Global Scientific JOURNALS

GSJ: Volume 10, Issue 2, February 2022, Online: ISSN 2320-9186
www.globalscientificjournal.com

OPTIMIZATION TECHNIQUES FOR LOAD BALANCING IN CLOUD COMPUTING

Muhammad Ahmad Ali

University of Engineering & Technology, Peshawar, Pakistan
Email: eng.ahmadalikhan@gmail.com

KeyWords
Cloud Computing, Energy Efficiency, Future Prediction Model, Geographical Load Balancing, Geographically Distributed Data Centers, Lyapunov Optimization technique, Online Algorithm.

ABSTRACT

Huge Electricity cost is the key issue for large-scale data center operators. These companies have deployed their data centers across different geographical locations for efficiency and reliability purposes. A Lot of research work is done for the minimization of the overall energy cost of these operators. A section of researchers haS focused on hardware-based techniques while otherS work on soft-ware based techniques. In this research, the software-based techniques and model are targeted and the various approaches used in this field is presented along with pros and cons. Furthermore, this research provides a simplified overview and categorization of the work done in the area of energy cost minimization for geographically distributed data centers. The overall work done is categorized in four major sections which are Future Predication models, Competitive Online Algorithm, Dynamic Programming, and Lyapunov optimization technique.

Introduction

With the passage of time, the demands of IT applications and online services are increasing day by day. In order to fulfill the user demands and provides them the services such as audio distribution, video distribution, web surfing, scientific simulation etc., large companies such as Google, Amazon have deployed their data centers across different geographical locations. These Data Centers consists of thousands of servers, network equipment and cooling systems. Every Data Center consumes a significant amount of brown energy to process the user requests. It is obvious that with increase in user demands/ requests, the energy consumptions of these datacenter will also grow. Tens of megawatts of electricity is required in order to keep the data center in the running state [1]. The datacenter operators pay millions of dollars annually in terms of their electricity bills and constitutes 30-50% of their operational cost [2]. According to a research conducted in 2010, the global electricity demand of all the datacenters were 70 TWH and based on these figure, the projected consumption of these datacenters in 2030 will be almost 1800 TWH [3]. Datacenters will be consuming 3-13 % of the global electricity as compared to 1% consumption in 2010 [3]. Fig. 1. represent the expected electricity consumption from 2010 to 2030 [3].

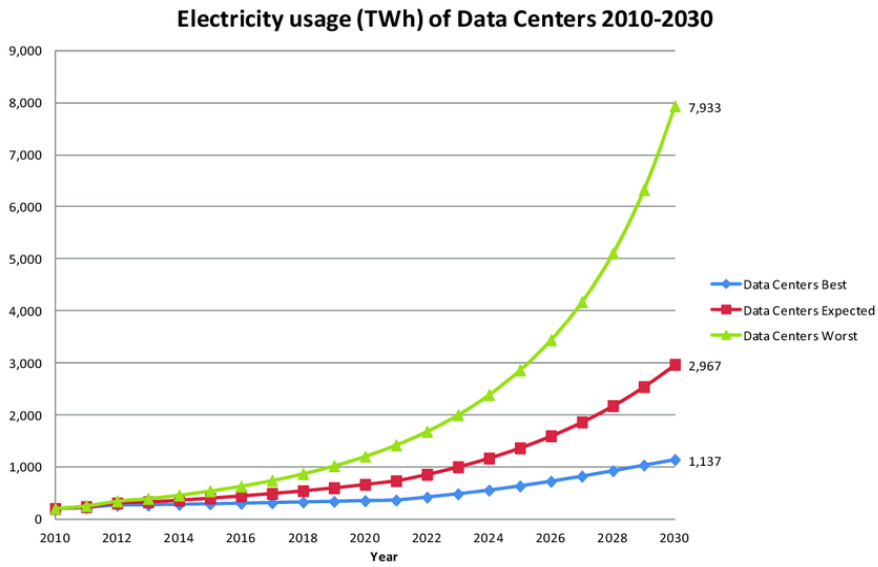

Figure 1: Electricity usage from 2010-2030 (Source: Andrae et al [3])

Datacenter operators have deployed their datacenters across various geographical location for two main purposes i.e. efficiency and reliability. The user request is initially reached at central location known as aggregator. The user request is of two types. 1) Delay Intolerant Workload: This type of workload is also known as interactive workload. This type of workload must be executed at the instance it arrived in the system. It does not tolerate any kind of delay. 2) Delay Tolerant Workload: This type of workload could be delayed up to certain extend. The typical system model is illustrated in Fig. 1.

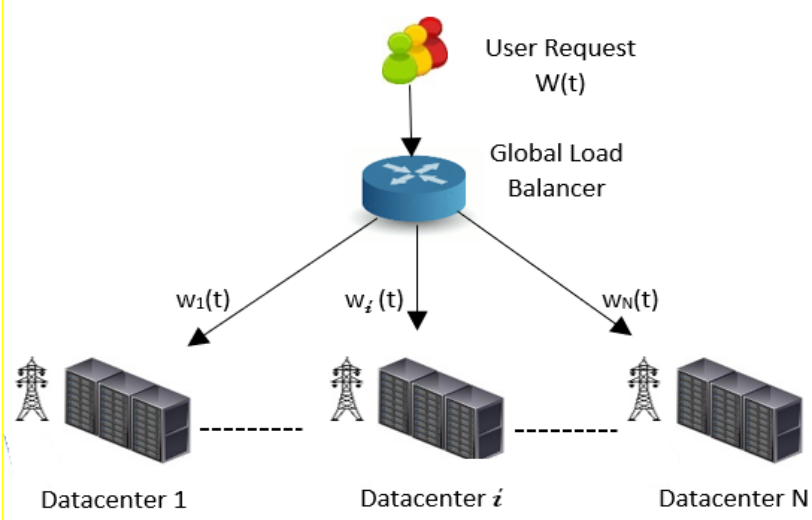

Fig. 1. Workload Distribution Model (Source: Khalil et al[4])

From here on, the workload is forwarded to one of the datacenter based on one or many objectives. The objective may be to transfer the workload to that datacenter which is the nearest, or where the renewable energy is sufficiently available, or where the per unit electricity cost is the minimum [5].

According to Qureshi et al [6], electricity prices varies across time and location. Researchers have exploit this point and tries to minimize the electricity cost by scheduling the workload in such a way that maximum workload could be served at that time and location when and where the electricity cost is minimum.

In order to lower down the electricity cost of the datacenters operator, researchers have worked on various aspect using different techniques. The work done for minimizing the overall cost of datacenter operator can be categorized in two broad categories. One section of researchers have work of software based techniques such as proactive demand response [7], visualization [8], dynamic power management [9], option pricing [10] etc. while other researchers have focused their attention on reducing the electricity cost of geographically distributed datacenters using hardware based techniques such as dynamic voltage, chip multiprocessor [11], and frequency scaling [12].

Large internet service providers have deployed various datacenters across different geographical locations for two key purposes i.e. efficiency and reliability. Electricity prices varies across time and location. This variation can be utilized to minimize the average electricity cost of geographically distributed datacenters. Qureshi et al. [6] was the pioneer of pointing out that variation exist in the electricity prices at different locations and at different times could be exploit intelligently for minimization the overall cost of the datacenters operator. The key for getting maximum benefit from this fluctuation is possible by forwarding the workload to those data centers where currently the electricity prices are lower.

Optimization Techniques

a. Future Prediction Model

In this model, the future inputs i.e. workload amount, workload deadline and per unit electricity cost are predicted on the basis of previous historical data [13]. As the prediction depends upon previous historical data, therefore the prediction accuracy is also heavily dependent to the stability and the pattern exist in the previous data. There are two major drawbacks/ limitation in using this model. First of all, the future can't be predicted with hundred percent accuracy. Secondly, the environment were previous data is not available, this model can't be applied directly.

Buchbinder et al. [14] investigated how to lower the total operational cost of distributed DCs for batch workloads. The proposed algorithm's main goal was to strike a balance between power and bandwidth costs when shifting workloads from one DC to another. The programme took into account not only the current price and availability of energy sources, but also the expected data for future prices and availability of energy sources. The fundamental disadvantage of predictive models is their inability to forecast the future.

b. Competitive Online Algorithm

In this model, an online algorithm is developed and then compare it against the performance of offline algorithm in worst case scenario [15, 16, 17, 18, 19]. This model do not needs the future information in advance. The performance is measured in terms of competitive ratio. The smaller the ratio, the better the online algorithm is considered. The main limitation of this model is that in real world one deal majorly with the average case scenario instead of worst case.

Toosi et al. [20] employed competitive online algorithms in order to solve the problem of minimizing the energy cost of data centers by utilizing the renewable energy. Based on the availability of renewable energy, the workload was dispersed among data centres located in various locations. Due to unknown workload amounts and unexpected weather conditions, this system has a severe flaw: it cannot totally rely on renewable energy.

c. Dynamic Programming

Dynamic Programming is also one of the technique for solving the cost minimization problems. But there are two major issues in solving the optimization problem through Dynamic Programming. One problem while using this technique is that the future statistics of the incoming workload and electricity prices does not exist. Even if we use the past historical information for making a prediction about the future data and try to solve the problem with various existing techniques such as the Dynamic programming. This way of solving the above optimization will not work will because these techniques suffers from the "curse of dimensionality" problem as the computational complexity grows higher with the size [21].

d. Lyapunov Optimization technique

It follow a greedy type approach in which it tries to achieve an objective by minimize the cost between the consecutive timeslots and hopes that at the end the overall cost would be minimized [22][23]. Normally there exist contradictory type of requirements in the objective function. For example, on one side the objective is to reduce the electricity cost while at the same time the aim is also to reduce the workload delay as well. Now, if the steps are taken to reduce the cost, then the algorithm may will waits till that moment where per unit cost is sufficiently lower, but by doing so the workload delay will definitely increases. On other side, if measures are taken to decrease the workload delay, then the workload will be executed irrespective of the current per unit cost and due to which the cost may increase. Here a fine tradeoff has to be made b/w these contradictory requirements. A tradeoff variable known as V is used for this purpose. The main feature of this technique is that it is easy to implement and does not depends upon the system sta-

tistics. Some of the work done in this area are as follows:

Urgaonkar et al. [23] suggested an online algorithm based on the Lyapunov Optimization technique for reducing electricity costs by employing storage devices with continuous power supply (UPS). The algorithm's logic is to charge the battery when market prices are lower and use it when electricity prices are higher. The biggest drawback is that the battery has a finite capacity and may only be utilised for a certain amount of time. Second, this strategy minimises the entire workload's average latency.
Polverini et al. [24] presented GreFar as an online algorithm for scheduling batch workloads across multiple data centres. The goal of the algorithm was to lower the cost of power while also allocating available resources fairly among different requests while meeting the average delay and maximum inlet temperature requirements. The algorithm was contrasted to an offline algorithm that knew what would happen next. A simulation-based technique was used to assess the algorithm's performance. The key disadvantage or constraint is that the GreFar reduce the average delay but do not guarantee that each work will be completed on time.
Researchers are also working on other aspects of reducing the energy cost of geographically distributed datacenters, such as minimising the cost in a multi-electricity market environment [25], proactive demand response for data centres [26], lowering the electricity price for interactive workload [27], and so on.

Conclusion and Future Work

Datacenter Operators such as Google, Facebook, Amazon etc. bear a major portion of their operational cost in the shape of electricity consumption. Many researchers work on hardware and software based techniques in order to minimize the electricity cost of those operators. One angle to minimize the electricity cost of those operator is to exploit the variation of per unit electricity cost that exist along the time and space. In this paper, we summarized and present the work done in the area of minimizing the overall electricity cost of datacenter operators by using software based techniques and model. There are mainly four models and techniques used in the literature for achieving the cost minimization objective. These are Future Predication models, Competitive Online Algorithm, Dynamic Programming and Lyapunov optimization technique. Pros and cons of each of them is also explained.
In future, we would like to investigate and includes the aspect of using the renewable source of energy, energy storage devices on various types of workloads for minimizing the electricity cost of the geographically distributed datacenters.

References

[1] F. Kong and X. Liu, "A Survey on Green-Energy-Aware Power Management for Datacenters", ACM Computing Surveys, vol. 47, no. 2, pp. 1-38, 2014. Available: 10.1145/2642708.

[2] Y. Guo, Z. Ding, Y. Fang, and D. Wu. "Cutting Down Electricity Cost in Internet Data Centers by Using Energy Storage." 2011 IEEE Global Telecommunications Conference - GLOBECOM 2011, 2011. doi:10.1109/glocom.2011.6134209.

[3] Andrae, A.S.G.; Edler, T. On Global Electricity Usage of Communication Technology: Trends to 2030. *Challenges* **2015**, *6*, 117-157. https://doi.org/10.3390/challe6010117

[4] M. I. K. Khalil, I. Ahmad and A. A. Almazroi, "Energy Efficient Indivisible Workload Distribution in Geographically Distributed Data Centers," in IEEE Access, vol. 7, pp. 82672-82680, 2019, doi: 10.1109/ACCESS.2019.2924085.

[5] Ahmad, I., Khalil, M. I. K., & Shah, S. A. A. (2020). Optimization-based workload distribution in geographically distributed data centers: A survey. International Journal of Communication Systems, 33(12), e4453.

[6] A. Qureshi, R. Weber, H. Balakrishnan, J. Guttag, and B. Maggs, "Cutting the electric bill for internet scale systems," in ACM SIGCOMM, Barcelona, Spain, August 2009.

[7] X. Zheng and Y. Cai, "Energy-awar load dispatching in geographically located Internet data centers," Sustainable Computing: Informatics and Systems, vol. 1, no. 4, pp. 275-285, 2011.

[8] R. Nathuji and K. Schwan, "VirtualPower", ACM SIGOPS Operating Systems Review, vol. 41, no. 6, p. 265, 2007. Available: 10.1145/1323293.1294287.

[9] D. Meisner, B. Gold and T. Wenisch, "PowerNap", *ACM SIGPLAN Notices*, vol. 44, no. 3, p. 205, 2009. Available: 10.1145/1508284.1508269.

[10] Khalil, M. I. K., Ahmad, I., Shah, S. A. A., Jan, S., & Khan, F. Q. (2020). Energy cost minimization for sustainable cloud computing using option pricing. *Sustainable Cities and Society*, *63*, 102440.

[11] L. Barroso, "The price of performance", Queue, vol. 3, no. 7, p. 48, 2005. Available: 10.1145/1095408.1095420.

[12] T. Horvath, T. Abdelzaher, K. Skadron and X. Liu, "Dynamic Voltage Scaling in Multitier Web Servers with End-to-End Delay Control", IEEE Transactions on Computers, vol. 56, no. 4, pp. 444-458, 2007. Available: 10.1109/tc.2007.1003.

[13] C.-K. Chau and L. Yang, "Competitive online algorithms for geographical load balancing in data centers with energy storage," in *Proceedings of the 5th International Workshop on Energy Efficient Data Centres*, 2016, p. 1: ACM.

[14] N. Buchbinder, N. Jain and I. Menache. "Online Job-Migration for reducing the electricity bill in the cloud," in J. Domingo-Pascual et al. (Eds.): NETWORKING 2011, Part I, LNCS 6640, pp. 172–185, 2011.@ IFIP International Federation for Information Processing 2011.

[15] A. Borodin and R. El-Yaniv. Online computation and competitive analysis. Cambridge University Press, 1998.

[16] C.-K. Chau, M. Khonji, and M. Aftab. Online algorithms for information aggregation from distributed and correlated sources. to appear in IEEE/ACM Trans. Networking, 2016. http://arxiv.org/abs/1601.03147.

[17] L. Lu, J. Tu, C.-K. Chau, M. Chen, and X. Lin. "Online energy generation scheduling for microgrids with intermittent energy sources and co-generation". In ACM SIGMETRICS, 2013.

[18] J. Iqbal and I. Ahmad, "Optimal online k-min search," EURO Journal on Computational Optimization, vol. 3, no. 2, pp. 147–160, 2015.

[19] Ahmad, Iftikhar, Marcus Pirron, and Günter Schmidt. "Analysis of threat based algorithm using different performance measures." *RAIRO: Recherche Opérationnelle* 55 (2021): 2393.

[20] A. Nadjaran Toosi, C. Qu, M. de Assunção and R. Buyya, "Renewable-aware geographical load balancing of web applications for sustainable data centers", Journal of Network and Computer Applications, vol. 83, pp. 155-168, 2017. Available: 10.1016/j.jnca. 2017.01.036.

[21] D. P. Bertsekas, Dynamic Programming and Optimal Control, 2nd ed. Athena Scientific, 2000.

[22] M. Lin, A. Wierman, L.L.H. Andrew, and E. Thereska, "Dynamic Right-Sizing for Power-Proportional Data Centers," Proc. IEEE INFOCOM, 2011.

[23] R. Urgaonkar, B. Urgaonkar, M. Neely and A. Sivasubramaniam, "Optimal power cost management using stored energy in data centers", ACM SIGMETRICS Performance Evaluation Review, vol. 39, no. 1, p. 181, 2011. Available: 10.1145/2007116.2007138.

[24] M. Polverini, A. Cianfrani, S. Ren and A. Vasilakos, "Thermal-Aware Scheduling of Batch Jobs in Geographically Distributed Data Centers", IEEE Transactions on Cloud Computing, vol. 2, no. 1, pp. 71-84, 2014. Available: 10.1109/tcc.2013.2295823.

[25] 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 INFOCOM, 2010 Proceedings IEEE, 2010, pp. 1-9: IEEE.

[26] H. Wang, J. Huang, X. Lin, and H. Mohsenian-Rad, "Proactive demand response for data centers: A win-win solution," IEEE Transactions on Smart Grid, vol. 7, no. 3, pp. 1584-1596, 2016.

[27] Y. Xu, Y. Zhan, and D. Xu, "Building cost efficient cloud data centers via geographical load balancing," in Computers and Communications (ISCC), 2017 IEEE Symposium on, 2017, pp. 826-831: IEEE.