IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

Cost Reduction for Big Data Processing In Physically Dispersed Data Centers

Miss. Sayyad K. M.[1], Mr. Gurav R. R.[2], Mr. Salunke S. S.[3], Miss. Mirgane P. V.[4], Prof. Shaikh N. K.[5]

Student, Computer Science & Engg, Bhagwant Institute of Technology, Barshi, India[1,2,3,4]

Assistant Professor, Computer Science & Engg, Bhagwant Institute of Technology, Barshi, India[5]

Abstract: Demand on big data is being rising day by day and also growing heavy burden on computation, storage and communication in data centers, which cause significant expenses to data center providers. So, cost reduction became an issue for the upcoming big data. One of the primary feature of big data is coupling of data and computation as computation assignment. Three obligations like data placement, task assignment and data movement impact the rate of facts centers. In this paper we study how to reduce the cost using joint optimization of these above three factors for big data service in geographically spreaded data centers. Right here we recommend 2-d markov chain to describe time to finish a particular undertaking with consideration of data transmission and computation to derive average challenge finishing touch time in closed time. The problem with the mixed integer nonlinear programming solved by linearizing it.

Keywords: Big data, markov chains, data centers, nonlinear programming, geo-distributed.

I. INTRODUCTION

In earlier years data explosion is leading for increasing demand in big data processing which are spreaded among different geographical centers. For example Google have 13 data centers over 8 countries in 4 continents [1]. Massive statistics analysis have already shown its potential to enhance decision-making, hazard minimization and to develop new products and services as large statistics has greater computation and communication sources, it reached excessive cost. [2]. 71% of worldwide data center hardware expenses will come from the big data processing, which is near to $126.2 billion which was surveyed by Gartner.

Therefore there is an immense need in focusing about cost reduction for big data processing in geographically spreaded centers. Many efforts have been made to minimize cost for data processing. To decrease the computation cost the number of activated servers are adjusted via task placement which is proposed by Data Center Resizing (DCR) [3]. In mild of DCR, some studies have investigated the geological appropriation nature of server farms and power cost heterogeneity to lower the energy price. The truth that the above arrangements have received a few fantastic results, they're a long way from carrying out the rate efficient good sized facts getting ready because of the accompanying shortcomings.

Second, the links in networks vary on the transmission rates and expenses according to their one of its kind of features [9], e.g., the distances and physical optical fiber services between data centers. However, the presented steering strategy among data centers fails to exploit the link variety of data center networks. Due to the storage and computation capacity constraints, not all tasks can be located onto the same server, on which their analogous data reside. It is inevitable that certain data must be downloaded from a remote server.

Third, the Quality-of-Service (QoS) of huge information everyday jobs has not been considered in existing work. Like customary cloud administrations, massive information applications likewise show Service-Level-Agreement (SLA) between an administration supplier and the requesters. To watch SLA, a specific level of QoS, normally in so far as undertaking conclusion time, should be ensured.

The QoS of any allotted computing undertakings is dictated through where they may be put and how many calculation belongings are apportioned. In addition, the transmission rate is another component since massive information assignments are information driven and the computation undertaking can't continue until the comparing information are reachable.

We have linearized whole number straight programming (MILP) to manage the high computational anomaly of solving MINLP. Through wide numerical concentrates on, we reveal the high efficiency of our projected joint-improvement based computation.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

II. RELATED WORK

2.1 Server Cost Minimization

Large-scale data centers have been spreaded all over the world providing services to hundreds of thousands of users. In keeping with [11], a data center can also consist of massive numbers of servers and eat megawatts of power[5], [11] [13]. Hundreds of thousands of bucks on energy price have posed a heavy burden at the running price to data center companies. Consequently, decreasing the electricity cost has acquired extensive interest from each academia and industry. Among the mechanisms which have been proposed to this point for facts center power control, the techniques that attract masses of interest are task assignment and DCR. DCR and task placement are usually jointly measured to match the computing requirement. Liu et al. [4] surveyed the similar problem by giving attention to network delay. Fan et al. [12] examine energy provisioning techniques on how lots computing gadget may be adequately and efficiently hosted within a given power price range

2.2 Big Data Management

To tackle the challenges of efficiently managing big data, many proposals were proposed to enhance the storage and computation procedure. The important thing with difficulty in big data control is reliable and powerful data placement. To achieve this goal, Sathiamoorthy et al. [16] present a novel family of erasure codes that are efficiently repairable and offer higher reliability in comparison with Reed-Solomon codes. They also analytically show that their codes are optimal on an identified tradeoff between locality and minimum distance. Yazd et al. [8] takes advantage of flexibility in the data block placement policy to increase energy efficiency in data centers and with consideration of energy efficiency in addition to fairness and data locality properties he proposed scheduling algorithm. Hu et al. [17] propose a mechanism allowing linked open data to take advantage of existing large-scale data stores to meet the requirements on distributed and parallel data processing.

2.3 Data Placement

Shachnai et al. [19] explored how to find out a assignment of Video-on-Demand (VoD) file copies on the servers and the amount of load capability given to each file copy so as to reduce the communique cost while ensuring the user experience. Agarwal et al. [20] put forward an automated data placement mechanism Volley for geo-located cloud services with the consideration of WAN bandwidth cost, data center capability limits, data inter-dependencies, etc. Cloud services uses Volley by submitting logs of data center desires. With the help of iterative optimization algorithm based on outputs migration, data access patterns and client locations recommendations back to the cloud service, the Volley analyzes the logs.

For data center cost optimization big data management or data placement mainly focuses on one or two factors. To contract with big data processing in geographically located data centers, we argue that it is fundamental to jointly consider data placement, task assignment and data flow routing in a systematical way.

III. SYSTEM MODEL

In this segment, we present the framework model.

3.1 Network Model

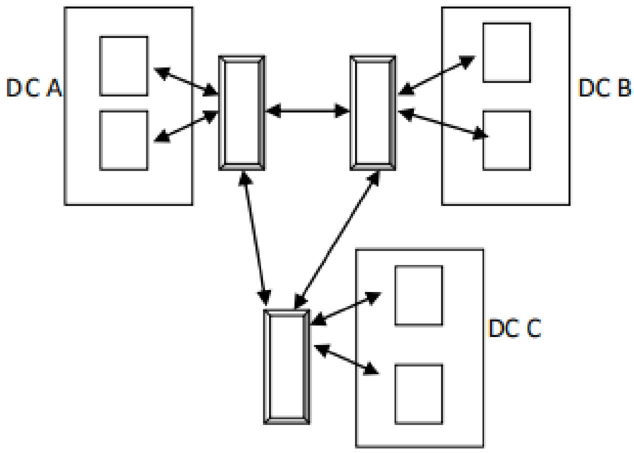

Fig -3.1: Data Center Network Model

DOI 10.17148/IJARCCE.2018.7367

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

The dispersed data centers are a system that spans many data centers at numerous places usually considered for storage. Every data center has many servers for storing and organization and retrieving data in data center. Task assignment, data loading and data migration are contained and supported by data center. The data centers A, B and C in the figure are connected with each other.

3.2 Task Model

We consider big data tasks targeting on data stored in a dispersed file system that is built on geographically spreaded data centers' of chunks are made of data. Each chunk k ∈ K has the size of $\Phi K(\Phi K \leq 1)$, which is normalized to the server storage ability. Our model uses P-way replica [19] there are Exactly P copies stored in the distributed file system for each chunk, for flexibility and fault-tolerance.

It has been broadly granted that the tasks coming at data centers during a time period can be viewed as a Poisson process [9], [21]. In particular, let λk be the usual task coming rate requesting chunk k. For the reason that those responsibilities can be disbursed to servers with a set opportunity, the undertaking arrival in each server may be additionally regarded as a Poisson manner. We denote the common arrival charge of project for bite ok on server j as $\lambda jk(\lambda jk \leq 1)$When a venture is shipped to a server in which its asked data chunk does not found, it desires to anticipate the information chunk to be transferred. Response should be in time D.

3.3 Markov Chain

A Markov chain is a process that contains fixed number of states and some well-known possibility having the property that, given the current state, the upcoming sate is self-governing of the past. The example of Markov chain is a simple random walk. A series of autonomous events satisfies the definition of Markov chain. To describe the rate inhibited figures and transmission in big data process, a two-dimensional Markov chain is applied and predicted task finishing time is calculated. The overall value may be calculated with the aid of summing up the fee on every server across all distributed data centers and this will be formulated as combined integer non-linear programming trouble and similarly it's far linearized to cope with high operational

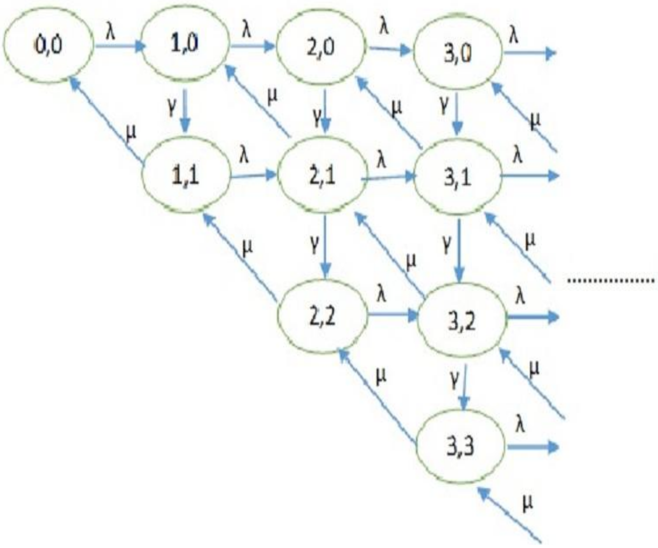

Fig 3.2: Markov Chain

IV. PROBLEM FORMULATION

Here, we former present the constraints of data and task placement, remote data loading, and QoS. Then, we give the absolute formulation of the cost minimization dilemma in a mixed-integer nonlinear programming form.

4.1 Constraints of Data and Task Placement

$$y_{jk}=\begin{cases} 1, \text{if chunk k is placed on server j,} \\ 0, \text{Otherwise.} \end{cases} \quad (1)$$

$$x_{j}=\begin{cases} 1, \text{if chunk k is placed on server j,} \\ 0, \text{otherwise} \end{cases} \quad (2)$$

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

4.2 Constraints of Data Loading

Here the nodes N in graph G which including switches as well as servers divided into three categories as follows:

- Source nodes u (u ∈ J).
- Relay nodes mi (mi ∈ M).
- Destination node j (j ∈ J).

The flow over the link (u, v) ∈ E carrying data of chunk k ∈ K which is meant to server j ∈ J, denoted by$f_{jk}^{(u,v)}$. The above three categories of nodes can be expressed as follows respectively with the constraints.

$$f_{jk}^{(u,v)} \leq Yuk.\lambda k.\phi_{k,}\ \forall\ (u, v) \in E,\ u, j \in J,\ k \in K \tag{3}$$

$$\sum_{(u,j)\epsilon E} f_{jk}^{(u,v)} - \sum_{(v,w)\epsilon E} f_{jk}^{(u,w)} = 0, \forall v\ \in\ M,$$

$$j\ \in\ J, k\ \ \in\ K. \tag{4}$$

$$\sum\nolimits_{(u,j)\epsilon E} f_{jk}^{(u,j)} = (1\ -\ yjk)\lambda jk\ \cdot\ \phi k, \forall j\ \in\ J, k\ \in\ K \tag{5}$$

4.3Constraints of QoS Satisfaction

$$u_{jk=}\begin{cases}1, \text{if this server is activated}\\ \quad\quad 0, \text{otherwise}\end{cases} \tag{6}$$

4.4 An MINLP Formulation

Total energy cost = cost on each server across all the geo-distributed data centers + communication cost

$C_{total} = \sum_{j\epsilon J} xj.Pj + \sum_{j\in J} \sum_{k\in K} \sum_{(u,v)\in E} f_{jk}^{(u,v)}.w^{(u,v)}$

MINLP: FORMULA

Min f(x,y)

St. Ci(x, y) = 0 ∀i∈E

Ci(x, y) ≤ 0 ∀i∈I

x ∈ X

y ∈ Y Integer

x_j , y_{jk}, z_{jk}, u_{jk} ∈ {0, 1}, ∀j ∈ J, k ∈ K

MINLP-2: FORMULA

s.t. :

$$\sum_{j\epsilon J} yjk\ =\ p, \forall k\ \epsilon\ K,$$

p ≥ 1,

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

$\delta jk = y_{jk}\lambda_{jk}$, ∀j ∈ J, ∀k ∈ K,

Now, mixed-integer linear programming (**MILP**) is linearized form of the **MINLP** problem.

MILP:

$$\begin{array}{rl} \min & \mathbf{c}^T\mathbf{x} \\ \text{subject to} & \mathbf{A}\mathbf{x}\ \{\geq,=,\leq\}\ \mathbf{b} \\ & \mathbf{l} \leq \mathbf{x} \leq \mathbf{u} \\ & \mathbf{x}_i \in \mathbb{Z}\quad \forall i \in \mathscr{S} \end{array}$$

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

V. PERFORMANCE EVALUATIONS

Now it’s turn to present the performance outcomes of our joint-optimization algorithm (“Joint”) which used MILP design. The isolated optimization arrangement algorithm (“Non-joint”), results in finding a least number of servers to be triggered and the traffic routing scheme by means of the network flow model is compared with “Joint”.

Different cost considered for performance evaluation

- Server cost
- Communication cost
- Overall cost
- With the above three cost we calculate the cost of geo distributed data center on the following basis:
- On the effort of the number of servers.
- On the effort of task arrival rate
- On the effort of data size
- On the effort of expected task completion delay
- On the effort of number of replica.

Consider,
Number of data center |J| = 3, with the same number of servers, Intra- and inter-data center link communication cost CL = 1 and CR = 4
The default settings are as follows:
Data center size= 20,
Number of data chunks |K| = 10,
Task arrival rates $\lambda k \in [0.01, 5]$, $\forall k \in K$,
Number of replicas P = 3,
Data chunk size $\phi k \in [0.01, 1]$, $\forall k \in K$, and
Delay D = 100.

On the effort of the number of servers:
Fig 5.1(a) shows that the server cost always keep constant on any data center size when server numbers varying from 36 to 60.Fig. 5.1(b), shows as servers increases from 36 to 48 communication costs of both algorithms decreases because more servers on same data center more tasks and data chunk can be placed on the same data center. After the number of server reaching 48, as increasing the number of servers will not affect the distributions of tasks or data chunks any more hence results the same as shown Fig 5.1 (c).

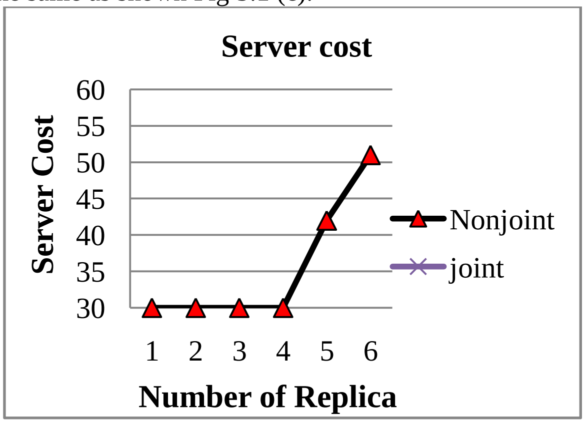

Fig.5.1 (a) Server Cost.

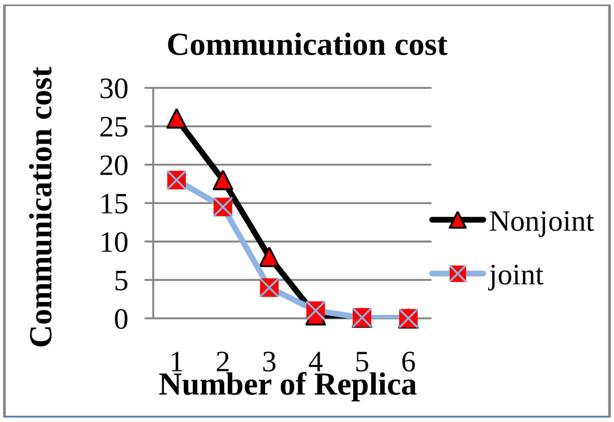

Fig.5.1 (b) Communication Cost

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

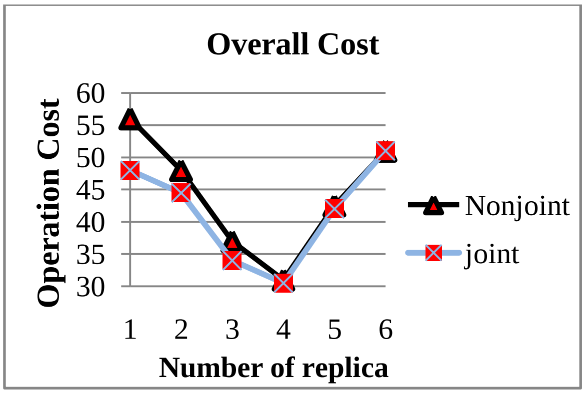

Fig.5.1 (c) Overall Cost

On the effort of task arrival rate

As the first phase of the "Non-joint" algorithm tries to lower the server cost hence Fig 5.2(a) shows "Joint" algorithm requires higher server cost than "Non-joint". The balance between server cost and communication cost is balanced by "Joint" algorithm hence lowers the complete cost as shown in Fig. 5.2(b) and Fig 5.2(c).

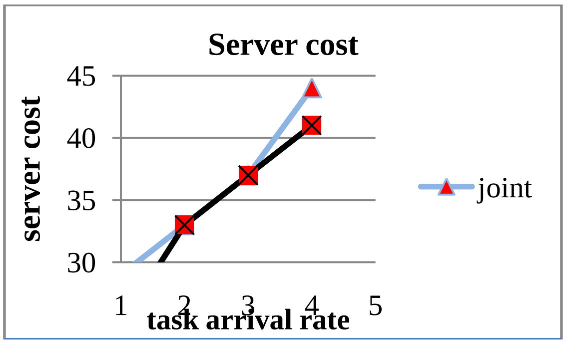

Fig 5.2 (a) Server cost

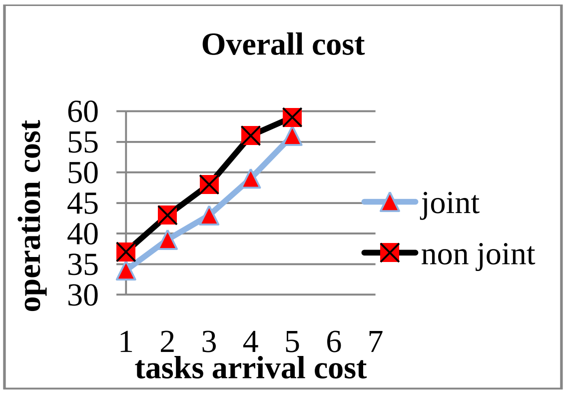

Fig.5.2 (b) Overall Cost

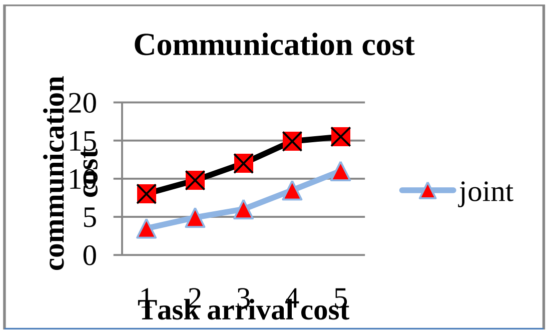

Fig.5.2 (c) Communication Cost

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

On the effort of data size
Larger chunk size (here data chunk size from 8.4 to 19) more servers need to be activated hence increases server cost as shown in Fig. 5.3(a).

Communication cost also increases when there is more traffic on data link, Fig 5.3(b).Fig. 5.3(c) shows increase in cost.

Fig. 5.3 (a).Server Cost.

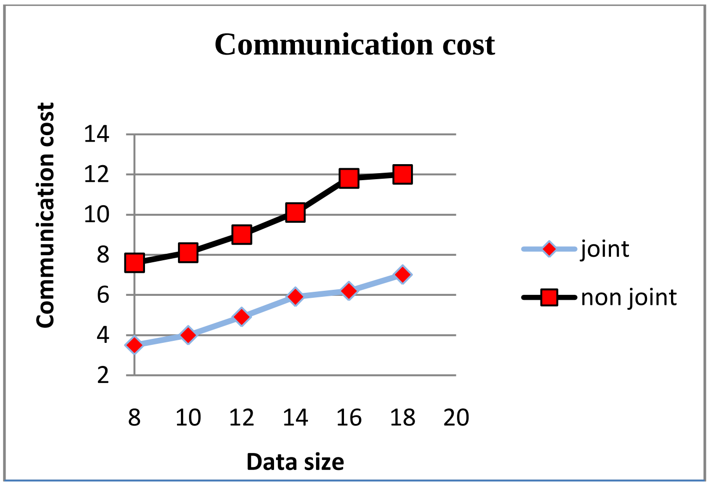

Fig. 5.3 (b) Communication Cost.

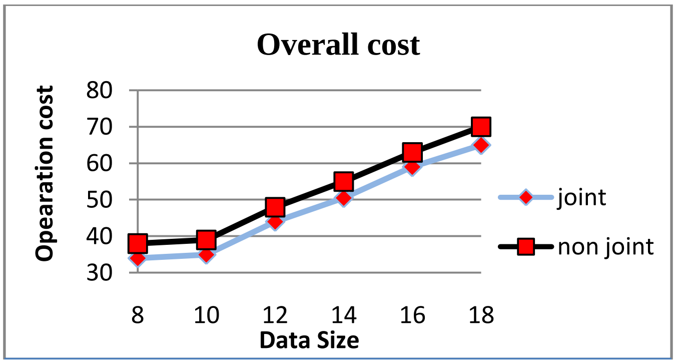

Fig 5.3 (c) Overall Cost.

On the effort of expected task completion delay.
Here response time D increases from 20 to 100.As delay constraints increases results in decrease in cost because less number of servers have to be activated to guarantee QoS shown by Fig 5.4(a).

A lesser QoS requirement also helps find cost effective routing policies as illustrated in Fig. 5.4(b). “Joint” over “Non-joint “can be always observed in Fig. 5.4(c).

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

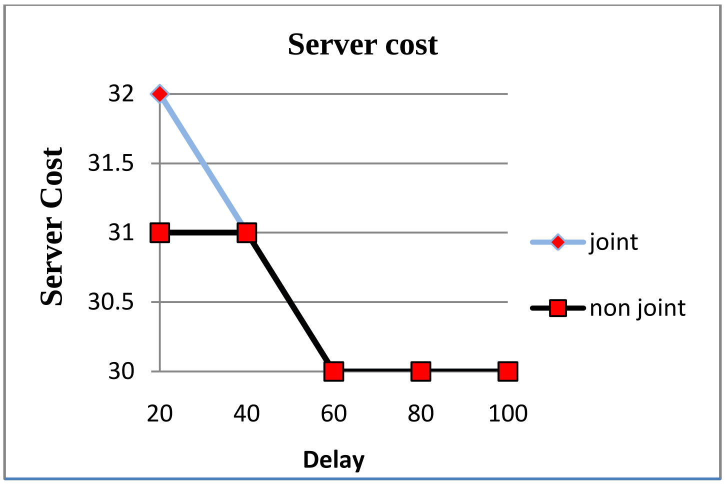

Fig.5.4 (a) Server Cost.

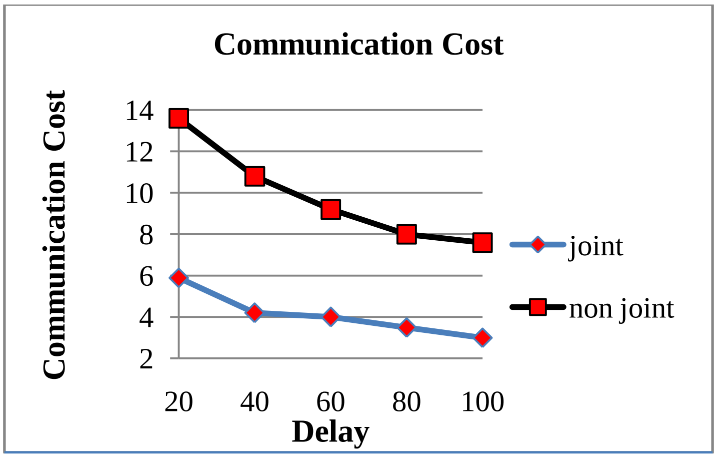

Fig.5.4 (b) Communication Cost.

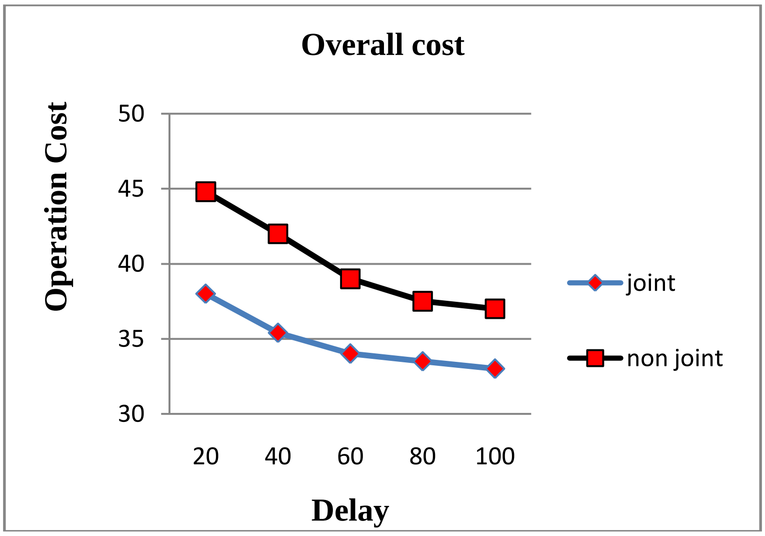

Fig.5.4(c) Overall Cost.

On the effort of number of replica.
The number of replicas for each data chunk set from 1 to 6. Increasing replica number from 1 to 4, there is limited number of activated servers are always enough for task processing, as shown in Fig. 5.5(a).

Possibly when task and its required data chunk are placed on the same server reduce the communication cost, as shown in Fig. 5.5(b).Fig. 5.5(c) is that the total cost first decreases and then increases with the increasing number of replicas.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

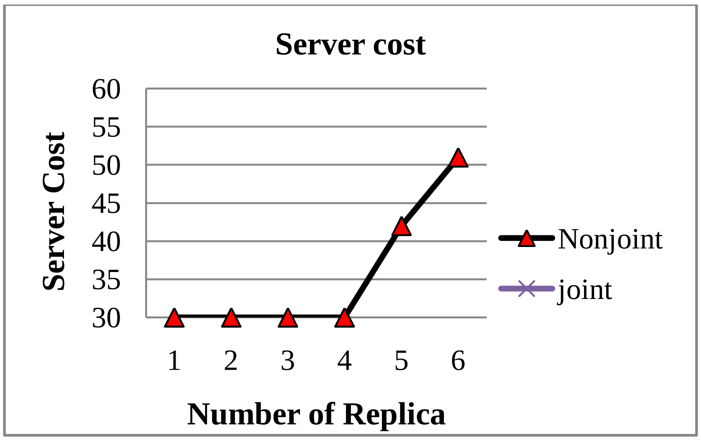

Fig.5.5 (a) Server Cost

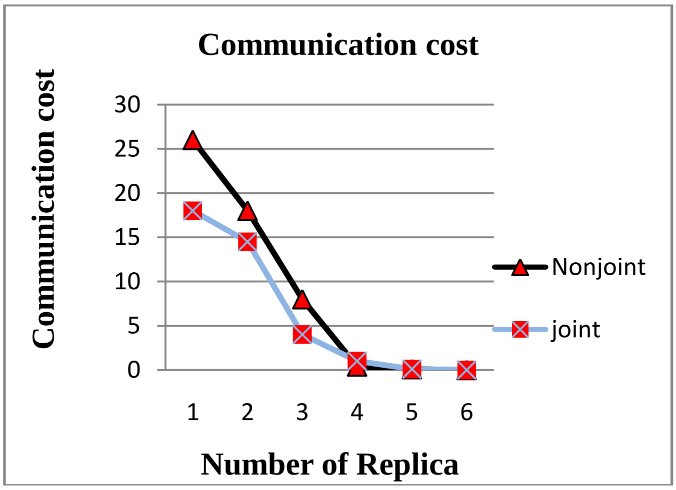

Fig.5.5 (b) Communication Cost.

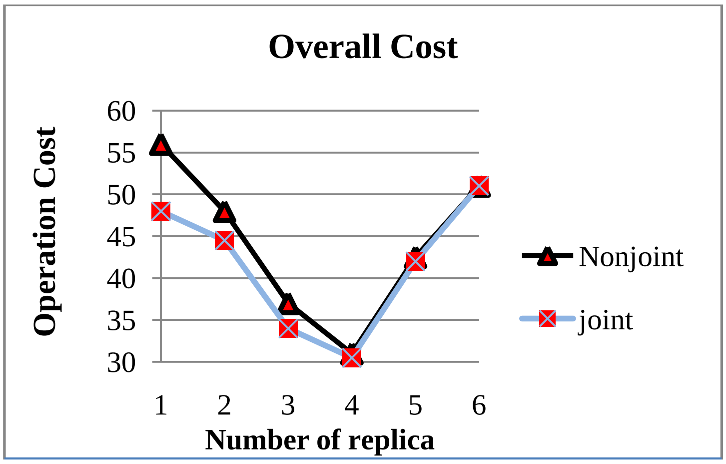

Fig.5.5 (c) Overall Cost.

VI. CONCLUSIONS

In this paper we study the issue related to geo distributed data centers. Also get introduced with the three factors of data placement, task assignment, data center resizing and routing to minimize the complete operational cost in large-scale geo-dispersed data centers for big data applications. First of all illustrated a 2-dimensional Markov chain used for the data processing process. Then found the predictable completion time, on which expressed the joint optimization as problem with the MINLP. To deal with high complexity of computation of solving our MINLP, it is linearized into an MILP problem. Concluded that with wide experiments, we show that our joint-optimization solution has considerable benefit over the tactic by isolated optimization.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

REFERENCES

[1] Data Center Locations," 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 Proceedings of the 13th InternationalConference on Architectural Support for Programming Languages andOperating 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 MultiElectricity-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 ComputerSystems (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,
[8] pp. 107–113, 2008.
[9] 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.
[10] Marshall and C. Roadknight, "Linking cache performance to user behaviour," Computer Networks and ISDN Systems, vol. 30, no. 223, pp. 2123 – 2130, 1998.
[11] H. Jin, T. Cheocherngngarn, D. Levy, A. Smith, D. Pan, J. Liu and N. Pissinou, "Joint Host-Network Optimization for EnergyEfficient Data Center Networking," in Proceedings of the 27th International Symposium on Parallel Distributed Processing (IPDPS), 2013, pp. 623–634.
[12] Qureshi, R. Weber, H. Balakrishnan, J. Guttag, and B. Maggs, "Cutting the Electric Bill for Internet-scale Systems," in Proceedings Of the ACM Special Interest Group on Data Communication (SIGCOMM). ACM, 2009, pp. 123–134.
[13] X. Fan, W.-D. Weber, and L. A. Barroso, "Power Provisioning forA Warehouse-sized Computer," in Proceedings of the 34th Annual International Symposium on Computer Architecture (ISCA). ACM,2007, pp. 13–23
[14] S. Govindan, A. Sivasubramaniam, and B. Urgaonkar, "Benefits and Limitations of Tapping Into Stored Energy for Datacenters," in Proceedings of the 38th Annual International Symposium on Computer Architecture (ISCA). ACM, 2011, pp. 341–352.
[15] 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 ofInternational Conference on Measurement and Modeling of Compute Systems (SIGMETRICS). ACM, 2012, pp. 175–186.
[16] R. Kaushik and K. Nahrstedt, "T*: A data-centric cooling energy costs reduction approach for Big Data analytics cloud," in 2012 InternationalConference for High Performance Computing, Networking,Storage and Analysis (SC), 2012, pp. 1–11.
[17] M. Sathiamoorthy, M. Asteris, D. Papailiopoulos, A. G. Dimakis R. Vadali, S. Chen, and D. Borthakur, "Xoring elephants: novelerasure codes for big data," in Proceedings of the 39th internationaconference on Very Large Data Bases, ser. PVLDB'13. VLDB Endowment, 2013, pp. 325–336.
[18] B. Hu, N. Carvalho, L. Laera, and T. Matsutsuka, "Towards big linked data: a large-scale, distributed semantic data storage," in Proceedings of the 14th International Conference on Information
[19] Integration and Web-based Applications & Services, ser. IIWAS '12.ACM, 2012, pp. 167–176.
[20] H. Shachnai, G. Tamir, and T. Tamir, "Minimal cost reconfiguration of data placement in a storage area network," Theoretical Computer Science, vol. 460, pp. 42–53, 2012.
[21] 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 ystems Design and Implementation (NSDI), 2010, pp. 17–32.
[22] S. Gunduz and M. Ozsu, "A poisson model for user accesses toweb pages," in Computer and Information Sciences - ISCIS 2003, ser.Lecture Notes in Computer Science. Springer Berlin Heidelberg,2003, vol. 2869, pp. 332–339.