ISSN NO. 2320-5407

Journal homepage: http://www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH

RESEARCH ARTICLE

CURTAIL THE EXPENDITURE OF BIG DATA PROCESSING USING MIXED INTEGER NON-LINEAR PROGRAMMING

R.Kohila[1], N.Sivaranjani[2]

1. Assistant professor, Department of Computer science and Engineering, V.S.B Engineering College, Karur.
2. Assistant Professor, Department of Information Technology, V.S.B Engineering College, Karur.

Manuscript Info

Manuscript History:

Received: 15 October 2015
Final Accepted: 22 November 2015
Published Online: December 2015

Key words:

Corresponding Author

R.Kohila

Abstract

Big data is a popular term used to describe the 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, 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 is 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 via a joint optimization of these three 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, we propose a two-dimensional Markov chain and derive the average task completion time in closed-form. Furthermore, we model the problem as a mixed-integer non-linear programming (MINLP) and propose an efficient solution to linearize it.

INTRODUCTION

Data explosion in recent year leads to rising demand for big data processing. It took 15 years for the internet to grow to 50 million users in 1998. Eleven years later, in 2009, there were 1 billion internet users around the world. Three years later, it doubled to over 2.1 billion users, and by 2013, 39% of the world's population was using the internet (2.7 billion people). E.g., Amazon works with 46,000 servers [1].

Data center re sizing (DCR) has been proposed to reduce the computation cost by adjusting the number of activated servers via task placement [2]. Based on DCR, some studies have explored the geographical distribution nature of data centers and electricity price heterogeneity to lower the electricity cost [3]. To reduce the communication expenditure, a few current studies make hard work to get better data locality by placing jobs on the servers where the input data reside to avoid remote data loading [4].

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 of a server with less popular data may stay at rest. The low resource efficacy further causes more servers to be activated and hence higher operating cost.

Second, the link in networks varies on the transmission rate and costs according to their unique features, e.g., the distance and physical optical fiber facilitates between data centers. 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 task has not considered in existing work. Similar to conservation cloud service, big data applications also exhibit Service Level Agreement (SLA) between a service provider and requestor. To observe SLA, a certain level of QoS, usually in terms of task completion time, shall be guaranteed.

The QoS of any cloud computing tasks is first determined by where they are placed and how many computation resources are allocated. Besides, the transmission rate is another influential factor since big data tasks are data-centric and computation task cannot proceed until the corresponding data are available.

RELATED WORKS

MINIMIZATION OF ENERGY EXPENDITURE

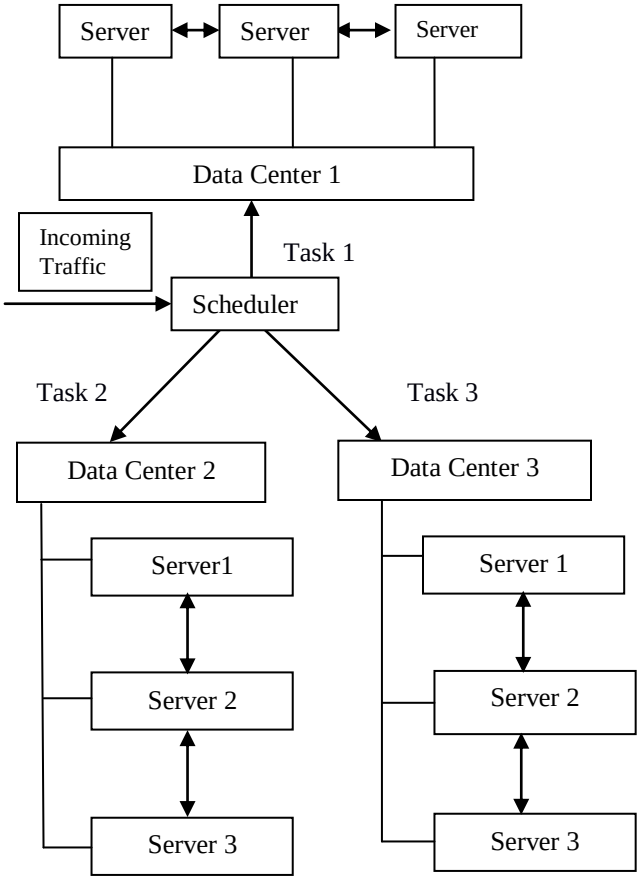
Energy expenditure is becoming an increasing important fraction of data center operating costs. At the same time the energy expenses per unit of computation can vary significantly between two different locations. A large data centers may requires many mega watts of electricity enough to power thousands of houses. A server becomes more energy efficient with various energy saving technique the data centers network has been accounting for 20% or more of the energy consumed by the entire data centers. Although there exists number of optimization solution for DCNs they consider only either host or network but both. Joint optimization scheme that simultaneously optimizes virtual machine placement and network flow routing to maximize energy saving.

Similarly among various mechanisms that have been proposed so far for data center energy management, the techniques that attract lots of attraction are task placement and DCR. These are usually jointly considered to match the corresponding requirement. Liu et al [5] re-examine the same problem by network delay into consideration. Rao et al.. [2] Investigated how to reduce electricity cost by using routing user requests to geo-distributed data centers with according updated sizes that match the requests.

MANAGING BIG DATA

To undertake the challenges of successfully managing big data, many decisions have been proposed to recover the storage and processing cost. The advantage in managing big data is reliable and efficient data placement. The use of flexibility in the data placement policy to boost energy efficiency in data centers and propose a scheduling algorithm. In addition, allocation of computer resources to task has also strained much concentration. Cohen et al. [6] developed new design attitude, techniques and knowledge providing a new magnetic, agile and deep data analytics for one of the world's major marketing networks at Fox Audience Network, by using Green plum parallel database system.

SYSTEM ARCHITECTURE



SYSTEM MODEL

NETWORK MODEL

Geo distributed data center topology is considered in which all the available servers of the same data center (DC) are related to their local switch, while the different data centers will communicate through switches. A set I data centers and each data center i€I consists of a set of servers Ji that are connected to a switch mi€M with a local transmission cost of CL. The entire system can be represented as a directed graph G=(N,E). The vertex set N=MUJ includes the set of all the servers and E is the directional edge sets. An server are connected to their local switch while the switches are connected via inter-data center link W(u,v) can be defined as

W (u,v)={CR if u, v€M, CL otherwise}.

TASK MODEL

We consider big data tasks targeting on data targeting on data stored in a distributed file system that is built on geo- distributed data centers. The data are divided into a set K of chunks. Each chunk k€K has the size of ϕk(ϕk≤1), which is normalized to the server storage capacity. P-way replica [7] is used in our model. That is, for each chunk, there are exactly P copies stored in the distributed file system for saliency and fault-tolerance.

It has been widely agreed that the tasks arrival at data centers during a time period can be viewed as a Poisson process, [8]. In particular, let λk be the average task arrival rate requesting chunk k. since these tasks will be distributed to a server with fixed probability, the task arrival in each server can be also regarded as a Poisson process. We denote the average arrival rate of task for chunk k on server j as $\lambda jk(\lambda j \leq 1)$. When a task is distributed to a server where its requested data chunks does not reside, it needs to wait for the data chunk to be transferred. Each task should be responded in time D.

Moreover, in practical data center management, many task predication mechanisms based on the historical ststistics have been developed and applied to the decision making in data centers [7]. To keep the data center settings up-to-date, data center operators may make adjustment according to the task prediction period by period [2][5]. This approach is also adopted in this paper.

PERFORMANCE

Select the big data and stored into the hadoop environment for the 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.

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. The packet processing system is specifically designed for dealing with the network traffic. Segmentation may be required when the data packet is larger than the maximum transmission unit supported by the network.

The Data Center should be selected according to computation and storage capacities 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 is assigned to data center according to Memory requirement for effectively processing of data.

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. Joint optimization scheme that simultaneously optimizes virtual machine (VM) placement and network flow routing to maximize energy savings.

The high computational server should not process the low population of data chunk. Because it increases 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 centers.

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.

FUTURE ENHANCEMENT

We include the fault tolerance in cost minimization of big data processing to avoid the data loss. Nodes share their load with alternative access point. A fault tolerance is a setup or configuration that prevents a computer or network device from failing in the event of an unexpected problem.

CONCLUSION

In this paper, we jointly study the data placement, task assignment, data center resizing and routing to minimize the overall operational cost in large-scale geo-distributed data centers for big data applications. We first characterize the data processing process using a two-dimensional Markov chain and derive the expected completion time in closed-form, based on which the joint optimization is formulated as an MINLP problem. To tackle the high computational complexity of solving our MINLP, we linearize it into an MILP problem. Through extensive experiments, we show that our joint-optimization solution has substantial advantage over the approach by two- step separate optimization. Several interesting phenomena are also observed from the experimental results.

REFERENCE

[1] Data Center Locations [Online]. Available: http: //www. google. com/ about/ datacenters/ inside/ locations/ index. Html.

[2] 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 Proc. 29th Int. Conf. Comput. Commun., 2010, pp. 1–9.

[3] Z. Liu, M. Lin, A. Wierman, S. H. Low, and L. L. Andrew, ‘‘Greening geographical load balancing,’’ in Proc. Int. Conf. Meas. Model. Comput. Syst., 2011, pp. 233–244. [4] J. Dean and S. Ghemawat, ‘‘Mapreduce: Simplified data processing on large clusters,’’ Commun. ACM, vol.51, no. 1, pp. 107–113, 2008.

[4] Z. Liu et al., ‘‘Renewable and cooling aware workload management for sustainable data centers,’’ in Proc. Int. Conf. Meas. Model. Comput. Syst., 2012, pp. 175–186.

[5] J. Cohen, B. Dolan, M. Dunlap, J. M. Hellerstein, and C. Welton, ‘‘Mad skills: New analysis practices for big data,’’ Proc. VLDB Endowment, vol. 2, no. 2, pp. 1481–

1492, 2009.

[6] R. Kaushik and K. Nahrstedt, ‘‘T*: A data-centric cooling energy costs reduction approach for Big Data analytics cloud,’’ in Proc. Int. Conf. High Perform. Comput., Netw., Storage Anal., 2012, pp. 1–11.

[7] S. Gunduz and M. Ozsu, ‘‘A poisson model for user accesses to web pages,’’ in Computer and Information Sciences-ISCIS (Lecture Notes in Computer Science). Berlin, Germany: Springer-Verlag, 2003, pp. 332–339.