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. 4, Issue 5, May 2016

Minimizing Cost of Big Data Processing using Data Centers

Kanchan Dendge[1], Pratiksha Kamble[2], Aishwarya Nerkar[3], Shweta Lakal[4,]Prof.Umesh Talware[5]

Students, Department of Information Technology, Dhole Patil College of Engineering, Pune, Maharashtra, India[1,2,3,4]

Professor, Department of Information Technology, Dhole Patil College of Engineering, Pune, India[5]

ABSTRACT:ABig data is a popular term used to describe exponential growth and availability of data both structured and unstructured. The explosive growth of demands on big data processing imposes a heavy burden on computation,storage, and communication in data centers, while hence incurs considerable operational expenditure to data center providers. Therefore cost minimization has become an emergent issue for upcoming big data era. Different from convention cloud services, one of the main features of big data services is the tight coupling between data and computation tasks can be conducted only when the corresponding data is available. As a result, the factors, i.e., task assignment, Data placement and data movement, deeply influence the operational expenditure of data centers.

KEYWORDS:Authorizationbig data, data flow, data placement, data centres, cost minimization.

I. INTRODUCTION

Cloud Data explosion in recent years leads to a raising demand for big data processing in modern data centers that usually distributed at different geographic regions Many efforts have been made to lower the computation or communication cost of data centers. Data center resizing has been proposed to reduce the computation cost by adjusting the number of activated servers via task placement. Based on DCR, some studies have explored the geographical distribution nature of data centers and electricity price heterogeneity to lower the electricity cost. Big data service frameworks eg.,, comprise a distributed file system underneath which distributes data chunks and their replicas across the data centers for fine-grained load-balancing and igh parallel access performance. To reduce the communication cost,a few recent studies make efforts to improve data locality by placing jobs on the servers where the input data reside to avoid remote data loading.

Although the above solutions have obtained some positive results, they are far from achieving the cost efficient bid data processing because of the following weakness First, data locality may result in a waste of resources. For example, most computation resource of 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 features, eg., 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. It is unavoidable that certain data must be downloaded from a remote server. In this case, routing strategy matters on transmission cost.

Third, the QoS of big data tasks has not beem considered in existing work. Similar to conventational cloud services, big data applications also exhibit Service-level-Agreement between a service provier and the requestors.

To conquer above weaknesses, we study cost minimization problem for big data processing via join optimization of task assignment, data placement, and routing in data centers. Specifically, we consider the following issues in our joint optimization. Servers are equipped with limited storage and computation resources. Each data chunk has a storage requirement and will be required by big data tasks. The data placement and task assignment are transparent to the data users with guaranteed QoS.

Our objective is to optimize the big data placement, task assignment, routing and DCR such that the overall computation and communication cost is minimized. To describe the rate-constrained computation and transmission in big data processing process, we propose a two dimensional Markov chain and derive the expected task completion time in closed for. To deal with the high computational complexity of solving MINLP, we linearize it as a mixed-integer

DOI: 10.15680/IJIRCCE.2016. 0405142

linear programming (MILP) problem, which can be solved using commercial solver. Through extensive numerical studies, we show the high efficiency of our proposed joint-optimization based algorithm.

II. LITERATURE SURVEY

Author, Xiaobo Fan, Wolf-Deitrich, Weber Luiz Andre Barrosoint their paper "Power provisioning for a warehouse-sized computer[1]" presented the aggregate power usage characteristics of large collections of server for different classes of applications over a period of approximately six months. Their observation allow us to evaluate opportunities for maximizing the use of the deployed power capacity of datacenters, and access the risks of over-subscribing it.

H.Shachnai, G.Tamir, and T.Tamir in their paper "Minimal cost recognition of data placement in storage area network[2]" showed the dynamic placement of web applications.

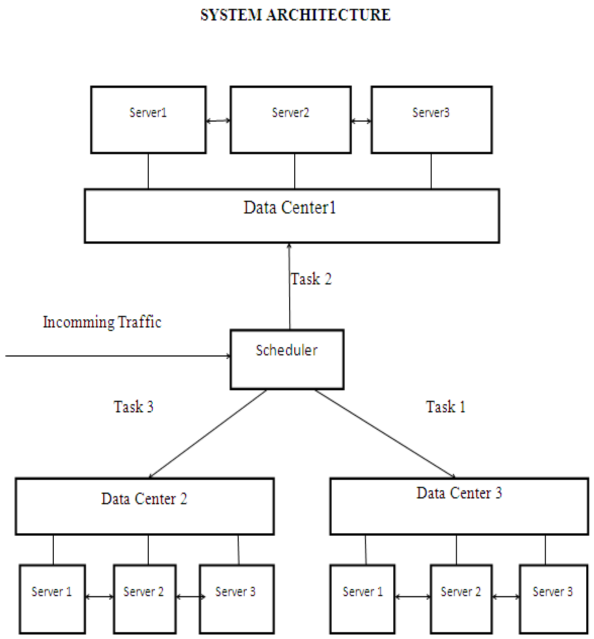
B.Hu, N. Carvalho, L.Laera suggested a mechanism allowing LOD to take advantage of existing large-scale data stores while suggesting its semantic nature in paper "Towards big linked data: a large-scale,distributed semantic data storage"[3]

H.Jin, T. Cheochemgngam, D. Levy, A. Smith, D. Pan, J.Liu proposed a joint optimization scheme that simultaneously optimizes virtual machine placement and network flow routing to maximize energy savings in their paper "Joint Host-Network Optimization for Energy-Efficient Data Center Networking"[6].

III. PROPOSED SYSTEM

To describe the rate-constrained computation and transmission in big data processing process, we propose a two dimensional Markov chain and derive the expected task completion time in closed form.

System Architecture



Modules

A.Data Uploading

Select the big data and stored into the hadoop environment for performing map reduce on hadoop. The data should be loaded into the VM server location. After Uploading the file the data segmentation is performed for further process.

B.Segmentation

Packet segmentation improves network performance by splitting the packets in received Ethernet frames into separate buffers. Packet segmentation may be responsible for splitting one into multiple so that reliable transmission of each one can be performed individually. Segmentation may be required when the data packet is larger than the maximum transmission unit supported by the network.

The packet processing system is specifically designed for dealing with the network traffic. Most networks, such as the Internet, are distributed and layered systems composed of hosts, workstations, switches and routers etc. The processing speed of edge equipment falls behind those in core network. Finally the access network connect s the terminals of a customer endpoint. And usually the bandwidth and line rate requirement is lowest among the three.

The packet processing system can be equipped in any layer of the network, either in the high end core routers or in the LAN switches. The flexibility of the system comes from the programmable elements within it, i.e. NPs. And a series of stacked network protocols guarantee its capability to achieve the performance specification.

C.Task Assignment

The Data Center should be selected according to computation and storage capacity of servers reside in the data center. Identification of Data Center is important matter for minimizing operational expenditure of servers reside in the each data centers. Data chunks can be placed in the same data center when more servers are provided in each data center. Further increasing the number of servers will not affect the distributions of tasks. Task should be assigned to data center where number of activated servers are optimal. Task assignment is deeply influence the operational expenditure of data center. Task is assigned to data center according to nearest data center for effectively processing of data. Each data chunk has a storage requirement and will be required by big data tasks.

D.Data Loading

A Data Placement on the servers and the amount of load capacity assigned to each file copy so as to minimize the communication cost while ensuring the user experience. Cloud services make use of Volley by submitting logs of datacenter requests. Volley analyzes the logs using an iterative optimization algorithm based on data access patterns and client locations, and outputs migration recommendations back to the cloud service. Invent Min Copy sets, a data replication placement scheme that decouples data distribution and replication to improve the data durability properties in distributed data centers. Recently, Jin et propose a joint optimization scheme that simultaneously optimizes virtual machine (VM) placement and network flow routing to maximize energy savings.

E. Processing of Task

The high computational server should not processing the low population of data chunk. Because it increase the operational expenditure of server, wastage of storage and transmission cost. The population of data is processed depend upon the computational capacity of servers reside in the data centres.

F. Evaluation process

We present the performance results of our joint-optimization algorithm using the MILP formulation. Evaluate server cost, communication cost and overall cost under different total server numbers.

IV. Conclusion

In this paper, 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.

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, 2008, pp. 48–59.
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,2010, pp. 1–9.

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. 4, Issue 5, May 2016

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, 2011, pp. 233–244.
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, 2011, pp. 221–232.
6. B. L. Hong Xu, Chen Feng, “Temperature Aware Workload
7. Management in Geo-distributed Datacenters,” in Proceedings of International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS). ACM, 2013, pp. 33–36.
8. J. Dean and S. Ghemawat, “Mapreduce: simplified data processing on large clusters,” Communications of the ACM, vol. 51, no. 1, pp. 107–113, 2008.
9. 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.
10. I. Marshall and C. Roadknight, “Linking cache performance to user behaviour,” Computer Networks and ISDN Systems, vol. 30, no. 223, pp. 2123 – 2130, 1998.
11. 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), 2013, pp. 623–634.
12. A. Qureshi, R. Weber, H. Balakrishnan, J. Guttag, and B. Maggs,“Cutting the Electric Bill for Internet-scale Systems,” in Proceedings of the ACM Special Interest Group on Data Communication (SIGCOMM). ACM, 2009, pp. 123–134.
13. X. Fan, W.-D. Weber, and L. A. Barroso, “Power Provisioning forA Warehouse-sized Computer,” in Proceedings of the 34th Annual International Symposium on Computer Architecture (ISCA). ACM,2007, pp. 13–23.
14. 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
P. X. Gao, A. R. Curtis, B. Wong, and S. Keshav, “It’s Not EasyBeing Green,” in Proceedings of the ACM Special Interest Group of Data Communication (SIGCOMM). ACM, 2012, pp. 211–222.Z. Liu, Y. Chen, C. Bash, A. Wierman, D. Gmach, Z. Wang,
15. M. Marwah, and C. Hyser, “Renewable and Cooling Aware Workload Management for Sustainable Data Centers,” in Proceedings of International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS). ACM, 2012, pp. 175–186.
16. M. Sathiamoorthy, M. Asteris, D. Papailiopoulos, A. G. Dimakis,R. Vadali, S. Chen, and D. Borthakur, “Xoring elephants: novelerasure codes for big data,” in Proceedings of the 39th internationalconference on Very Large Data Bases, ser. PVLDB’13. VLDBEndowment, 2013, pp. 325–336.
17. B. Hu, N. Carvalho, L. Laera, and T. Matsutsuka, “Towards biglinked data: a large-scale, distributed semantic data storage,”
18. in Proceedings of the 14th International Conference on InformationIntegration and Web-based Applications & Services, ser. IIWAS ’12.ACM, 2012, pp. 167–176.
19. J. Cohen, B. Dolan, M. Dunlap, J. M. Hellerstein, and C. Welton, “Mad skills: new analysis practices for big data,” Proc. VLDBEndow., vol. 2, no. 2, pp. 1481–1492, 2009.
20. R. Kaushik and K. Nahrstedt, “T*: A data-centric cooling energycosts reduction approach for Big Data analytics cloud,” in 2012 International Conference for High Performance Computing, Networking,Storage and Analysis (SC), 2012, pp. 1–11
21. .F. Chen, M. Kodialam, and T. V. Lakshman, “Joint scheduling ofprocessing and shuffle phases in mapreduce systems,” in Proceedingsof the 29th International Conference on Computer Communications (INFOCOM). IEEE, 2012, pp. 1143–1151.
22. H. Shachnai, G. Tamir, and T. Tamir, “Minimal cost reconfiguration of data placement in a storage area network,” TheoreticalComputer Science, vol. 460, pp. 42–53, 2012.
23. 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 NetworkedSystems Design and Implementation (NSDI), 2010, pp. 17–32

BIOGRAPHY

Ms.Kanchan Dendge pursuing her Degree Course in Bachelor of Engineering from Dhole Patil College of Engineering, Pune, Maharashtra, India

Ms.Pratiksha Kamble, pursuing her Degree Course in Bachelor of Engineering from Dhole Patil College of Engineering, Pune, Maharashtra, India

Ms.Aishwarya Nerkar pursuing her Degree Course in Bachelor of Engineering from Dhole Patil College of Engineering, Pune, Maharashtra, India

Ms.Shweta Lakal pursuing her Degree Course in Bachelor of Engineering from Dhole Patil College of Engineering, Pune, Maharashtra, India

Prof.Umesh Talware is a Professor in Department of Information Technology in Dhole Patil College of Engineering, Pune, Maharashtra, India.

DOI: 10.15680/IJIRCCE.2016. 0405142