iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VI **Month of publication:** June 2022

DOI: https://doi.org/10.22214/ijraset.2022.43747

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VI June 2022- Available at www.ijraset.com

Big Data Processing in Geo-Distributed Data Centers Using Cost Minimization Methods

Senthil Kumar K MCA.,Ph.D.,[1], Karthiga P[2]
[1]*Assistant Professor,* [2] *Scholar, Department of MCA, Karpagam College of Engineering, Coimbatore, India*

Abstract: ***The explosive growth of demands on huge process imposes a crucial burden on computation, storage, and communication in data centers, that so incurs wide operational expenditure to data center suppliers. Therefore, Price decrease has become Associate in Nursing rising issue for the longer term huge data era. All totally different from customary cloud services, one in each of the foremost choices of huge data services is that the tight coupling between data and computation as computation tasks is also conducted provided that the corresponding data is obtainable.***
As a result, 3 factors, i.e., task assignment, data placement and data movement, deeply the operational expenditure of data centers. Throughout this paper, we tend to tend to unit of measurement meant to see the worth decrease downside via a joint improvement of these three factors for big data services in geo-distributed data centers.
To elucidate the task completion time with the thought of every data transmission and computation, we tend to tend to propose a two-dimensional stochastic process and derive the standard task completion time in closed-form. Moreover, we tend to tend to model the matter as a mixed-integer non-linear programming (MINLP) and propose economical resolution to correct it. The high efficiency of our proposal is valid by thorough simulation primarily based studies
Keywords— Nonlinear Programming, Big Data, Computer Centres, Integer Programming, Minimisation, Mixed Integer Nonlinear Programming, Cost Minimization, Data Centers, Big Data Services, Operational Expenditure, Task Assignment, Big Data

I. INTRODUCTION

Data explosion in recent years leads to a rising demand for giant process in stylish information centers that square measure typically distributed at fully totally different geographic regions, e.g., Google's 13 information centers over eight countries in four continents. Huge information analysis has shown its nice potential in unearthing valuable insights of information to spice up deciding, minimize risk and develop new merchandise and services. On the alternative hand, huge information has already Translated Into Huge price as results of its high demand on computation and communication resources. Gartner predicts that by 2015, seventy one amongst worldwide information center hardware payment will come back from the huge process, which is able to surpass $126.2 billion. Therefore, it\'s imperative to review the worth reduction downside for big information process in geo-distributed information centers. Several efforts square measure created to lower the computation or communication worth of information centers. Information center resizing (DCR) has been projected to chop back the computation price by adjusting the number of activated servers via task placement. Supported DCR, some studies have explored the geographical distribution nature of information centers and electricity price no uniformity to lower the electricity worth. Huge information service frameworks, e.g. comprise a distributed file system to a lower place that distributes information chunks and their replicas across the data centers for fine-grained load-balancing and high parallel data access performance. to cut back the communication price, variety of recent studies produce efforts to spice up information section by inserting jobs on the servers where the computer file reside to avoid remote information loading .

Though the on A-one of solutions have obtained some positive results, they\'re faraway from achieving the cost effective massive process as a result of the following weaknesses. First, information neck of the woods may finish in a very waste of resources. as an example, most computation resource of a server with less well-liked information may keep idle. The low resource utility any causes further servers to be activated and therefore higher expense. In interest of performance, Huge information analytics information section constraint restricts the server selections in thermal aware computation placement techniques to exclusively the servers that host a replica of the data to be computed upon; thereby, reducing the potential cooling energy savings. On the other hand, neglecting data-locality results in higher cooling energy savings at the worth of performance. Completely different cooling management techniques use computation migration; they reactively migrate computations from a server with high run-time temperature to lower temperature servers. Computation migration is viable solely servers are state-less; in Huge information analytics cloud servers have important state. To boot, computation migration to a server that doesn't host a replica of the information results in nonlocal information accesses that comes at a performance worth.

II. PROJECT ANALYSIS

A. Objective

The main objective of this system is to give exact details about the fare of the cab to its users which will ultimately help them in choosing the best available transport facility thereby reducing the expense. This will help in reducing the deceive that cab driver impose on passengers. Various studies on existing system has made clear that most of the riders order fare higher than estimation rate which when considered is actually fraudulent. Since passenger cannot set path of route of travel there are chances that rider takes passenger through long distance route thereby ask for increased fare.

B. Existing System

In 2015, 71% of worldwide data center hardware spending will come from the big data processing, it's predicted by Gartner. In Data center resizing (DCR) data locality may result in a waste of resources. Less popular data may stay idle and the low resource utility causes more servers to be activated and hence higher operating cost. Links in networks vary on the transmission rates and costs according to their unique features. If the routing strategy among data centers fails then it is unavoidable to download from a remote server. In this case, routing strategy matters on the transmission cost. The Quality-of-Service (QoS) of big data tasks has not been considered in existing work.

C. Proposed System

The worth minimization downside of large process with joint thought of data placement, task assignment and data routing. to elucidate the rate-constrained computation and transmission in vast method process method, we've got a bent to propose a a pair of dimensional Markov process and derive the expected task completion time in closed sort. supported the closed-or expression, we've got a bent to formulate the worth minimization downside throughout a method of mixed range nonlinear programming (MINLP) to answer the next questions: 1) the simplest way to put these data chunks among the servers, 2) the simplest way to distribute tasks onto servers whereas not violating the resource constraints.3) the simplest way to size data centers to appreciate the operation price minimization goal. To touch upon the high procedure complexity of finding MINLP, we've got a bent to correct it as a mixed-integer maths (MILP) downside, which can be resolved victimization business thinker. Through comprehensive numerical studies, we've got a bent to indicate the high efficiency of projected joint-optimization based totally algorithm.

III. SERVER COST MINIMIZATION

Large-scale data centers have been deployed all over the world providing services to hundreds of thousands of users. According to [11], a data center may consist of large numbers of servers and consume megawatts of power. Millions of dollars on electricity cost have posed a heavy burden on the operating cost to data center providers. Therefore, reducing the electricity cost has received significant attention from both academia and industry [5], [11]–[13]. Among the mechanisms that have been proposed so far for data center energy management, the techniques that attract lots of attention are task placement and DCR. DCR and task placement are usually jointly considered to match the computing requirement. Liu et al. [4] re-examine the same problem by taking network delay into consideration. Fan et al. [12] study power provisioning strategies on how much computing equipment can be safely and efficiently hosted within a given power budget. Rao et al. [3] investigate how to reduce electricity cost by routing user requests to geo-distributed data centers with accordingly updated sizes that match the requests. Recently, Gao et al. [14] propose the optimal workload control and balancing by taking account of latency, energy consumption and electricity prices. Liu et al. [15] reduce electricity cost and environmental impact using a holistic approach of workload balancing that integrates renewable supply, dynamic pricing, and cooling supply.

IV. SYSTEM MODEL

A. Network Model

Networking provides the possibility of orchestrating all resources towards different optimization goals. For data transferring between the storage units and the processing units in big data processing.A communication cost of large volume data transferring is non-ignorable and shall be carefully addressed in the consideration of cost efficiency.A communication cost diversity in geo-distributed data centers towards big data processing cost efficiency and propose a scheduling algorithm that can be incorporate into the scheduler module in cloud networking.

B. Task Assignment

Further increasing the number of servers will not affect the distributions of tasks.Task should be assigned to data centre where number of activated servers is optimal.Task assignment is deeply influence the operational expenditure of data center. Task is

assigned to data centre according to nearest data centre for effectively processing of data. Each data chunk has a storage requirement and will be required by big data tasks.

V. PERFORMANCE EVALUATION

Our discovery that the optimal number of chunk replicas is equal to 4 under the network setting above is verified one more time by solving the formulation that is to minimize the number of replicas with the minimum total cost. Additional results are given under different settings via varying the task arrival rate and chunk size in the ranges of [0.1, Z_$\lambda _{U}$ _Z] and [Z_$\phi _{L}$ _Z, 1.0], respectively, where a number of combinations of Z_$(\lambda _{U}, \phi _{L})$ _Z are shown in Fig5.1. We observe that the optimal number of replica a non-decreasing function of the task arrival rate under the same chunk size while a non-increasing function of the data chunk size under the same task arrival rate.

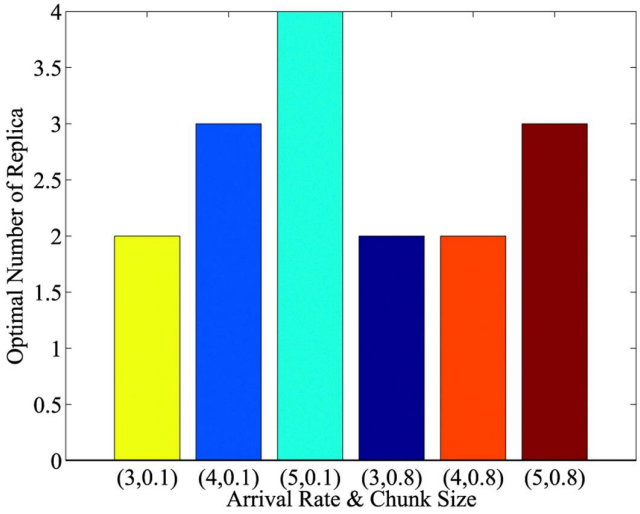


Fig 5.1 Optimal number of replica

VI. CONCLUSION

This project has been developed using .Net as front end and SQL server as backend. It is well planned and organized successfully. Even though some difficulties arise in the conversion of manual work to computerization, we complete and implemented it successfully. In this project each module is thoroughly tested and found working properly.
The most important matter in this project is the simplicity and applicability of the system to the factual cases. This system must be simple, reliable, flexible and convenient to the nature of the application area.

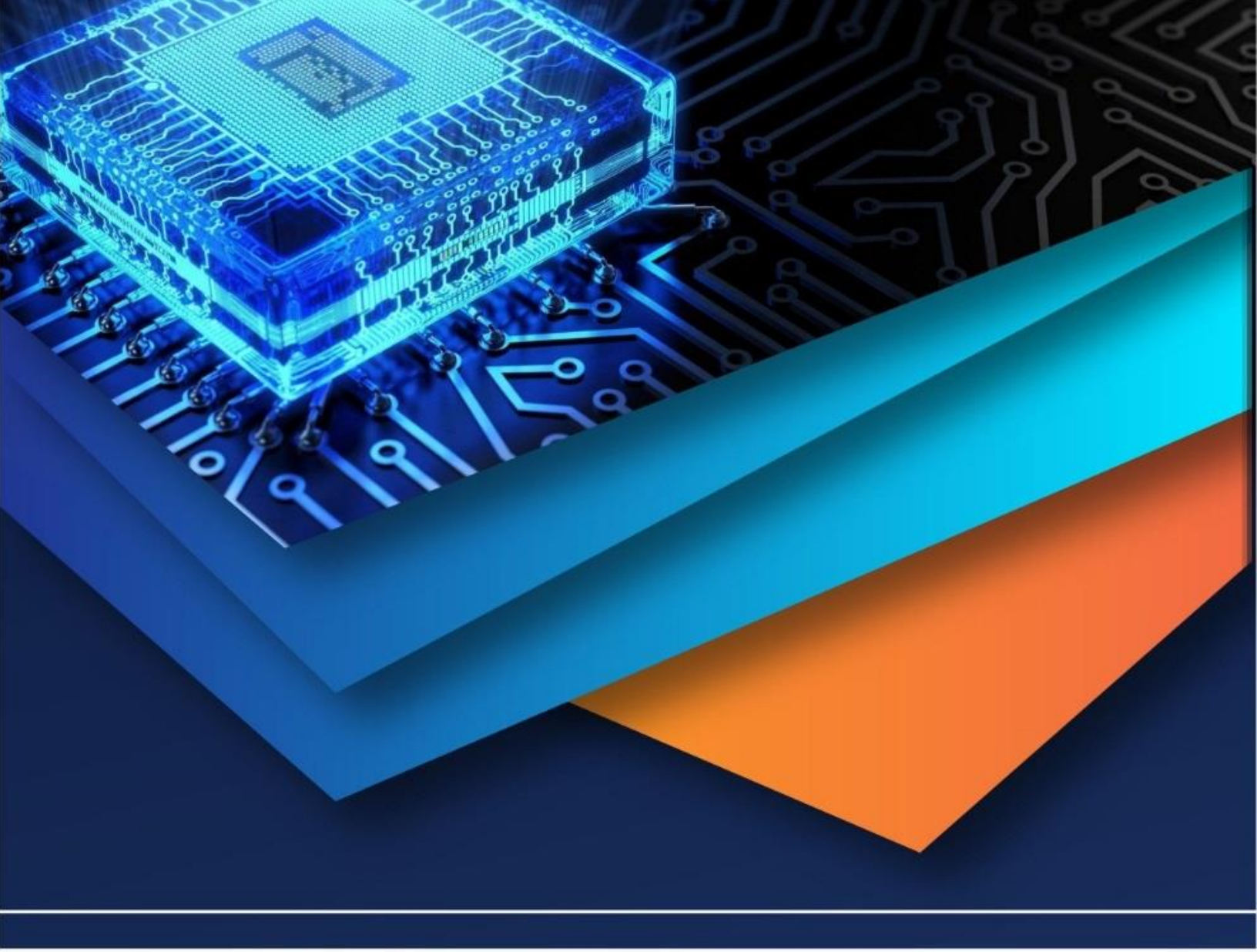
VII. ACKNOWLEDGMENT

This Research Article was supported by Department of MCA, Karpagam College of Engineering, Coimbatore. I have great satisfaction in presenting this article on "BIG DATA PROCESSSING IN GEO DISTRIBUTED DATA CENTERS USING COSI MINIMIZATION METHODS". I take this opportunity to express my sincere thanks to my guide, Prof **Mr.SENTHIL KUMAR MCA.,Ph.D.**, for providing the technical guidelines and suggestions regarding the line of this work. I want to convey my gratitude for his constant encouragement, support and guidance throughout the project's development. I am grateful to **Dr. K.Anuradha** (Director In-Charge, Department of MCA); my project would not have shaped up without their support. I wish to express my deep gratitude toward all my Professor at Karpagam College of Engineering, Coimbatore, for their encouragement.

REFERENCES

[1] (2013). *Data Center Locations* [Online]. Available: http://www.google.com/about/datacenters/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 *Proc. 13th Int. Conf. Archit. Support Program. Syst.*, 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 *Proc. 29th Int. Conf. Comput. Commun.*, 2010, pp. 1–9.
[4] 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.
[5] R. Urgaonkar, B. Urgaonkar, M. J. Neely, and A. Sivasubramaniam, "Optimal power cost management using stored energy in data centers," in *Proc. Int. Conf. Meas. Model. Comput. Syst.*, 2011, pp. 221–232.

[6] H. Xu, C. Feng, and B. Li, "Temperature aware workload management in geo-distributed datacenters," in *Proc. Int. Conf. Meas. Model. Comput. Syst.*, 2013, pp. 33–36.
[7] J. Dean and S. Ghemawat, "Mapreduce: Simplified data processing on large clusters," *Commun. 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," *Comput. Netw. ISDN Syst.*, vol. 30, no. 223, pp. 2123–2130, 1998.
[10] H. Jinet al., "Joint host-network optimization for energy-efficient data center networking," in *Proc. 27th Int. Symp. Parallel Distrib. Process.*, 2013, pp. 623–634.
[11] A. Qureshi, R. Weber, H. Balakrishnan, J. Guttag, and B. Maggs, "Cutting the electric bill for internet-scale systems," in *Proc. ACM Special Interest Group Data Commun.*, 2009, pp. 123–134.
[12] X. Fan, W.-D. Weber, and L. A. Barroso, "Power provisioning for a warehouse-sized computer," in *Proc. 34th Annu. ISCA*, 2007, pp. 13–23.
[13] S. Govindan, A. Sivasubramaniam, and B. Urgaonkar, "Benefits and limitations of tapping into stored energy for datacenters," in *Proc. 38th Annu. ISCA*, 2011, pp. 341–352.
[14] P. X. Gao, A. R. Curtis, B. Wong, and S. Keshav, "It's not easy being Green," in *Proc. ACM SIGCOMM*, 2012, pp. 211–222.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY