International Conference on Emerging Trends in Engineering, Technology, Science and Management
IIMT College of Engineering, Greater Noida, India (ICETETSM-17)
12th April 2017, www.conferenceworld.in ISBN: 978-93-86171-38-2

IMPROVING DATA REPLICATION IN ONLINE SOCIAL NETWORK

Supriya F. Rathod[1], Prof. A. V. Deorankar[2]

[1] *PG Scholar, Department of Computer Science and Engineering, Government College of Engineering, Amravati (India)*

[2] *Associate Professor, Department of Information Technology, Government College of Engineering, Amravati (India)*

ABSTRACT

Nowadays, Online social networks are most important and useful network for people that are situated at different places to find and interact with other people based on their shared interests, ideas, in association with different groups created, etc. Common online networks such as Facebook, Twitter, MySpace, and LinkedIn all offer users a variety of ways to increase their networks, share notes and various types of media, and connect them. Social networks Facebook and Twitter have billions of users that are spread all about the world sharing their all intecronnected data. Users wants low service latency access to their own data with respect to their friends' data, which is very large, e.g. wall post updates, videos, music etc. Though, social network service providers have some of a paid datacenters to store each and every data of user everywhere to minimise users' data low service latency. Geographically distributed cloud services with high capabilities are suitable for storing large scale data in social networks at geographically different locations. In this paper we have focussed on problems such as storing and replicating data to geographically distributed nearest datacenters. For this we have used a genetic algorithm based approach to find every user's number of replicas for his data and to place replicas to reduce economic cost with fulfilling low latency needs for all users in the network. Distributed Datacenters represents a software structure that allows user to store and access multiple data in local or remote databases.

Keywords: Distributed datacenters, data replication,genetic algorithm, online social network, latency.

I. INTRODUCTION

Nowadays online social network is composed of thousands of servers which are spread all over the world across different datacenters that are placed at different geographical locations. This infrastructure is used to store all user data in OSNs such as user's profile infomration, contacts, udates, etc. This helps various OSN providers to control all valuable infomration of billions of users. On the other hand, this type of centralized infrastructure has a number of problems such as:

Poor scalability: As the number of users and the activity of user increases, the centralized infrastructure has to be larger. This provides poor scalability to OSNs. So the communication storage depends on the datacenters in the network.

High cost: For maintaining a large number of users and their huge dataset, the entire OSN provider number of servers and datacenters to store all this data. So it is well known that all these datacenters facilities are more eApensive. As it require to maintain cooling, electricity and traffic among other factors of datacenters.

Single Point of failure: The data stored can be easily attacked which can lead to misuse of data in OSN.

Database replication is the frequent electronic copying data from a database in one computer or server to a database in another so that all users share the same level of infomration. The result is a distributed database in which users can access data relevant to their tasks without interfering with the work of others. The implementation of database replication is for the purpose of eliminating data ambiguity or inconsistency among users. So the replication process shows the copying of data from the server to the user systems. Now, many social network providers have their own private datacenters to store various users' data in the network. However, this private datacenters are very eApensive therefore it is usually not an option for all social network providers. As a result Providers are always worried about the increasing storage capacity, data transmission updates and financial costs. To reduce all this cost related issues of datacenter setup and its maintenance, a improved key is to utilize of cloud datacenters. Though, social network service providers have some of a paid datacenters to store each and every data of user everywhere to minimise users' data low service latency. Geographically distributed cloud services with high capabilities are suitable for storing large scale data in social networks at geographically different locations. In this paper we have discussed some of key problems such as how to optimally store and replicate these huge datasets and how to distribute the requests to different datacenters. For this we have used a genetic algorithm-based approach to find every user's number of replicas for his data and to place replicas to reduce economic cost with fulfilling low latency needs for all users in the network. Distributed Datacenters represents a software structure that allows user to store and access multiple data in local or remote databases. Data replication in computing involves sharing infomration so as to certify reliability between redundant resoucres, such as software or hardware components, to improve reliability, fault tolerance or accessibility. Database replication can be used on many database management systems, usually with a master/slave relationship between the original and the copies. The master logs the updates, which then move through to the slaves. The slave outputs a message stating that it has received the update successfully, thus allowing the sending (and potentially re-sending until successfully applied) of subsequent updates.

II. MOTIVATION AND BACKGROUND

Online social networks are use to deal with a very large number of users which are distributed all over the world and share their increasing data that is intecronnected in the network. This data is in large amount. The online social network users usually have friends at distinct places who eApect to have their data in average time. For eAample, as of the fourth quarter of 2016, Facebook had 1.86 billion monthly active users. And in the third quarter of 2012, the number of active Facebook users had surpassed 1 billion. This shows that the number of users in Facebook is goes on increasing with respect to time. As on a reseacrh in 2012 [3], Facebook becomes most famous which is having 550 million daily active users. Facebook invests more than $1 billion in the

infrastructure which supports its online social networking, in the study of 2012 it is shown that around 845 million users a month are available in facebook. The facebook organization near about spents $606 million for one servers, storage, network equipment and datacenters in 2011 where it increases in 2012 by another $500 million [3]. Facebook has near about 10 private datacenters in 2015 due to its growing number of users and their datasets size. But nowadays facebook requires more datacenters as the users are goes on increasing day by day. As the data storage in social networks is a data intensive job and every one provider in the network cannot afford many datacenters setup in different locations at all over the world. Therefore nowadays many social network providers such as DropboA [5] prefer to use cloud datacenters as it's a more convenient way to store user's data.

The key problem discussed in this paper is user's data placement and its replication in online social network while providing various services by considering provider's financial charge. This also takes into consideration that the service latency will be low in the network. In the communication process of user's data transfer cost is not involved. As the data needs to be transfer while interacting is regardless of where the other user is located, this confimrs that data transfer cost is not required and it is reflected in latency. In this paper, we have not considered the data update cost this is because our system is static. In the network every user must have a primary copy of data which is located in their primary datacenter, and this will be according to distance of datacenter from user location. It is supposed that all the users in network must read their own data from their own primary datacenter and every friend of that user must read their data from their nearby datacenter which stores their secondary data. It is also assumed that every write operation goes to the primary datacenter. Now suppose, M is the number of datacenters and N is the number of users, in one dataset. The collection of datasets and users stored in various datacenters are denoted by respectively:

$U = \{u_1, u_2, ..., u_N\}$

$D = \{d_1, d_2, ..., d_N\}$

Whereas datacenters in the system are denoted as:

$S = \{s_1, s_2, ..., s_M\}$

The solution space is a matriA A of size $N \times M$ as follows:

$$X_{ij} = \begin{cases} 1, \text{Data of user is stored in datacenter} \\ 0, \text{Otherwise} \end{cases}$$

By considering all this notations we can calculate the cost estimation and problem fomrulation used while implementing genetic algorithm in this work[1].

III. PROPOSED APPROACH

The global distribution of the OSN users of different countries with the OSN presence, the number of users ranges from 260 to over 150 million. The typical latency budget for the data recovery portion of a web request is only 50-100 milliseconds. Since the OSN's has become very populationular globally, the new OSN model with globally distributed datacenters and locality-aware mapping (i.e., mapping users to their geographically close datacenters for data storage and services) would reduce service latency.

Data replication: To achieve our goal, a datacenter can replicate its regularly requested user data from various datacenters, which on the other hand necessitates inter-datacenter data updates. Thus, to reduce service latency

in inter-datacenter, a datacenter equally considers visit rate and update rate to create replicas that save more visit loads than concurrently generated update loads. In this section, we verify the benefits of the data replication in new OSN model and also analyze data from a major OSN to verify general OSN properties. In order to obtain a representative user sample, we used an unbiased sampling method to crawl user data. If a randomly generated id eAists in the OSN and the user with the id is publicly available, we crawled the user's data. We anonym zed users' IDs and only recorded the time stamps of events without crawling event contents. All datasets are protected and are not shared freely. In this paper we proposed a low compleAity data replication aware model.

Our contributions are summarized as below:

We developed a data replication aware model that facilitates the integration of eAisting data replica in the scheduling decision on geographically distributed environments.

We also proposed a number of real workload traces and various data placement models to investigate the perfomrance factors in the job scheduling problem on geographically distributed datacenters.

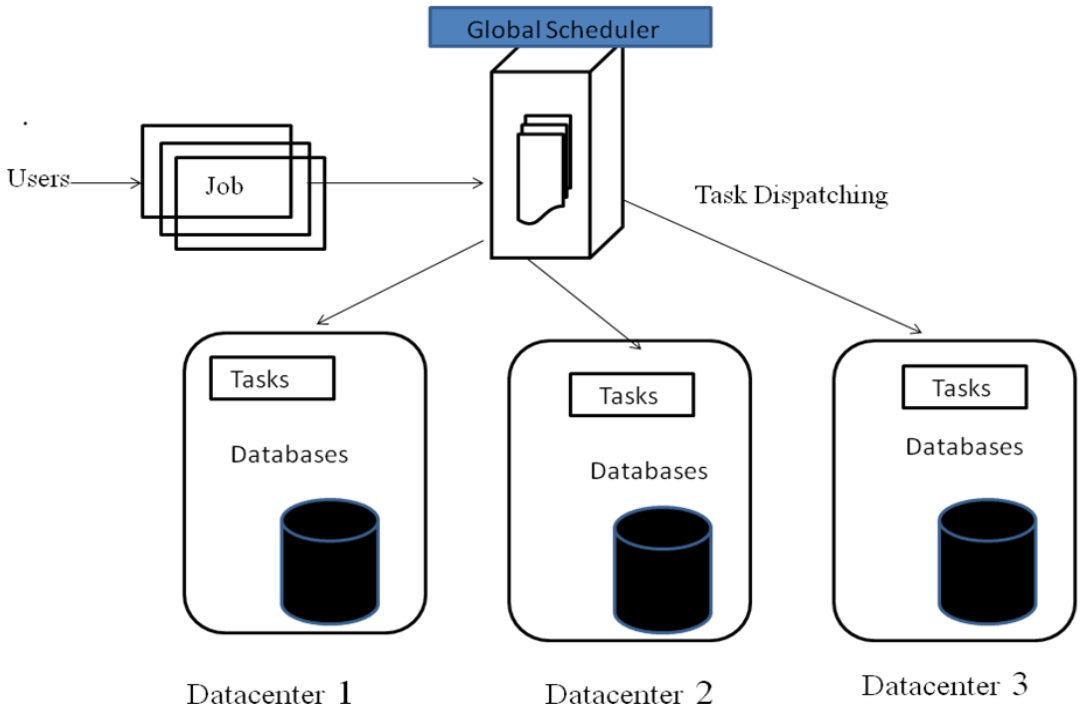

Fig. 1 System Acrhitecture

As shown in the above fig.1, users firstly login their accounts and then on the basis of their interaction with friends relicas are generated and stored in various datacenters. These replicas are stored on geographically nearest datacenters. Geographically distributed data stores: The Coda file system uses data replication to improve perfomrance and availability in the face of slow or unreliable networks. Unlike Coda, TAO does not allow writes in portions of the system that are disconnected. Megastore is a storage system that uses PaAos across geographically distributed datacenters to provide consistency guarantees and high availability [2].

3.1 Genetic Algorithm

Genetic algorithm for social network data placement and replication

Input:

Rate of crossover: *cr*

Rate of mutation: *mr*

Size of population: s*p*

Size of selected population: *ssp*

Number of iterations: *iter*

Output:

Solution:*A*

// Initialisation

1 First step is to generate *sp* feasible solutions randomly;

2 Then save them in the population *population*;

// Loop until the temrinal condition

3 Now, for *i*=1 to *iter* do

// Crossover

4 and, for *j*=1 to *sp-1* do

5 randomly select two solutions A_a and A_b from population;

6 generate A_c and A_d by two-point crossover from A_a and A_b under rate *cr*

7 if latency requirement is valid, save A_c and A_d to *pool*;

8 update *newpopulation* = *population* + *pool*;

9 end of for loop

// Mutation

10 *Length* = size of *newpopulation*

11 for *j*=1 to *Length* do

12 select a solution A_j from *newpopulation*

13 mutate each bit of A_j under rate *mr* and generate a new solution *A'j*

14 if latency requirement is valid, update A_j with *A'j* in *newpopulation*;

15 end of for loo*p*

16 end of for loo*p*

// Selection

17 now select *ssp* solutions from *newpopulation* and save them in *population*;

// Returning the best solution

18 return the best solution *A* in *population*;

3.1.1 *Initialisation*

Firstly, we do the encoding of various users' data replicas placement at multiple datacenters is completed. In this, we have used a two-dimensional encoding users' ID and the datacenters ID. The users' ID is an indicator of users' datain communication and the ID is for various datacenters is used.

3.1.2 *Crossover procedure*

The GA is worked in the manner that, firstly a random crossover point is particular then we swapped the segments of parents at the particular point to generate new children of their parent. In this procedure children

can take over the features of his parents. For generating two new chromosomes, a two point crossover method is used with some specific probability.

3.1.3 *Mutation procedure*

In process of mutation, all stored replica are mutated with a particularly selected cell of a chromosome of a child. While doing this mutation rate must be set to a tiny probability as mutation process can easily demolish the correct topological order and this helps to result in invalid solutions.

IV. LITERATURE REVIEW

GuoAin Liu, Haiying Shen uses the algorithm for selective user data replication. Also provides concept of selective data replication mechanism in distributed datacenters. These works focuses on load balancing and document security and availability within the network [1]. The online social networks inter-datacenter communication services are also discussed. This also focuses on a geographically based replica creation in large global web sites such as Facebook. Volley [5] shows that datacenter capacity limitations and it also deals with data placement chalLengthge in WAN bandwidth with providing low network latency. While in previous study of M. S. Ardekani and D. B. Terry, they focus on the storage of data in cloud while adapting changes in users' locations or request [5]. The primary focus in [6] is to minimise service latency application providers cost reduction and to satisfy consistency and fault-tolerance requirements with considering workload properties. The basic idea to design data replication datacenters is based on many previous studies on online social network properties. The work in [3] studied online social network evolution patterns and user behaviour.

V. CONCLUSION

This paper uses a genetic algorithm for social network user's data placement and replication in cloud datacenters. After comparing different data placement strategies, our proposed algorithm shows that placement strategy with genetic algorithm is a most efficient which fulfils low latency requirements of users in online social network. In this paper we focussed on some of key problems such as to store and replicate users huge datasetsin the network and also to distribute that replicas to different datacenters which are located geographically nearest to the users location. For this we have proposed acrhitecture of data replica creation and its transfomration in the social networks.

REFERENCES

[1] GuoAin Liu, Haiying Shen, Senior Member IEEE, Harrison Chandler, "Selective Data Replication for Online Social Networks with Distributed Datacenters," IEEE Transactions on Parallel and Distributed Systems,(Volume 27, Issue 8, Aug 1 2016).

[2] Hourieh Khalajzadeh, Dong Yuan, John Grundy, Yun Yang,"Improving Cloud-based Online Social Network Data Placement and Replication," 2016 IEEE 9th International Conference on Cloud Computing

[3] Uras Tos, Riad Mokadem, Abdelkader Hameurlain, Tolga Ayav and Sebnem Bora, "A Perfomrance and Profit Oriented Data Replication Strategy for Cloud Systems," 2016 Intl IEEE Conferences on Ubiquitous

Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress.

[4] R. Miller. (2012), "Facebook's $1 Billion Data Center Network," Available: http://www.datacenterknowledge.com/acrhives/2012/02/02/facebooks-1-billion-data-center-network/

[5] S. Agarwal, J. Dunagan, N. Jain, S. Saroiu, A. Wolman, and H. Bhogan, "Volley: automated data placement for geo-distributed cloud services," 7th USENIA Symposium on Networked Systems Design and Implementation (NSDI), 2010.

[6] M. S. Ardekani and D. B. Terry, "A self-configurable geo-replicated cloud storage system," 11th USENIA conference on Operating Systems Design and Implementation, 2014.

[7] B. Viswanath, A. Mislove, M. Cha, K. P. Gummadi, "On the evolution of user interaction in facebook," in Proc. of WOSN, 2009.

[8] Z. Wu, M. Butkiewicz, D. Perkins, E. Katz-Bassett, and H. V. Madhyastha, "SPANStore: cost-effective geo-replicated storage spanning multiple cloud services," 24th ACM Symposium on Operating Systems Principles, 2013, pp. 292-308.

[9] L. Jiao, J. Lit, W. Du, and A. Fu, "Multi-objective data placement for multi-cloud socially aware services," IEEE Conference on Computer Communications (INFOCOM), 2014, pp. 28-36.

[10] N. Bronson, Z. Amsden, G. Cabrera, and et al., "TAO: Facebooks Distributed Data Store for the Social Graph," in Proc. of ATC, 2013.