International Journal of Engineering Research-Online
A Peer Reviewed International Journal
Articles available online http://www.ijoer.in

Vol.3., Issue.5., 2015
(Sept.-Oct.)

RESEARCH ARTICLE

ISSN: 2321-7758

EFFORT DIMINISH FOR BIG DATA HANDLING IN DISTRIBUTED SERVER FARMS

S.THYLASHRI[1], C.SHANMUGANATHAN[2]

[1]II Year M.E., CSE Department of Computer Science & Engineering, St.Peter's College of Engineering & Technology, Chennai, India

[2]Assistant Professor, Department of Computer Science & Engineering, St.Peter's College of Engineering & Technology, Chennai, India

ABSTRACT

The explosive growth of demands on massive processing imposes an important burden on computation, storage, and communication in information centers, that therefore incurs respectable operational expenditure to information center suppliers. Cloud computing security is developing at a fast pace which incorporates laptop security, network security, data security, and information privacy. Cloud computing plays a really very important role in protective information, applications and also the connected Infrastructure with the assistance of policies, technologies, controls, and large information tools. Therefore, value reduction has become an associate in nursing emerging issue for the approaching massive information area completely different from standard cloud services, one in every of the most options of massive information services is that the tight coupling between information and computation as computation tasks may be conducted only the corresponding information is accessible. As a result, three factors, i.e., task assignment, information placement and information movement, deeply influence the operational expenditure of knowledge centers. In this paper, we tend to square measure impelled to review the price reduction downside via a joint improvement of those three factors for giant information services in geo-distributed information centers. To explain the task completion time with the thought of each information transmission and computation, we tend to propose a two-dimensional Markov process and derive the common task completion time in closed-form. What is more, we learn to model the matter as a Mixed-Integer Non-Linear Programming (MINLP) associate in propose an efficient resolution to line arise it.

Index Terms—big data, data flow, data placement, distributed data centers, cost minimization, task assignment

1. INTRODUCTION

In order to investigate advanced knowledge and to spot patterns it's vital to firmly store, manage and share massive amounts of advanced knowledge. Cloud comes with a certain security challenge, i.e. the information owner may not have any management of wherever the information is placed the rationale behind this management issue is that if one desires to induce the advantages of cloud computing, he/she should conjointly utilize the allocation of resources and conjointly the planning given by the controls. Thus it's needed to safeguard the information within the interior of undependable processes. Since cloud involves intensive quality, we tend to believe that instead of providing a holistic resolution to securing the cloud, it might be ideal to create noteworthy enhancements in securing the cloud that may ultimately offer U.S.A. with a secure cloud. Knowledge explosion in recent year's results in a rising demand for large processing in fashionable knowledge centers that square measure typically distributed at completely different geographic regions, e.g., Google's thirteen knowledge centers over eight countries in four continents [1]. Massive knowledge analysis has shown its nice potential in unearthing valuable insights of knowledge to enhance decision making, minimize risk and develop new product and services. On the opposite hand, massive knowledge has already translated into massive worth owing to its high demand on computation and communication resources [2]. Gartner predicts that by 2015, seventy one of worldwide knowledge center hardware defrayal can come back from the large processing, which can surpass $126.2 billion. Therefore, it's imperative to check the value step-down drawback for large processing in geo-distributed knowledge centers. Several efforts are created to lower the computation or communication price of knowledge centers. Knowledge center resizing (DCR) has been projected to scale back the computation price by adjusting the quantity of activated servers via task placement [3]. Supported DCR, some studies have explored the geographical distribution nature of knowledge centers and electricity worth non uniformity to lower the electricity price [4]–[6]. Massive knowledge service frameworks, e.g., [7], comprise a distributed file system beneath, that distributes knowledge chunks and their replicas across {the knowledge the info the information} centers for fine-grained load-balancing and high parallel data access performance. to scale back the communication price, some recent studies build efforts to enhance knowledge vicinity by inserting jobs on the servers wherever the input file reside to avoid remote data loading [7], [8]. Though the on top of solutions has obtained some positive results, they're off from achieving the cost efficient massive processing owing to the subsequent weaknesses. First, knowledge vicinity might lead to a waste of resources. As an example, most computation resource of a server with less in style knowledge might keep idle. The low resource utility any causes a lot of servers to be activated and thus higher overhead. Second, the links in networks vary on the transmission rates and prices in step with their distinctive options [9], e.g., the distances and physical optical fiber facilities between knowledge centers. However, the prevailing routing strategy among knowledge centers fails to use the link diversity of knowledge center networks. Owing to the storage and computation capability constraints, not all tasks are placed onto identical server, on that their corresponding knowledge reside. It's inescapable that sure knowledge should be downloaded from a far off server. During this case, routing strategy matters on the transmission price. The lot of link used, the upper price is going to be incurred. Therefore, it's essential to lower the quantity of links used whereas satisfying all the transmission needs. Third, the Quality-of-Service (QoS) of huge knowledge tasks has not been thought-about in existing work. Almost like typical cloud services, massive knowledge applications conjointly exhibit Service-Level-Agreement (SLA) between a service supplier and also the requesters. To watch SLA, an explicit level of QoS, typically in terms of task completion time, shall be bonded. The QoS of any cloud computing tasks is first determined by wherever {they square measure placed and the way several computation resources are allocated. Besides, the transmission rate is another influential issue since massive knowledge tasks square measure knowledge-centric and also the computation task cannot proceed till the corresponding data square measure on the market. Existing studies, e.g., [3], on general cloud computing tasks chiefly specialize in the computation capability constraints, whereas ignoring the constraints of transmission rate. To conquer higher than weaknesses, we have a tendency to study the price minimization drawback for large processing via joint optimization of task assignment, information placement, and routing in geo-distributed information centers. Specifically, we have a tendency to contemplate the subsequent problems in our joint optimization. Servers area unit equipped with restricted storage and computation resources.

Every information chunk incorporates a storage demand and can be needed by huge information tasks. The info placement and task assignment area unit clear to the info users with bonded QoS. Our objective is to optimize the large information placement, task assignment, routing and DCR such the general computation and communication value is decreased. Our main contributions area unit summarized as follows:

• To our greatest data, we have a tendency to area unit the primary to contemplate the price minimization drawback of massive processing with joint thought of knowledge placement, task assignment and information routing. To explain the rate-constrained computation and transmission in huge method process, we have a tendency to propose a two-dimensional Mark off process and derive the expected task completion time in closed type.

• supported the closed-form expression, we have a tendency to formulate the price minimization drawback in an exceedingly type of mixed integer nonlinear programming (MINLP) to answer the subsequent questions: 1) a way to place these information chunks within the servers, 2) a way to distribute tasks onto servers while not violating the resource constraints, and 3) a way to size information centers to realize the operation value minimization goal.

• To handle the high process complexness of finding MINLP, we have a tendency to correct it as a mixed-integer linear programming (MILP) drawback, which might be resolved mistreatment business thinker. Through in depth numerical studies, we have a tendency to show the high potency of our projected joint-optimization based mostly algorithmic rule.

2 RELATED WORKS

2.1 Server Cost Diminish

Large-scale knowledge centers are deployed everywhere the planet providing services to many thousands of users. in step with [11], a knowledge center might accommodates giant numbers of servers and consume megawatts of power. Lots of bucks on electricity price have posed an important burden on the overhead to knowledge center suppliers.

Therefore, reducing the electricity price has received vital attention from each world and trade [5], [11]–[13]. Among the mechanisms that are projected thus far for knowledge center energy management, the techniques that attract ample attention square measure task placement and DCR. DCR and task placement square measure typically together thought of to match the computing demand. Liu et al. [4] canvas identical drawback by taking network delay into thought.

The study power provisioning ways on what quantity computing instrumentation may be safely and expeditiously hosted at intervals a given power budget. Rao et al. [3] investigate the way to scale back electricity price by routing user requests to geo-distributed knowledge centers with consequently updated sizes that match the requests. Recently, Gao et al. [14] propose the optimum employment management and leveling by taking account of latency, energy consumption and electricity costs.

2.2 Huge knowledge Management

To tackle the challenges of effectively managing huge knowledge, several proposals are projected to boost the storage and computation method. The key issue in huge knowledge management is reliable and effective knowledge placement propose a programming formula, that takes under consideration energy potency additionally to fairness and knowledge vicinity properties, The [17] propose a mechanism permitting coupled open knowledge to require advantage of existing large-scale knowledge stores to fulfill the wants on distributed and parallel processing.

Moreover, the way to apportion the computation resources to tasks has conjointly drawn a lot of attention. Cohen et al. [18] gift new style philosophy, techniques and skill providing a replacement magnetic, agile and deep knowledge analytics for one in every of the world's largest advertising networks at Fox Audience Network, victimization the Green plum parallel info system. All [19] propose a unique, data-centric formula to scale back energy prices and with the guarantee of thermal-reliability of the servers. The [20] take into account the matter of together programming all 3 phases, i.e., map, shuffle and scale back, of the Map Reduce method and propose a sensible experiential to combat the high programming complexness.

2.3 Information Placement

In this [21] investigate how to determine a placement of Video-on-Demand (VoD) file copies 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. The [22] propose an automated data placement mechanism Volley for geo-distributed cloud services with the consideration of WAN bandwidth cost, data center capacity limits, data inter-dependencies, etc. Cloud services make use of Volley by submitting logs of datacenter requests. Volley analyzes the logs using an iterative optimization algorithm based on data access patterns and client locations, and outputs migration recommendations back to the cloud service. Cidon et al. [23] invent MinCopysets, a data replication placement scheme that decouples data distribution and replication to improve the data durability properties in distributed data centers.

Recently, the [10] propose a joint optimization scheme that simultaneously optimizes virtual machine (VM) placement and network flow routing to maximize energy savings. Existing work on data center cost optimization, big data management or data placement mainly focuses on one or two factors. To deal with big data processing ingeo-distributed data centers, we argue that it is essential to jointly consider data placement, task assignment and data flow routing in a systematical way.

3 SYSTEM MODEL

In this section, we tend to introduce the system model. For the convenience of the readers, the most important notations

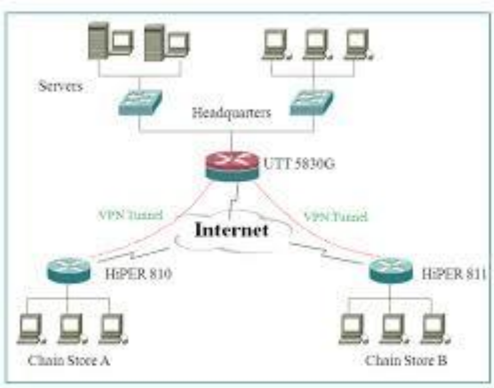

Fig. 1. Data center topology

3.1 Network Model

We think about a geo-distributed knowledge center topology as shown in Fig. 1, within which all servers of constant knowledge center (DC) square measure connected to their native switch, whereas knowledge centers square measure connected through switches. There square measure a group I of knowledge centers, and every knowledge center $i \in I$ consists of a group Islamic Group of servers that square measure connected to a switch $m_i \in M$ with a neighborhood transmission price of CL. In general, the transmission price Cr for inter-data center traffic is bigger than CL, i.e., CR > CL.

While not loss of generality, all servers within the network have constant computation resource and storage capability, each of that square measure normalized to 1 unit. We tend to use J to denote the set of all severs, i.e., $J = J1 \cup J2 \cdots \cup J|I|$. the total system are often sculptured as a directed graph $G = (N,E)$. The vertex set $N = M \cup J$ includes the set M of all switches and therefore the set J of all servers, and E is that the directional edge set.

All servers square measure connected to, and solely to, their native switch via intra-data center links whereas the switches square measure connected via inter-data center links determined by their physical association. the burden of every link $w(u;v)$, representing the corresponding communication price, are often outlined as $w(u;v) = \{Cr$, if $u, v \in M$, CL, otherwise. (1)3.2 Task Model we think about huge knowledge tasks targeting on knowledge hold on during a distributed classification problem in a mixed-integer nonlinear programming form. System that's engineered on geo-distributed knowledge centers. The info square measure divided into a group K of chunks. Every chunk $k \in K$ has the scale of $\varphi k(\varphi k \leq 1)$, that is normalized to the server storage capability. P-way replication [19] is employed in our model. That is, for every chunk, there square measure precisely P copies hold on within the distributed classification system for resiliency and fault-tolerance. It's been wide in agreement that the tasks arrivals at knowledge centers throughout a period are often viewed as a Poisson method [9], [24]. especially, let λk be the typical task arrival rate requesting chunk k.

Since these tasks are distributed to servers with a set chance, the task arrival in every server are often additionally considered a Poisson method. we tend to denote the typical arrival rate of task for chunk k on server j as $\lambda jk(\lambda jk \leq 1)$. Once a task is distributed to a server wherever its requested knowledge chunk doesn't reside, it must look forward to the info chunk to be transferred. Every task ought to be responded in time D.

Moreover, in sensible knowledge center management, several task declaration mechanisms supported the historical statistics are developed and applied to the choice creating in knowledge centers [19]. to stay the info center settings up-to-date, knowledge center operators might create adjustment consistent with the task declaration amount by amount [3], [14], [15]. This approach is additionally adopted in this paper.

4 PROBLEM FORMULATIONS

In this section, we first present the constraints of data and task placement, remote data loading, and QoS. Then, we give the complete formulation of the cost minimization

4.1 Constraints of Data and Task Placement

We define a binary variable *yjk*to denote whether chunk*k* is placed on server *j* as follows,*yjk*={1, if chunk *k* is placed on server *j*,0, Otherwise.

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

$$\sum_{j\in J}^{n} yjk \ \cdot = P, \forall k \in K. \quad (3)$$

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

$$\sum_{yjk}^{n} yjk \cdot \phi k \le 1, \forall j \in J. (4)$$

As for task distribution, the sum rates of task assignedto each server should be equal to the overall rate,

$$\lambda k = \sum_{j\in J}^{n} y\lambda jk, \forall k \in K. (5)$$

Finally, we define a binary variable *xj*to denote whether server *j* is activated, i.e., *xj*={1, if this server is activated,0, otherwise.

4.2 Constraints of Data Loading

Note that when a data chunk *k* is required by a server *j*, it may cause internal and external data transmissions. This routing procedure can be formulated by a flow model. All the nodes *N* in graph *G*, including the servers and switches, can be divided into three categories:

- Source nodes *u*(*u ∈J*). They are the servers with chunk *k* stored in it. In this case, the total outlet flows to destination server *j* for chunk *k* from all source nodes shall meet the total chunk requirement per time unit as *λjk· ϕk*.
- Relay nodes *mi*(*mi ∈M*). They receive data flows from source nodes and forward them according to the routing strategy.
- Destination node *j*(*j ∈J*). When the required chunkis not stored in the destination node, i.e., *yjk*= 0,it must receive the data flows of chunk *k* at a rate *λjk· ϕk*.

4.3 Constraints of QoS Satisfaction

Let *μjk* and *γjk* be the processing rate and loading rate for data chunk *k* on server *j*, respectively. The processing procedure then can be described by a two-dimensional Markov chain, where each state (*p, q*) represent s*p* pending tasks and *q* available data chunks. We let *ϑjk* denote the amount of computation resource (e.g., CPU) that chunk *k* occupies. The processing rate oftasks is proportional to its computation resource usage ,i.e., *μjk*= *αj* · *ϑjk, ∀j ∈J, k ∈K,* (11)

Where *αj* is a constant relying on the speed of server *j*.

4.4 An MINLP Formulation

The total energy cost then can be calculated by summing up the cost on each server across all the geo-distributed data centers and the communication cost, i.e.,

MINLP:

min :

xj ,yjk, zjk, ujk∈{0, 1}*, ∀j ∈J, k ∈K*

where *Pj* is the cost of each activated server *j*.

5 PERFORMANCE EVALUATIONS

The processing speed depends on the quantity of records being gotten to in lesser time. The large amount of information is in this way being put away into the distributed servers as opposed to the traditional server. This shows extensive upgrades as far as time utilization for processing of this vast information

Fig. 2 shows the result after performance evaluation. The x-axis gives the number of arrival tasks to the server in a given measure of time and the y-axis indicates the cost of each task that has touched base to the server. The diagram gives the reasonable shot of the cost taken by the server relying upon their arrival tasks to the server.

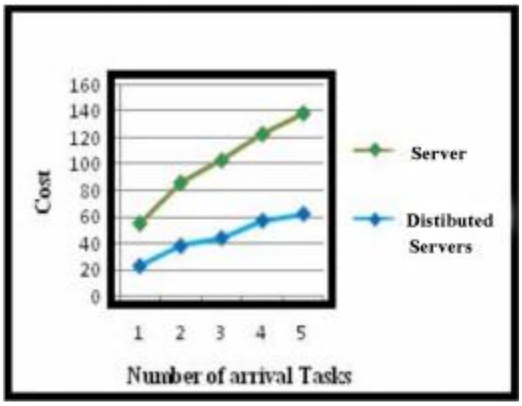

Fig.2. Analysis graph of cost exploitation for various servers

The next graph i.e. Fig. 3, the correlation between processed records got to against the time or the time taken for processing these records. The graph clearly indicates how the distributed servers are ended up being quicker when contrasted with that of the traditional server.

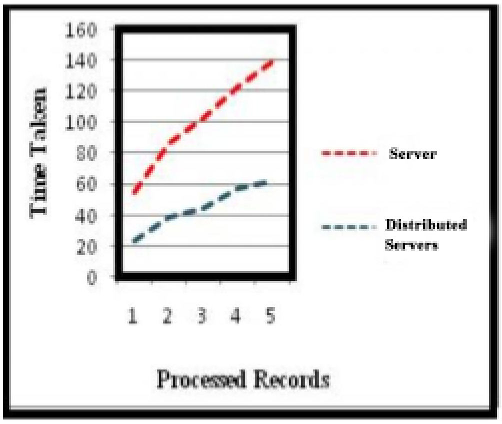

Fig. 3.Analysis graph for Time Consumption of different servers

6.CONCLUSION

This paper, minimizes the cost that happens along with the enormous information considering so as to handle two fundamental huge information administration issues i.e., information placement and task assignment by proposing the MILP. We have a tendency to conjointly study the data placement, task assignment, knowledge center resizing and routing to reduce the operational value in large-scale geo-distributed knowledge centers for giant knowledge applications. we have a propensity to initial characterize the info process method employing a two-dimensional Mark off process and derive the expected completion time in closed-form, supported that the joint improvement is developed as associate degree MINLP drawback. To tackle the high machine complexness of finding our MINLP, we have a tendency to set it into associate degree MILP drawback. Through intensive experiments, we have aaffinity to show that our joint-optimization answer has substantial advantage over the approach by ballroom dance separate improvement. Many fascinating phenomena are determined from the experimental results.

7 REFERENCES

[1]. Lin Gu, Deze Zeng, Peng Li, and Song Guo "Cost minimization for big data processing in geo-distributed data centers" IEEE Transactions on emerging topics in computing, volume 2, no.3, September 2014.

[2]. (2013). Data Center Locations [Online]. Available:http://www.google.com/about/datacenters/inside/locations/index.html.

[3]. M.Sathiamoorthyet al., ''Xoring elephants: Novel erasure codes forbig data,'' in Proc. 39th Int. Conf. Very Large Data Bases, 2013,pp. 325–336.

[4]. H. Shachnai, G. Tamir, and T. Tamir, ``Minimal cost reconfiguration of data placement in a storage area network,'' Theoretical Comput. Sci., vol. 460,pp. 42-53, 2012

[5]. 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

[6]. F. Chen, M. Kodialam, and T. V. Lakshman, ``Joint scheduling of processing and shuffle phases in mapreduce systems,'' in Proc. 29th Int. Conf.Comput. Commun. 2012, pp. 1143-1151.

[7]. 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.

[8]. S. Agarwal, J. Dunagan, N. Jain, S. Saroiu, A. Wolman, and H. Bhogan,"Volley: Automated Data Placement for Geo-Distributed Cloud Services,in The 7th USENIX Symposium on

Networked Systems Design and Implementation (NSDI), 2010, pp. 17–32
[9]. 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 ComputerSystems (SIGMETRICS). ACM, 2011, pp. 233–244.
[10]. J. Dean and S. Ghemawat, "Mapreduce: Simplified data processing on large clusters," Commn. ACM, vol. 51, no. 1, pp. 107-113, 2008.
[11]. "IBM What Is Big Data: Bring Big DataEnterprise,http://www01.ibm.com/software/data/bigdata/, IBM, 2012
[12]. A. Rajaraman and J. Ullman, Mining of Massive Data Sets.Cambridge Univ. Press, 2011
[13]. J. Cohen, B. Dolan, M. Dunlap, J. M. Hellerstien and C. Welton, "Mad skills: New analysis practices for big data," Proc. VLDB Endowment, vol. 2, no. 2, pp. 1481_1492, 2009.
[14]. 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.
[15]. R. Raghavendra, P. Ranganathan, V. Talwar, Z. Wang, and X. Zhu, ''No'power' struggles: Coordinated multi-level power management for the datacenter,'' in Proc. 13th Int. Conf. Archit. Support Program. Lang. Oper. Syst.,2008, pp. 48–59.