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 3, March 2015

Big Data Processing of Data Services in Geo Distributed Data Centers Using Cost Minimization Implementation

A. Dhineshkumar, M.Sakthivel

Final Year MCA Student, VelTech HighTech Engineering College, Chennai, India

Assistant Professor, Department of MCA, VelTech HighTech Engineering College, Chennai, India

ABSTRACT: The huge demands on big data processing imposes a heavy load on computation, storage, and communication in data centers, which hence incurs considerable operational expenditure to data center providers. Therefore, cost minimization has become an emergent issue for the upcoming big data era. Different from conventional cloud services, one of the main features of big data services is the tight coupling between data and computation as computation tasks can be conducted only when the corresponding data are available. As a result, three factors, i.e., task assignment, data placement, and data movement, deeply influence the operational expenditure of data centers. In this paper, we are motivated to study the cost minimization problem and optimization of these factors for big data services in geo-distributed data centers. To describe the task completion time with the consideration of both data transmission and computation.

KEYWORDS: big data , geo distributed data center, cost minimization, data placement.

I. INTRODUCTION:

Big data is an any collection of data sets so large and complex. It becomes difficult to process them using traditional data processing applications. Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, manage, and process data within a tolerable elapsed time. Big data is a set of techniques and technologies that require new forms of integration to uncover large hidden values from large data sets. First, the proposed Data center resizing (DCR) to reduce the computation cost by adjusting the number of activated servers via task placement. They distributes data in entire geographical data centers to lower the electricity cost. They distributes data bases on no of user and distribution of industries.

Second, the links in networks vary on the transmission rates and costs according to their unique features e.g.,the distances and physical optical fiber facilities between data centers. However, the existing routing strategy among data centers fails to exploit the link diversity of data center networks. Due to the storage and computation capacity constraints, not all tasks can be placed onto the same server, on which their corresponding data reside. Third, the Quality-of-Service (QoS) of big data tasks has not been considered in existing work. Similar to conventional cloud services, big data applications also exhibit Service-Level-Agreement (SLA) between a service provider and the requesters. Existing studies, on general cloud computing tasks mainly focus on the computation capacity constraints, while ignoring the constraints of transmission rate.
we are the first to consider the cost minimization problem of big data processing with joint consideration of data placement, task assignment and data routing. To describe the rate-constrained computation and transmission in big data processing. To reduce the communication cost, a few recent studies make efforts to improve data Cost Minimization for big data processing.

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 3, March 2015

II. RELATED WORK

Raghavendra.P,Z.Wang proposed the key challenges in data center environments are Power delivery, power 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 proposed the Data center 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 proposed the Nowadays services grow to span more and more globally distributed datacenters, so we need urgent automated mechanisms to place application data across these datacenters. 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 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, 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 load shifting 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, 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, 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. Cost minimization Data centre resizing (DCR) has been proposed to reduce the computation cost by adjusting the number of activated servers via task placement. To describe the rate-constrained computation and transmission in big data processing process, a two dimensional Markov chain and derive the expected task completion time in closed form has been proposed. To deal with the high computational complexity of solving MINLP, a mixed-integer linear programming (MILP) problem is linearized, which can be solved using commercial solver.DCR and task placement are
usually jointly considered to match the computing requirement

III. SYSTEM MODEL

Big Data Processing in Geo-Distributed Data Centers and communication resources Gartner predicts that by of worldwide data center hardware spending will come from the big data processing. 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 data. 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..Google's and Amazon using this geo distributed data centers for storing and managing their data. Recently Data explosion leads to a growing demand for big data processing in contemporary data centers that are usually distributed at different geographic regions in data centers. Big data analysis has shown its great potential in unearthing valuable insights of data to improve decision making, minimize risk and develop new products and services. Therefore, it is imperative to study the cost minimization problem for big data processing in geo-distributed data centers. Many efforts have been made to lower the computation Cost Minimization for Big Data Processing in Geo-Distributed Data Centers or communication cost of 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.

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 3, March 2015

Data center resizing has been proposed to reduce the computation cost by adjusting the number of activated servers via task Based on DCR, some studies have explored the geographical distribution nature of data centers and electricity price heterogeneity to lower the electricity Big data service frameworks comprise a distributed file system underneath, which distributes data chunks and their replicas across the data Cost Minimization for Big Data Processing in Geo-Distributed Data Centers for fine-grained load-balancing and high parallel data access performance.

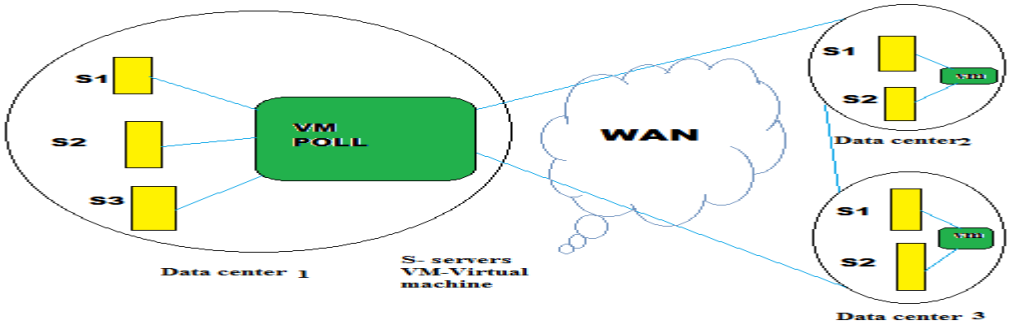

Figure1. Geo-Distributed Data Centers in WAN environment

IV. EXISTING SYSTEM

Big Data Processing in Geo-Distributed Data Centers locality by placing jobs on the servers where the input data reside to avoid remote data Although the above solutions have obtained some positive results, they are far from achieving the cost efficient big data processing because of the following weaknesses.

First, data locality may result in a waste of resources. For example, most computation resource Cost Minimization for Big Data Processing in Geo-Distributed Data Centers of a server with less popular data may stay idle. The low resource utility further causes more servers to be activated and hence higher operating cost. Second, the links in networks vary on the transmission rates and costs according to their unique the distances and physical optical fiber facilities between data centers.

However, the existing routing strategy among data centers fails to exploit the link diversity of data center networks. Due to the storage and computation Cost Minimization for Big Data Processing in Geo Distributed Data Centers capacity constraints, not all tasks can be placed onto the same server, on which their corresponding data reside. It is unavoidable that certain data must be Cost Minimization for Big Data Processing in Geo-Distributed Data Centers downloaded from a remote server. In this case, routing strategy matters on the transmission cost.

As indicated by Jin et al. the transmission cost energy, nearly proportional to the number of network link used. The more link used, the higher cost will be incurred. Therefore, it is essential to lower the number of links used while satisfying all the transmission requirements. Third, the Quality-of-Service of big data tasks has not been considered in existing work.

Drawbacks of Existing System:

- data locality may result in a waste of resources
- Does not support flexibility and computational task.
- Data center resizing difficulty in cloud environment

.

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 3, March 2015

V. PROPOSED SYSTEM

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].

Power cost minimization: The power 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 minimization : 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 dimensional markov chain. To efficiently manage the Data center resizeing.proposed the optimal workload and balancing of latency, electricity prices and the energy consumption. key issue in Large-scale data centers is electricity cost and operating cost Therefore, reducing the electricity cost has received significant attention from both academia and industry.so reduce electricity cost by routing user requests to geo-distributed data centers with accordingly updated sizes that match the requests using a holistic approach of workload balancing.

Big data management: The main key issue in big data management is reliable and effective data placement. To achieve this goal they propose a scheduling algorithm, which takes into account energy efficiency in addition to fairness and data locality properties. and new design philosophy providing a new agile and deep data analytics for one of the world's largest networks at Fox Audience Network, using the Green plum parallel database system.

VI. COST MINIMIZATION USING MAPREDUCE ALGORITHM

Numerous algorithms were defined earlier in the analysis of large data set. Will go through the different work done to handle Big Data. In the beginning different algorithm was used earlier to analyze the big data. In work done by Hall. et al. there is defined an approach for forming the rules of the large set of training data. The approach is to have a single decision system generated from a large and independent n subset of data. Here we use cost minimization using map reduce algorithm as follows ,

Cost Minimization using MapReduce Algorithms. Denote by S the set of input objects for the underlying problem. Let n, the problem cardinality, be the number of objects in S, and t be the number of machines used in the system. Define m = n/t, namely, m is the number of objects per machine when S is evenly distributed across the machines. Consider an algorithm for solving a problem on S.

We say that the algorithm is minimal cost if it has all of the following properties.

• Minimum footprint: at all times, each machine uses only O(m) space of storage.

• Bounded net-traffic: in each round, every machine sends and receives at most O(m) words of information over the network.
Constant round: the algorithm must terminate after aconstant number of rounds.

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 3, March 2015

• Optimal computation: every machine performs only O(Tseq/t) amount of computation in total (i.e., summing over all rounds), where Tseq is the time needed to solve the same problem on a single sequential machine. Namely, the algorithm should achieve a speedup of t by using t machines in parallel.

Each machine M has at most 2 groups remaining, i.e., with keys kmin (M) and kmax (M), respectively. Hence, there are at most 2t such groups on all machines. To handle them, we ask each machine to send at most 4 values toM1 (i.e., to just a single machine). The following elaborates how:

Map-shuffle (on each Mi, $1 \leq i \leq t$):

Step 1. Obtain the total weight Wmin(Mi) of group kmin(Mi), i.e., by considering only objects in Mi.
Step 2. Send pair (kmin(Mi),Wmin(Mi)) toM1.
Step 3. If kmin (Mi) 6= kmax (Mi), send pair (kmax (Mi),Wmax (Mi)) to M1, where the definition of kmax (Mi) is similar to kmin (Mi).

Reduce (only on M1):

Let (k1,w1), ..., (kx,wx) be the pairs received in the previous phase where x is some value between t and 2t. For each group whose key k is in one of the x pairs, output its final aggregate Pj|kj=k wj . The minimality of our group-by algorithm is easy to verify. It suffices to point out that the reduce phase of the last round takes O(t log t) = O(nt log n) time (since $t \leq m = n/t$).

VI. CONCLUSION

In this paper we study the geo distributed data centres issues. We jointly study the data placement, data centre resizing and data routing to reduce the operational cost in geo distributed data centres for big data processing..To minimize the cost of data centre. We jointly study the data placement, task assignment, data centre resizing and routing to minimize the overall operational cost in large-scale geo-distributed data centres for big data applications. For example:-most computation resource of a server with less popular data may stay idle. The low resource utility further causes more servers to be activated and hence higher operating cost.

REFERENCES

[1] 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.

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7].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.

[8] 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.

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

[10].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.

[11] 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.