time construction costs (Cap-ex). They develop peak reduction algorithms that combine the UPS battery knob with existing throttling based techniques for minimizing power costs in datacenter .***Sharad Agarwal, John Dunagan et al[2]***Proposed the Now a days services grow to span more and more globally distributed datacenters, so we need urgent automated mechanisms to place application data across these datacenters. We present the Volley algorithm to a system that addresses these challenges***Jeffrey Dean and Sanjay Ghemawat et al[13]***Proposed the MapReduce is a programming model and its associated with implementation for processing and to generating large data sets. MapReduce runs on a large cluster of commodity machines and is highly scalable and its support to Programmers for the system easy to use.

Kuangyu Zheng, Xiaodong Wang et al[14]Proposed the Data center power optimization has recently received a great deal of research attention .Traffic consolidation has one to recently proposed to save energy for data center networks (DCNs). we propose PowerNetS, a power optimization strategy that leverages workload correlation analysis to jointly minimize the total power consumption of servers. ***Dan Xu Xin Liu , Bin Fan[17]*** The goal is to achieve an optimal tradeoff between energy efficiency and service performance over a set of distributed IDCs with dynamic demand. Dynamically adjusting server capacity and performing load shifting in different time scales. We propose three different loadshifting and joint capacity allocation schemes with different complexity and performance. Our schemes leverage both stochastic multiplexing gain and electricity-price diversity.

Zhenhua Liu, Minghong Lin[18] Energy expenditure has become a significant fraction of data center operating costs. Recently, "geographical load balancing" has been suggested to reduce energy cost by exploiting the electricity price differences across regions. However, this reduction of cost can paradoxically increase total energy use. This paper explores whether the geographical diversity of Internet-scale systems can additionally be used to provide environmental gains. Geographical load balancing can encourage use of "green" renewable energy and reduce use of "brown" fossil fuel energy. ***Hong Xu, Chen Feng, Baochun Li[19]*** For geo-distributed datacenters workload management approach that routes user requests to locations with cheaper and cleaner electricity to reduce the electricity cost.they using two factors for reducing the energy cost in datacenters. The factors are energy-gobbling cooling and location independent.Temperature diversity can be used to reduce the overall cooling energy overhead.

Disadvantages of Existing System:

- Used only 2 factors
- Existing system not support flexibility
- Data center resizing difficult

III. PROPOSED SYSTEM

Geo-Disributed Datacenter:

Geo distributed data center means many data centers are geo graphically distributed and connected through the WAN environment .In recently many organizations move to this geo distributed data center.Because they stored large or massive volume of datas.If they are using our own data center means only limited storage will be there so only many of them used this geo distributed data centers.

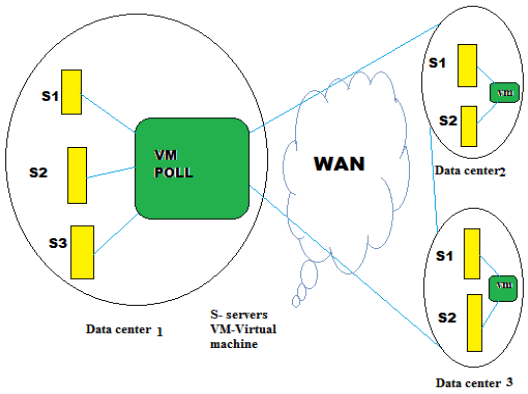

Fig Geo distributed datacenters

Google's and Amazon using this geo distributed data centers for storing and managing their data.

Markov chain:

A Markov is a mathematical system that undergoes transitions from one state to another on a state space. It is a random process usually characterized as memoryless: the next state depends only on the current state and not on the sequence of events that preceded it. memorylessness is called the Markov property. Markov chains have many applications as statistical models of real-world processes.A Markov chain is a sequence of random variables X_1, X_2, X_3, with the Markov property, namely that, given the present state, the future and past states are independent[16]. Formally, if both conditional probabilities are well defined, i.e. if

$$\Pr(X_1 = x_1, \ldots, X_n = x_n) > 0$$

The possible values of X_i form a countable set S called the **state space** of the chain.n-dimensional Markov chain works depends upon the processing , loading and task distribution.

Data Placement:

The data placement is another big issue in the geo distributed data centers.Because Where the datas are placed in the servers and how they can be accessed and calculate the latency time of that particular data transition and migrate user data to the closest datacenter.

However, the simple heuristic ignores two major sources of cost to datacenter operators: WAN bandwidth between data centers, and over-provisioning datacenter capacity to tolerate highly skewed datacenter utilization. In this paper, we show that a more sophisticated approach can both dramatically reduce these costs and still further reduce user latency[2]. Proposed using Volley algorithm for automated data placement in goe distributed data centers.once the data can be created means volley to analyze the migration data periodically.

Recursive Step:

wsm({fwi,xi}Ni=1=interp(wN ,wi; xn,wsm({fwi,xiN-1})i=1

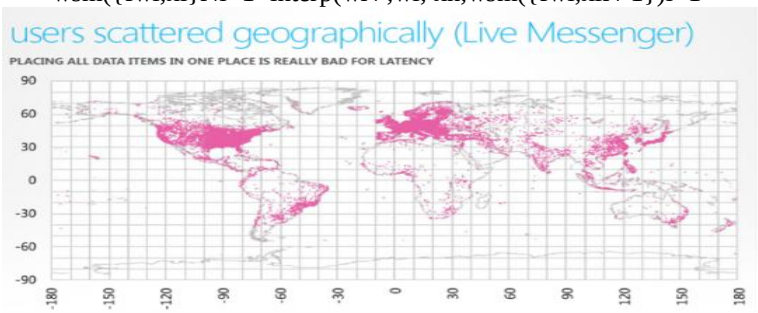

Fig 3

Volley algorithm contains three phases such as compute initial placement , iteratively move data reduce latency, iteratively collapse data to datacenters.

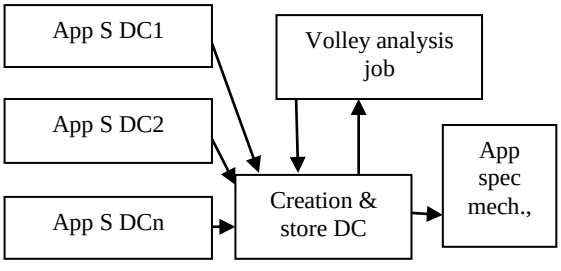

Fig 4 volley algorithm

Electricity cost:

The electricity cost another burden in geo-distributed data centers. Because more energy will be using in data centers . All the hardware's work without electricity. proposed a novel,data-centric algorithm used to reduce energy costs and with the guarantee of thermal-reliability of the servers in geo distributed data centers[19].And also using the n-dimensional markov chain algorithm to reduce the electricity cost.

Server cost:

In geo distributed data centers hundred's of servers used. Because of this automatically the server cost will be increases[10]. How to reduce the server cost means using communications and data placement and task assignment approach. Number of sever will be reduced means at a mean time the energy cost also decrease[11].

Server cost reduced using the joint optimization of these three factors such as task assignment , data placement and data routing via a n-dimensional markov chain. To efficiently manage the Datacenter resizeing.proposed the optimal workload and balancing of latency, electricity prices and the energy consumption.

IV. SYSTEM IMPLEMENTATION

This system output is a simulation based studies.It will be evaluated using NS2 tool.The evaluation base on the effect of the number of servers, on the effect of task arrival rate, on the effect of data size, on the effect of expected task completion delay, and on the effect of number replicas. Based on these we using joint optimization of task assignment ,data placement and data routing via a n-dimensional markov chain algorithm.To reduce the over all server, electricity and data placement cost in geo distributed data centers.

V. CONCLUSION

In this paper we study the geo distributed data centers issues. We jointly study the data placement , data center resizing and data routing to reduce the operational cost in geo distributed datacenters for big data processing.And we characterize the data processing and task assignment using n-dimensional markov chain based on joint optimization of that three factors.To minimize the cost of data center.

REFERENCES

[1] *"DataCenterLocations*,"http://www.google.com/about/datacentersinside/locations/index.html.

[2] S. Agarwal, J. Dunagan, N. Jain, S. Saroiu, A. Wolman, and H. Bhogan*,"Volley: Automated Data Placement for Geo-Distributed Cloud Services*," in The 7th USENIX Symposium on Networked Systems Design and Implementation (NSDI), 2010, pp. 17–32.

[3] GAO, P. X., CURTIS, A. R., WONG, B., ANDKESHAV, S. *It's not easy being green*.In Proc ACMSIGCOMM(2012).

[4] *"IBM What Is Big Data: Bring Big DataEnterprise*,"http://www01.ibm.com /software/data/bigdata/, IBM, 2012.

[5] *"Twitter Blog, Dispatch from the Denver Debate,"*http://blog.twitter.com/2/10/dispatch-from denvedebate.html,oct 2012.

[6] *Cost Minimization for Big Data Processing in Geo-Distributed Data Centers* Lin Gu, Student Member, IEEE, Deze Zeng, Member, IEEE, Peng Li, Member, IEEEand Song Guo, Senior Member,IEEEDOI10.1109/TETC.2014.2310456, IEEE Transactions on Emerging Topics in Computing2014.

[7] IEEE Network July/August 2014.

[8] A. Rajaraman and J. Ullman, *Mining of Massive Data Sets*.Cambridge Univ. Press, 2011

[9] INTEL INC. Reducing data center cost with an air reconomizer, August 2008.

[10] M. Sathiamoorthy, M. Asteris, D. Papailiopoulos, A. G. Dimakis,R. Vadali, S. Chen, and D. Borthakur, *"Xoring elephants: novel erasure codes for big data*," in Proceedings of the 39th international conference on Very Large Data Bases, ser. PVLDB'13. VLDB Endowment, 2013, pp. 325–336.

[11] B. Hu, N. Carvalho, L. Laera, and T. Matsutsuka, "*Towards big linked data: a large-scale, distributed semantic data storage,"* in Proceedings of the 14th International Conference on Information Integration and Web-based Applications & Services, ser. IIWAS '12. ACM, 2012, pp. 167–176.

[12] S. Govindan, A. Sivasubramaniam, and B. Urgaonkar, "*Benefits and Limitations of Tapping Into Stored Energy for Datacenters,"*in Proceedings of the 38th Annual International Symposium on Computer Architecture (ISCA). ACM, 2011, pp. 341–352.

[13] J. Dean and S. Ghemawat. MapReduce*: simplified data processing on large clusters.OSDI*, 2004.

[14] *Joint Power Optimization of Data Center Network and Servers with Correlation Analysis* Kuangyu Zheng, Xiaodong Wang, Li Li, and Xiaorui Wang The Ohio State University, USA*{zheng.722, wang.3570, li.2251, wang.3596}@osu.edu.*

[15] R. Raghavendra, P. Ranganathan, V. Talwar, Z. Wang, and X. Zhu*,"No "Power" Struggles: Coordinated Multi-level Power Management for the Data Center,"* 13th International Conference on (ASPLOS). ACM, 2008, pp. 48–59.

[16] http://en.wikipedia.org/wiki/Markov_chain

[17] 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 InternationalConference on Computer Communications (INFOCOM). IEEE,2010.

[18] 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 ComputerSystems (SIGMETRICS). ACM, 2011, pp. 233–244.

[19] B. L. Hong Xu, Chen Feng, *"Temperature Aware Workload Management in Geo-distributed Datacenters,"* in Proceedings of International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS). ACM, 2013, pp. 33–36.

[20] *"Data Mining with Big Data*"Xindong Wu, Fellow, IEEE, Xingquan Zhu, Senior Member, IEEE,Gong-Qing Wu, and Wei Ding, Senior Member,IEEE transactions on knowledge and data engineering, vol. 26, no. 1, january 2014.

[21] www.google.com/bigdata /images/definition of big data

