International Conference on "Emerging Trends in Computer Engineering, Science and Information Technology"- 2015
Special Issue of International Journal of Electronics, Communication & Soft Computing Science and Engineering, ISSN: 2277-9477

A New Effective Approach to Move Huge Data to Cloud Minimizing Cost

Leena B. Dhangar **Prof. Amitkumar Manekar**

***Abstract* — Big Data is a promising focus in the cloud computing arena that is getting a lot of significance. Information is progressively more important in our daily lives. We access the internet everyday to perform searches, and other applications. So processing of such huge data requires well suited resources as well as lot of time. In many scenarios, data are, however, geographically distributed across multiple data centers. So sometimes it becomes necessary to transfer data from one server to other due to some reasons. Therefore cost minimization for processing data has become an important issue in big data era. Here considered three factors like Data assignment, data placement and data movement which manipulate the operational expenses of data centers. Hence, this paper presents an online lazy migration (OLM) algorithm and a randomized fixed horizon control (RFHC) algorithm which minimizes the cost of data transfer to the cloud.**

***Key Words* — Big Data, Cloud Computing, Cost Minimization, Data Movement**

I. INTRODUCTION

Big data is a term that refers to data sets or combinations of data sets whose size (volume) [4], complexity (variability), and rate of expansion (velocity) make them difficult to capture, manage, process or analyze by usual technology and tools. Whereas the size used to determine whether a particular data set is considered big data is not firmly defined and continues to transform over era, nearly all analysts and practitioners at present refer to data sets from terabytes to multiple petabytes as big data [1].

Over the past several years there has been a tremendous increase in the amount of data being transferred between Internet users. Escalating usage of streaming multimedia [3] and other Internet based applications has contributed to this surge in data transmission. An additional facade of the augment is due to the expansion of Big Data [18], which refers to data sets that are an order of magnitude larger than the standard file transmitted via the Internet. Big Data can range in size from hundreds of gigabytes to petabytes [11].

Today everything is being stored digitally. Within the past decade, everything from banking transactions to medical history has migrated to digital storage. This change from physical documents to digital files [12] has necessitated the creation of large data sets and consequently the transfer of large amounts of data. There is no sign that the amount of data being stored or transmitted by users is steady or even decreasing. Every year average Internet users are moving more and more data through their Internet connections [12]. Depending on the bandwidth of these connections and the size of the data sets being transmitted, the duration of transfers could potentially be measured in days or even weeks.

There exists a need for an efficient transfer technique that can move large amounts of data quickly and easily without impacting other users or applications. Thus Big Data has translated already into the big price because of its high demands on computation and communication resources [14]. Accordingly considering the bandwidth of these connections and size of these data sets that is being transmitted, duration of data transfer could be measured in terms of days or even weeks. Therefore it becomes necessary to invent an approach that will minimize the cost [1] of processing of this big data.

Thus Big data analysis is one of the key challenges of current era. The restrictions to what able to be done are often times due to how much amount of data can be processed in a given period of time. Big data sets innately occur due to applications generating more information to get better operation, performance; general applications like social networks supports every individual users in producing massive amounts of data. The cloud computing paradigm enables rapid on-demand provisioning of server resources (CPU, storage, bandwidth) to users, with minimal managing hard work. Recent cloud platform, as exemplify by Amazon EC2 and S3, Microsoft Azure, Google App Engine, Rack space, etc., organize a shared pool of servers from multiple data center, and provide their users with virtualization technology.

The elastic and on-demand nature of resource provisioning makes a cloud platform attractive for the execution of various applications, especially computation-intensive ones [2], [3]. More and more data-intensive Internet applications such as the Human Genome Project [4], are relying on the clouds for processing and analyzing their petabyte-scale data sets,

with a computing framework such as MapReduce and Hadoop [5], [6]. For example, Facebook-like social media sites collect their Web server logs, Internet click data, social activity reports, etc., from various locations over time, and parse them using Mapreduce/Hadoop to uncover usage patterns and hidden correlation, in order to make possible marketing decision.While most efforts have been devoted to designing better computing models for big data analytics, an important issue has largely been left out in this respect: How does one move the massive amounts of data to a cloud, in the very first place? The current observe is to copy the data into large hard drives for physically transportation to the data centre [7], [8], or even to move entire machines [9].

Such physical transportation incurs unwanted delay and possible service downtime. It is also less secure, and is prone to infection of malicious programs and damages from road accidents. Therefore, a safer and more flexible data migration approach is in need, to minimize any potential service downtime.

A. Geo-Distributed Data Centre's

Data centers that are spread at different geographic regions, e.g., Google's 13 data centres over 8 countries in 4 continents[1].the main objective of this paper is to place these data chunks in the server to distribute tasks onto servers. Second step is to move data between data centres. In

Step 1 Volley system [2] is used that chooses the data centre to place the data chunks into the server.

Step 2 chooses the optimized server to process the data that is stored in the data centre. In step data chunk is processed in the first data centre and the output is passed as an input to the next data centre.

Likewise all the data chunks are processed in different geographically distributed data centre. Many hard works have been made to minimize the computation or data movement cost of data centres. Data centre resizing (DCR) has been projected to minimize the processing cost by adjusting the number of activated servers through task placement [3].

Although there are some advantages, it is far from achieving the cost efficient big data processing because of the following weaknesses like data locality may result in waste of resources.

The link in network may vary on transmission rate and costs. To overcome above problems, we study the cost minimization problem of big data computation through joint optimization of data assignment, data placement, and data movement in geographically distributed data centres. Servers are equipped with limited storage and computation resources. Our objective is to optimize the big data placement, task assignment, routing and DCR such that the overall computation and communication cost is minimized.

II. Literature Survey

Most existing studies in literature on geographical request allocation subject to diverse electricity prices and bandwidth costs focused on solving a constrained optimization problem with one or multiple constraints. However, finding an exact solution usually takes a much longer time due to highly computational complexity. The solution based on the MIP thus is only suitable for a small or medium network size and not scalable. Even if such a solution is found, it may not be applicable in the reality due to time varying nature of both electricity prices and request rates in the system. A series of recent work studies application migration to the cloud.

Develop an optimization model for migrating enterprise IT applications onto a hybrid cloud. Advocate deploying social media applications into clouds, for leveraging the rich resources and pay-as-you-go pricing. These projects focus on workflow migration and application performance optimization, by carefully deciding the modules to be moved to the cloud and the data caching/replication strategies in the cloud. The very first question of how to move large volumes of application data into the cloud is not explored. Few existing work discussed such transfer of big data to the cloud. Design Pandora, a cost-aware planning system for data transfer to the cloud provider, via both the Internet and courier services.

he same authors later propose a solution to minimize the transfer latency under a budget constraint. Different from our study, they focus on static scenarios with a fixed amount of bulk data to transfer, rather than dynamically generated data; in addition, a single cloud site is considered, while our study pays attention to multiple data centres.A number of online algorithms have been proposed to address different cloud computing and data centre issues. For online algorithms without future information, investigate energy-aware dynamic server provisioning, by proposing a Lazy Capacity Provisioning algorithm with a 3-competitive ratio. Investigate load balancing among geographically-distributed data centres with a receding horizon control (RHC) algorithm, and show that the competitive ratio can be reduced substantially by leveraging the predicted future information.

A. Minimizing Server cost

Numerous of large-scale data centres are deployed providing services to large users. As suggested, a data centre may contain big servers and guzzle high power. Millions of dollars cost on electricity have cause a serious trouble on the operating cost to data centre providers. Therefore, minimizing the electricity cost has established major attention from both academia and industry. Data Centre Resizing and data placement are generally jointly measured to match the processing requirement.

Suggest the best workload control by taking account of latency, energy expenditure and electricity cost.

B. 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. Use of flexibility in the data placement policy to boost energy efficiency in data centres and propose a scheduling algorithm.

III. PROPOSED SYSTEM

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.

Geo Distributed Data Centres

Geo distributed data centre means many data centres are geo graphically distributed and connected through the WAN. Recently many organizations move to this geo distributed data centre. They stored large or massive volume of data's. If they are using our own data centre means only limited storage will be there so only many of them used this geo distributed data centres.

A. Data Placement

The data placement is another big issue in the geo distributed data centres. Because where the data's are placed in the servers and how they can be accessed and calculate the latency time of that particular data transition and migrate user data to the closest data centre. However, the simple heuristic ignores two major sources of cost to data centre operators: WAN bandwidth between data centres, and over-provisioning data centre capacity to tolerate highly skewed data centre utilization. In this paper, we show that a more sophisticated approach can both dramatically reduce these costs and still further reduce user latency.

B. Electricity cost

The electricity cost another burden in geo-distributed data centres. Because more energy will be using in data centres. All the hardware's work without electricity. Proposed a novel, data-centric algorithm used to reduce energy costs and with the guarantee of thermal-reliability of the servers in geo distributed data centres. And also using the n- dimensional markov chain algorithm to reduce the electricity cost.

C. Server cost

In geo distributed data centres hundred's of servers used. Because of this automatically the server cost will be increases. How to reduce the server cost means using communications and data placement and task assignment approach.

Number of sever will be reduced means at a mean time the energy cost also decrease. Server cost reduced using the joint optimization of these three factors such as task assignment, data placement and data routing via a n-dimensional markov chain. To efficiently manage the Data centre resizing. Proposed the optimal workload and balancing of latency, electricity prices and the energy consumption.

D. Cloud System

Consider a cloud consisting of number of geo-distributed data centers

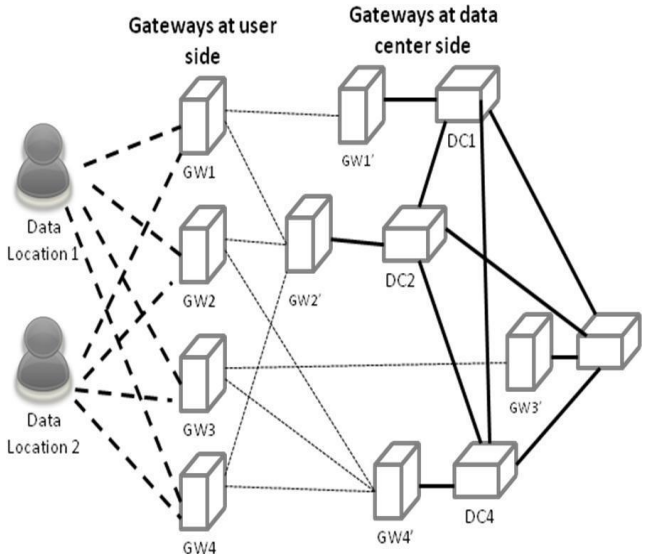

Figure 1: Cloud System

Cloud System is illustrated as shown in above fig. 1 in which a cloud user continuously produces large volumes of data of multiple geographic locations. The user connects to the data centers from different data generation locations. Inter data center connections within a cloud are usually dedicated high bandwidth lines.

IV. METHODOLOGY

A. Algorithms

A fast approximate solution is proposed to the minimum operational cost problem in a distributed cloud computing environment, through a reduction to the minimum cost multi commodity flow problem in Gf , where there are M commodities to be routed from their source nodes WPj to a common destination node t0 with demands rj , represented by a triple (WPj, t0; rj) for all j with $1 \leq j \leq M$ such that the system throughput is maximized while the associated operational cost is minimized, where $0 \leq \lambda \leq 1$.

a. Offline Algorithm

The offline optimization problem of minimizing the overall cost of data upload and processing over a time interval. We first present an offline algorithm, which derives the theoretical minimum cost given complete knowledge of data generation in both temporal and spatial domains.

b. Online Algorithm

A straightforward algorithm solves the above optimization in each time slot, based on the previous time slot. This can be far from optimal due to premature data migration. For example, assume data centre k was selected at t 1, and migrating data from k to j is cost-optimal at t according to the one-shot optimization (e.g., because more data are generated in region j in t); the offline optimum may indicate to keep all data in k at t, if the volume of data generated in k in t + 1 surges. We next explore dependencies among the selection of the aggregation data centre across consecutive time slots, and design a more judicious online algorithm accordingly. It avoids aggressive switches of the aggregation data centre, to prevent moving a large amount of data back and forth too often.

Excessive "laziness" is also avoided. Parameters 2 > 0 and 1 > 0 control the "laziness" and "aggressiveness" of the algorithm: a large 2 prolongs the inter-switch interval, while a large 1 invites more frequent switches.

V. EXPERIMENTAL ANALYSIS

The performance of the algorithms is evaluated under the default, fixed prices for data storage and processing in the data centers.

In Table I, *RFHC*(*x*) represent a RFHC Algorithm with look ahead window $l = x$. As compared to the simple algorithm,

TABLE I

CONSTANT PRICES, $P = 0.25$, $L = 0.01$

	Simple	OLM	RFHC(0)	Offline	Proposed
Overall Cost	20577	17264	18503	16712	16204
Ratio	1023	1.03	1.11	1	1.005

All proposed algorithms achieve considerable cost reduction and performance of the OLM algorithm and the RFHC algorithm with $l = 1$ approaches that of the offline optimum rather well. Then assume 100% prediction accuracy after investigating the following two considerations:

i) Spot instance pricing, P=0.5, L=0.01

ii) Spot instance pricing P=0.25, L=0.01

TABLE II

SPOT INSTANCE PRICING, $P = 0.25$, $L = 0.01$

	Simple	OLM	RFHC(0)	Offline	Proposed
Overall Cost	24709	16780	16676	16327	16204
Ratio	1.55	1.06	1.05	1.02	1.005

Thus predicted input is generated with error through adjusting the accurate input.

TABLE III

CONSTANT PRICES, $P = 0.25$, $L = 0.01$

	Simple	OLM	RFHC(0)	Offline	Proposed
Overall Cost	18503	17237	17050	16828	6778
Ratio	1.11	1.03	1.02	1.01	1.003

The above table describes the performance of RFHC with different Look ahead window sizes. Thus when error rate is adequately low, the cost is very close to the one with 100% accurate prediction.

CONCLUSION

Previous technologies developed an optimization model for migrating enterprise IT applications on a hybrid cloud. They focus on workflow migration and application performance optimization by deciding modules to be moved to the cloud. And also they focus on static scenarios with fixed amount of bulk data transfer instead of considering dynamically generated data. This paper presents an online approach for moving big data to the cloud which minimizes the cost of data transfer.

ACKNOWLEDGMENT

The authors would like to acknowledge Computer Engineering department, SITRC and all the people who provided with the facilities being required and conductive conditions for completion of the review paper.

REFERENCES

[1] Lin Gu, Deze Zeng, Peng Li, and song guo, "Cost Minimization for Big Data Processing in Geo-Distributed Data Centers" pp. 2168-6750, 2013.

[2] 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, pp. 33–36, 2013.

[3] 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.

[4] 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 27thInternational Symposium on Parallel Distributed Processing (IPDPS), pp. 623– 634, 2013.

[5] L Y. Wu, C. Wu, B. Li, L. Zhang, Z. Li, and F. Lau, "Scaling SocialMedia Applications into Geo-Distributed Clouds," in *Proc. IEEE INFOCOM*, Mar. 2012.

[6] M. Lin, Z. Liu, A. Wireman, and L. Andrew, "Online Algorithms For Geographical Load Balancing," in *Proc. IEEE IGCC*, 2012.

[7] T. Lu and M. Chen, "Simple and Effective Dynamic Provisioning for Power-Proportional Data Centers," in *Proc. IEEE CISS*, Mar. 2012.

[8] Mathew R. Sitaraman and P. Shenoy, "Energy-aware Load Balancing In Content Delivery Networks" Proc. IEEE INFOCOM, Mar 2012.

[9] Z. Liu, Y. Chen, C. Bash, A. Wierman, D. Gmach, Z. Wang, 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, pp. 175–186, 2012.

[10] R. Kaushik and K. Nahrstedt, "T*: A data-centric cooling energy costs reduction approach for Big Data analytics cloud," in 2012International Conference for High Performance Computing, Networking, Storage and Analysis (SC), pp. 1–11, 2012.

[11] M. Cardosa, C. Wang, A. Nangia, A. Chandra, and J. Weissman, "Exploring MapReduce Efficiency with Highly-Distributed Data", in *Proc. ACM MapReduce*, 2011.

[12] B. Cho and I. Gupta, "Budget-Constrained Bulk Data Transfer Via Internet and Shipping Networks," in *Proc. ACM ICAC*, 2011.

[13] M. Lin, A. Wierman, L. L. Andrew, and E. Thereska, "Dynamic ightsizing for Power- proportional Data Centers," in *Proc. IEEE INFOCOM*, April 2011.

[14] X. Cheng and J. Liu, "Load-Balanced Migration of Social Media to Content Clouds," in *Proc. ACM NOSSDAV*, June 2011

[15] Liu, M. Lin, A. Wierman, S. H. Low, and L. L. Andrew, "Greening eographical Load Balancing," in Proceedings of International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS). ACM, pp. 233–244, 2011.

[16] B. Cho and I. Gupta, "New Algorithms for Planning Bulk Transfer Via Internet and Shipping Networks," in *Proc. IEEE ICDCS*, 2010.

[17] D. Logothetis, C. Olston, B. Reed, K. C. Webb, and K. Yocum, "Stateful Bulk Processing for Incremental Analytics," in *Proc. ACM SoCC*, 2010.

[18] S. Pandey, L. Wu, S. Guru, and R. Buyya, "A Particle Swarm Optimization (PSO)-based Heuristic for Scheduling Workflow Applications in Cloud Computing Environment," in *Proc. IEEE AINA*, 2010.

[19] E. E. Schadt, M. D. Linderman, J. Sorenson, L. Lee, and G. P. Nolan "Computational Solutions to Large-scale Data Management and Analysis," *Nat Rev Genet*, vol. 11, no. 9, pp. 647–657, 09 2010.

[20] M. Hajjat, X. Sun, Y. E. Sung, D. Maltz, and S. Rao, "Cloudward Bound: Planning for Beneficial Migration of Enterprise Applications to the Cloud," in *Proc. ACM SIGCOMM*, August 2010

AUTHOR'S PROFILE

Author's Name : Leena B. Dhangar

B.E.Computer (Savitribai Phule Pune University)
PG Student,
Savitribai Phule Pune University,
SITRC College,
Nashik-422213
E-mail id: leena.dhangar@gmail.com

Author's Name: Asst. Prof. Amitkumar Manekar

Work Experience 1.6 Yrs in Company and 5 Yrs in teaching PhD Pursuing in CSE as specialization in Big Data Analytic and Cloud Computing Domain. Manekar A.S. Malti Nalge, et al." Sifting Of A Potent Convex Hull Algorithm For Scattered Point Set Using Parallel Programming", 2013 5th International Conference on Computational Intelligence and Communication Networks, IEEE / DOI 10.1109 / CICN.2013.121