ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 3, Issue 3, pp: (272-278), Month: July - September 2015, Available at: www.researchpublish.com

A Three Parameter Line of Attack to Downplay the Expenses of Data Processing

[1]Jayanth N R, [2]Sujatha Ramesh P

[1]Student, [2]Assistant Professor, [1,2] Dept. of CSE Sambhram Institute Of Technology Bengaluru, Karnataka, India

***Abstract*: 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 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 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 2-D Markov chain and derive the average task completion time in closed-form. Furthermore, we model the problem as a mixed-integer nonlinear programming and propose an efficient solution to linearize it. The high efficiency of our proposal is validated by extensive simulation-based studies.**

***Keywords:* Big data, data flow, data placement, distributed data centers, cost minimization, task assignment.**

I. INTRODUCTION

Information blast as of late prompts a rising interest for enormous information preparing in present day server farms that are normally conveyed at diverse geographic locales, e.g., Google's 13 server farms more than 8 nations in 4 landmasses [1]. Huge information examination has demonstrated its extraordinary potential in uncovering profitable experiences of information to enhance choice making, minimize hazard what's more, and grow new items and administrations. Then again, huge information has effectively deciphered into huge cost because of its high request on reckoning and correspondence assets [2]. Gartner predicts that by 2015, 71% of overall server farm equipment spending will originate from the enormous information handling, which will surpass $126.2 billion. In this way, it is basic to concentrate on the expense minimization issue for enormous information processing in geo-circulated server farms.

Numerous endeavors have been made to bring down the processing alternately correspondence expense of server farms. Server farm resizing (DCR) has been proposed to decrease the reckoning cost by changing the quantity of actuated servers by means of undertaking situation [3]. Taking into account DCR, a few studies have investigated the land conveyance nature of server farms and electricity value heterogeneity to bring down the power cost [4] [6]. Enormous information administration systems, e.g., [7], involve a circulated file framework underneath, which circulates information lumps and their copies over the server farms for one-grained burden adjusting and high parallel information access execution. To decrease the correspondence cost, a couple of late studies make endeavors to enhance information area by setting occupations on the servers where the info information live to stay away from remote information stacking [7], [8].

Despite the fact that the above arrangements have gotten some positive results, they are a long way from accomplishing the expense efficient huge information preparing due to the accompanying shortcomings. In the first place, information territory may bring about a misuse of assets. Case in point, most reckoning asset of a server with less prevalent information may stay unmoving. The low asset utility further causes more servers to be initiated and consequently higher working cost.

Second, the connections in systems shift on the transmission rates and expenses as indicated by their remarkable components [9], e.g., the separations and physical optical fiber offices between server farms. On the other hand, the current steering system among server farms neglects to abuse the connection differing qualities of information center systems. Because of the stockpiling and reckoning limit imperatives, not all errands can be set onto the same server, on which their relating information dwell. It is unavoidable that certain information must be downloaded from a remote server. For this situation, steering method matters on the transmission cost.

Third, the Quality-of-Service (QoS) of huge information assignments has not been considered in existing work. Like customary cloud administrations, huge information applications additionally display Service- Level-Agreement (SLA) between an administration supplier and the requesters. To watch SLA, a certain level of QoS, as a rule as far as assignment finishing time, might be ensured. The QoS of any distributed computing assignments is first dictated by where they are set and what number of processing assets are assigned. In addition, the transmission rate is another influential component since enormous information assignments are information driven and the calculation errand can't continue until the comparing information are accessible. Existing studies, e.g., [3], on broad cloud processing errands for the most part concentrate on the reckoning limit limitations, while overlooking the requirements of transmission rate.

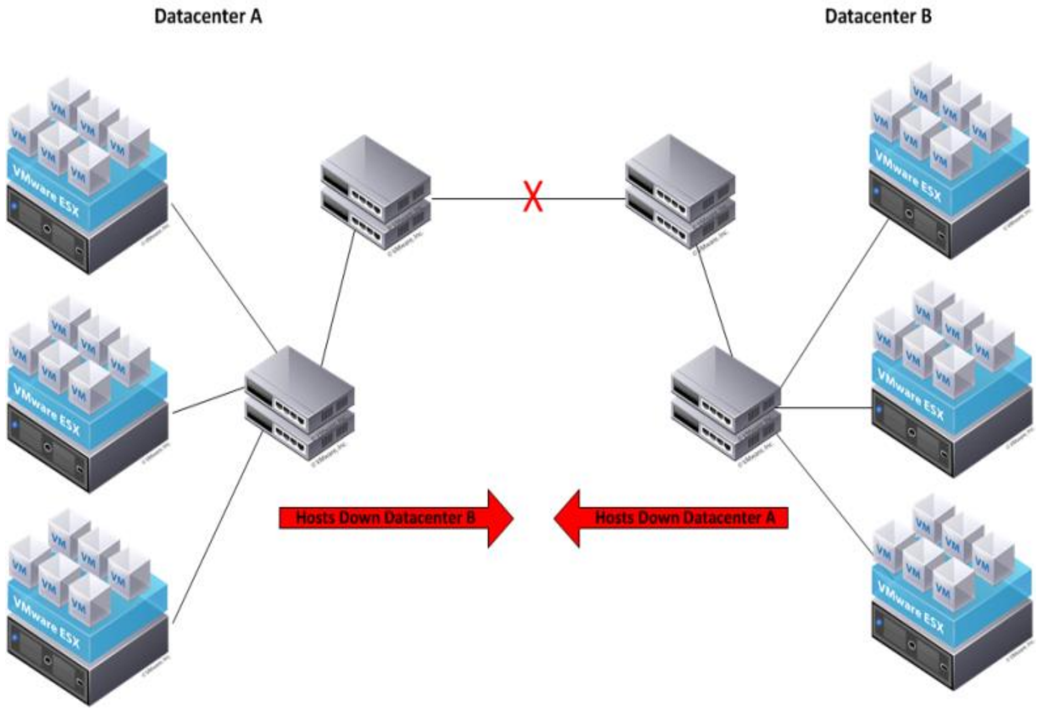

Fig 1 Datacenter Topology

To vanquish above shortcomings, we mull over the expense minimization issue for huge information preparing through joint optimization of undertaking task, information arrangement, and directing in geo-disseminated server farms. Specifically, we consider the accompanying issues in our joint streamlining. Servers are outfitted with constrained stockpiling and processing assets.

Every information lump has a stockpiling prerequisite and will be needed by enormous information errands. The information position and errand task are straightforward to the information clients with ensured QoS. Our goal is to improve the huge information position, errand task, steering and DCR such that the general processing and correspondence expense is minimized. Our principle commitments are outlined as takes after:

To our best information, we are the first to consider the cost minimization issue of enormous information handling with joint thought of information position, errand task what's more, information directing. To portray the rate-obliged computation and transmission in enormous information handling procedure, we propose a two-dimensional Markov chain and infer the normal errand culmination time in shut structure.

Based on the shut structure expression, we figure the cost minimization issue in a type of blended number nonlinear programming (MINLP) to answer the take aftering inquiries: 1) how to put these information lumps in the servers, 2) how to disperse errands onto servers without abusing the asset limitations, and 3) how to resize server farms to accomplish the operation cost minimization objective.

To manage the high computational multifaceted nature of solving MINLP, we linearize it as a blended whole number straight programming (MILP) issue, which can be illuminated utilizing business solver. Through broad numerical studies, we demonstrate the high efficiency of our proposed joint-streamlining based calculation.

Whatever is left of the paper is composed as takes after. Segment II outlines the related work. Area III represents the proposed system. The framework need to implement the methodology is discussed in Section IV. The hypothetical findings are verified by tests in Segment V. At last, Section VI concludes our work.

II. RELATED WORK

2.1 Minimizing Server cost:

Several of substantial scale server farms are sent giving administrations to substantial clients. As recommended in [11], a server farm may contain huge servers and swallow high power. A large number of dollars cost on power have cause a genuine inconvenience on the working expense to server farm suppliers. Along these lines, minimizing the power expense has set up real consideration from both the educated community and industry [4], [5]–[7]. Server farm Resizing and information position are for the most part together measured to coordinate the preparing necessity. [8]. Suggest the best workload control by taking record of inertness, vitality use and power cost.

2.2 Managing Big Data:

To embrace the difficulties of effectively overseeing enormous information, numerous choices have been proposed to recoup the capacity and preparing expense. The point of interest in overseeing enormous information is dependable and productive information position. [9] Utilization of adaptability in the information position strategy to support vitality productivity in server farms and propose a booking calculation.

Likewise, distribution of PC assets to undertaking has moreover strained much fixation. Cohen et al. [10] grew new outline disposition, systems and learning giving another attractive, dexterous and profound information investigation for one of the world's real promoting systems at Fox Audience Network, by utilizing Greenplum parallel database framework.

III. PROPOSED SYSTEM

In light of the investigation of information situation, undertaking task, server farm resizing and directing, the general operational cost in huge scale geo-appropriated server farms for enormous information applications will be minimized. First portray the information preparing procedure utilizing a two-dimensional Markov chain and infer the normal consummation time in shut structure, in view of which the joint improvement is figured as a MINLP issue. To handle the high computational multifaceted nature of understanding MINLP, linearize it into a MILP issue. Through broad tests, joint-enhancement arrangement has significant point of interest over the methodology by two-stage separate enhancement. K most limited way calculation is used to perform the base most brief way for steering.

IV. SYSTEM MODEL

The system model can be categorised into three main divisions namely, Big data flow or movement, Data Placement and Task assignment which are briefed below.

4.1 Big data and Data Flow:

Collecting dataset for big data is the first task. The whole system can be modelled as a directed graph G = (N;E).Receive data flows from source nodes and forward them according to the routing strategy. The weight of each link w(u;v), representing the corresponding communication cost, can be defined as

$$\omega^{(u,v)} = \begin{cases} C_R, & if\ u, v \in M, \\ C_L, & Otherwise. \end{cases} \qquad (1)$$

Where CR and CL, and are the inter-data centre traffic and local transmission cost such that CR> CL.

4.2 Data placement:

We define a binary variable yjk to denote whether chunk k is placed on server j as follows,

$$y_{jk} = \begin{cases} 1, & if\ chunk\ k\ is\ placed\ on\ the\ server, \\ 0, & Otherwise. \end{cases} \qquad (2)$$

In the distributed file system, we maintain P copies for each chunk k < K, which leads to the following constraint:

$$\sum_{j \in J} y_{jk} = P, \forall_k \in K \qquad (3)$$

Furthermore, the data stored in each server j belongs to J cannot exceed its storage capacity, i.e.

$$\sum_{k \in K} y_{jk} . \emptyset_k \le 1, \forall_j \in J \qquad (4)$$

Finally, we define a binary variable *xj* to denote whether server *j* is activated, i.e.

$$x_j = \begin{cases} 1, & if\ server\ is\ activated, \\ 0, & Otherwise. \end{cases} \qquad (5)$$

A server shall be activated if there are data chunks placed onto it or tasks assigned to it.

The data placement and task assignment are transparent to the data users with guaranteed QOS.

4.3 Task Assignment:

We consider big data tasks 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 $k \in K$ has the size of $\emptyset_k$ ($\emptyset_k \le 1$), which is normalized to the server storage capacity. *P*-way replica [19] is used in our model. That is, for each chunk, there are exactly *P* copies stored in the distributed file system for resiliency 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 [9], [24]. In particular, let $\lambda\, k$ be the average task arrival rate requesting chunk *k*. Since these tasks will be distributed to servers with a 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 $\lambda\, jk$ ($\lambda\, jk \le 1$). When a task is distributed to a server where its requested data chunk 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 statistics have been developed and applied to the decision making in data centers [19]. To keep the data center settings up-to-date, data center operators may make adjustment according to the task predication period by period [3], [14], [15]. This approach is also adopted in this paper.

V. EXPERIMENTAL EVALUATION AND RESULT

We analysed the performance of joint-optimization algorithm and also we compare it with another optimization scheme algorithm (“Non-joint”), which it finds minimum number of servers to be activated and the data routing scheme using the network model.

The given data is preprocessed first and the preprocessed data is given into the geographically distributed data center. We consider |J| = 2 data centers, each of which is with the same number of servers. The data center link communication cost are set as CL = 1 and CR = 4, respectively. The cost Pj on each activated server j is set to 1. The data size, storage requirement, and task arrival rate are all randomly generated.

5.1 Performance based on number of servers:

When the total number of servers increases as shown in fig. 2 the communication cost of both joint and non-joint optimization algorithm will decreases significantly. This is because more tasks and data chunks are placed in the same data center when more servers are placed in the data centers. Hence the communication cost greatly reduced.

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 3, Issue 3, pp: (272-278), Month: July - September 2015, Available at: www.researchpublish.com

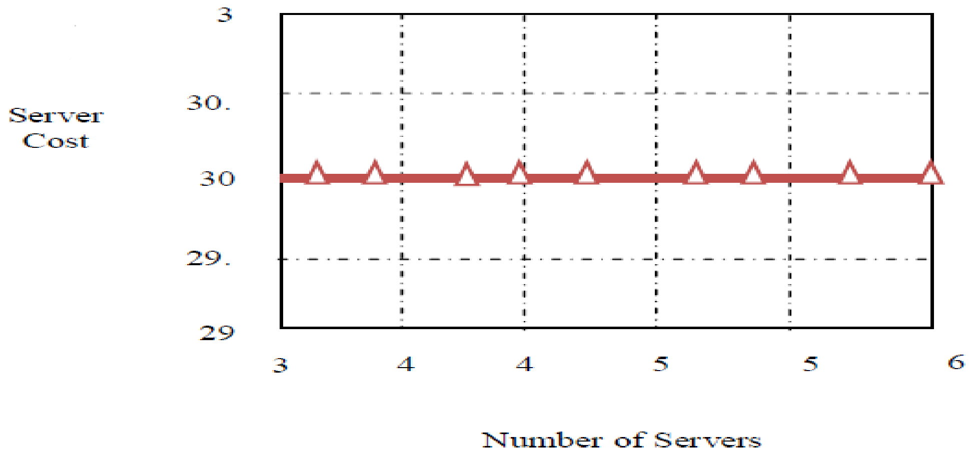

Fig.2 Server Cost

More tasks and their corresponding data chunks can be placed in the same data center, or even in the same server. Further increasing the number of servers will not affect the distributions of tasks or data chunks any more.

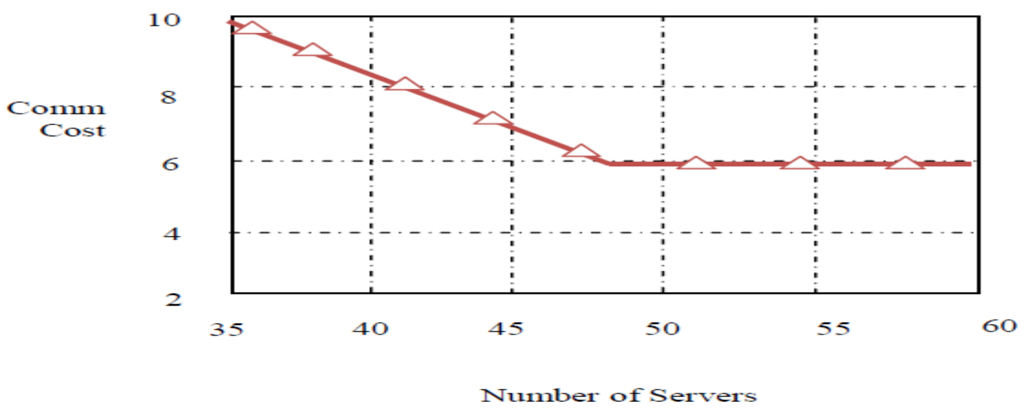

Fig.3 Communication Cost.

5.2 Performance Based on Data Size:

Fig. 4 shows the cost as a function of the total data lump size from 8.4 to 19. Larger chunk size leads to activating new servers with increased server cost. At the same time, fig. 5 shows more resulting traffic over the links creates higher communication cost.

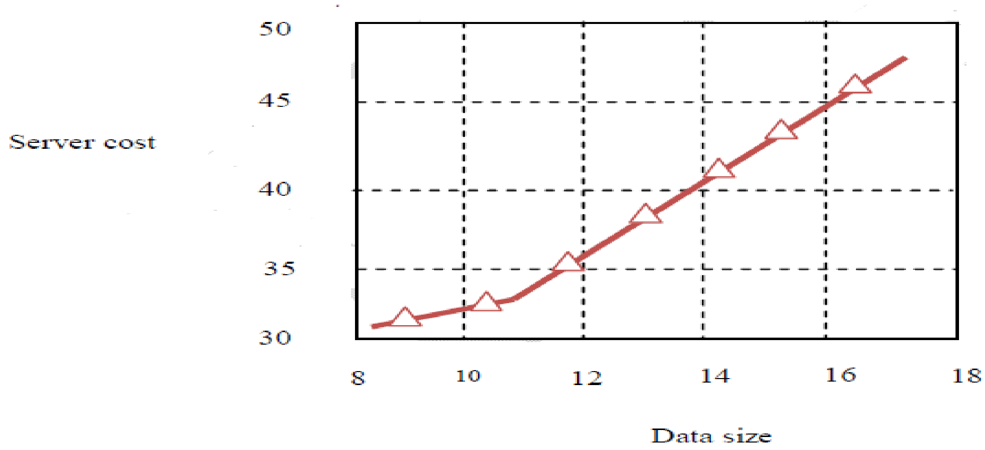

Fig.4 Server Cost

The data chunks are placed in multiple servers and each data chunk is processed separately. When the delay requirement is very small, more servers will be activated to guarantee the QoS. Therefore, the server costs of both algorithms decrease as the delay constraint.

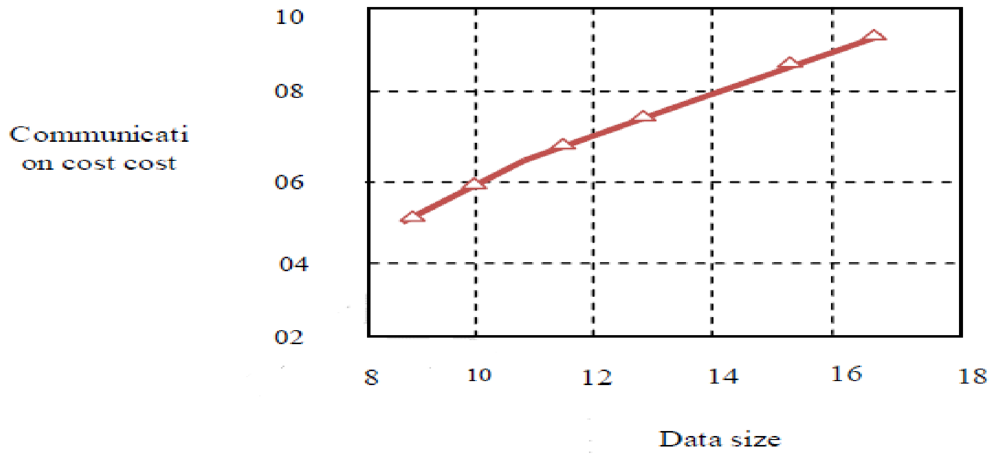

Fig.5 Communication Cost.

VI. CONCLUSION AND FUTURE WORK

In this paper, we mutually concentrate on the information arrangement, undertaking assignment of task, server farm resizing and steering to minimize the by and large operational cost in expansive scale geo-disseminated server farms for huge information applications. We first portray the information processing procedure utilizing a two-dimensional Markov chain and determine the normal finish time in shut structure, in light of which the joint streamlining is figured as a MINLP issue. To handle the high computational many-sided quality of fathoming our MINLP, we linearize it into a MILP issue. Through broad trials, we demonstrate that our joint-streamlining arrangement has significant favorable position over the methodology by two-step separate improvement. A few fascinating phenomena are additionally saw from the test results.

REFERENCES

[1] "Cost Minimization for Big Data Processing in Geo-Distributed Data Centers," in IEEE Transactions on Emerging Topics in Computing, volume 2, No 3, 2014, pp 314-323.

[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.

[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 Management in Geo-distributed Datacenters," in Proceedings of International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS). ACM, 2013, pp. 33–36.

[7] J. Dean and S. Ghemawat, "Mapreduce: simplified data processing on large clusters," Communications of the 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," Computer Networks and ISDN Systems, vol. 30, no. 223, pp. 2123 – 2130, 1998.

[10] 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.

[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, ``Bene_ts and limi- tations 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.

[15] 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.

[16] M. Sathiamoorthy et al., ``Xoring elephants: Novel erasure codes for big data," in Proc. 39th Int. Conf. Very Large Data Bases, 2013, pp. 325_336.

[17] B. Hu, N. Carvalho, L. Laera, and T. Matsutsuka, ``Towards big linked data: A large-scale, distributed semantic data storage," in Proc. 14th Int. Conf. Inf. Integr. Web Based Appl. Services, 2012, pp. 167_176.

[18] 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.

[19] 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.

[20] F. Chen, M. Kodialam, and T. V. Lakshman, ``Joint scheduling of processing and shuf_e phases in mapreduce systems," in Proc. 29th Int. Conf. Comput. Commun., 2012, pp. 1143_1151.

[21] A. Qureshi, R.Weber, H. Balakrishnan, J. Guttag, and B.Maggs, "Cutting the electric bill for Internet-scale systems," in Proc. ACM SIGCOMM, Aug. 2009, pp. 123–134.

[22] A. Beloglazov, R. Buyya, Y. C. Lee, and A. Zomaya, "A taxonomy and survey of energy-efficient data centers and cloud computing systems," Adv. Comput., vol. 82, no. 2, pp. 47–111, 2011.

[23] Amazon Web Services. http://aws.amazon.com.

[24] K. Amiri, D. Petrou, G. Ganger, and G. Gibson. Dynamic Function Placement for Data-intensive Cluster Computing. In USENIX Annual Technical Conference, 2000.