International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

Automatic Data Replication Using Distributed Datacenters 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, Head of Department, Department of Information Technology, Government College of Engineering, Amravati, India

Abstract: ***Online Social Network (OSN) is a network which consists of real users that interact with each other in various ways. Nowadays the number of users in an OSN increases diversely with respect to time. In this paper we are introducing the new OSN model, which distributes datacenters worldwide, to help decrease service latency which leads higher inter-datacenter communication load. For example in Facebook, recently it becomes very famous and most visited online network site. Each datacenter of Facebook has all the data, which are updated by the master datacenter, leading to remarkable load in the new model. This model uses datacenters to store data at their geographically nearest datacenters. The regular interactions in online network between the various users can generate long service latencies. In this paper, we proposed location based load balancing concept using which user will be redirected to the geographically nearest server and achieving low service latency. For this we proposed the concept of Automatic Data replication concept in Distributed Datacenters (AD3). In this model, AD3 jointly considers update rates and visit rates to create replicas; these replicas are created using selected data for replication.AD3 also includes users various types of data such as friends request, status update, friend post, music or videos update; etc for replication, while considering all these updates AD3 can lower the service latency. AD3 also includes replica updates, replica deactivation, anddocument security, availability.***

Keywords: Online Social Networks, Datacenter, Data replication

1. Introduction

In the recent years, Online Social Networks (OSNs) are becomes very powerful and useful which are spread all over the world. According to recent studies of 2016, Facebook is one of the popular and major worldwide OSNs which have near about 1.79billion monthly active real users and this study also shows that about 80% of users are outside of the US. However, datacenters of Facebook are deployed densely in the US. In this paper, we proposed a location based load balancing concept using which user will be redirected to the geographically nearest server and achieving low service latency.This model also includes creation of replica and its updation in datacenters for Online Social Network with distributed datacenters and it also leads to minimize inter-datacenter interaction load.The user data set is made up of various types of data which includes various friend pairs, wall posts, music, comments, personal info, photos, videos, etc. The data stored in datacenter is the data in the form of documents while the photos and videos are stored in content delivery network (CDN) partners of Facebook. We then proposed an Automatic Data replication concept in Distributed Datacenters (AD3) model for Online Social Networks that completes all the mentioned features.

Replication is the process of creating copy of documents; this document should be in the form of XML document on a local data server this decreases the load of database which automatically increases performance of model. This happens because as the time required fetching XML document data is less than that of the document in the database. As our model includes features such as security as well as availability of documents in case of any attack, the XML document made it easier. The replica created will be stored at various datacenters Data replica. These datacenters are located worldwide but replicas should be placed to their geographically nearest server. Weare proposing a location based load balancing concept using which user will be automatically redirected to their geographically nearest server and in case of any damage, system will automatically recover its damaged file from other datacenters as replica documents are stored in distributed datacenters.

1.1 Motivation

Motivation behind this project is to provide more security in the online social networks and also to make availability of documents to the users by creating replica. This also leads to replicate data in Online Social Network with distributed datacenters so as to minimize inter-datacenter communication load in the network. In this paper, we proposed a location based load balancing concept using which user will be redirected to the geographically nearest server and achieving low service latency.This model also includes creation of replica and its updation in datacenters for Online Social Network with distributed datacenters and it also leads to minimize inter-datacenter interaction load.In AD3, we focus on users friendship, every datacenter must refers to the actual user interactions with consideration of the update load and saved visit load while creating replica in datacenters in order to achieve low service latency in inter datacenter communications. As each user have different data this increases update rates. Each datacenter only replicates that data which saves inter-datacenter communications, rather than replicating a user's all data. AD3 also has a replica deactivation scheme which improves storage capacity and performance of the model. This helps to overcome network load in the datacenters.

2. Literature Survey

Previous study of A. Thomson, D. J. Abadi [2015] uses the concept of reliability protection of replicas over geographically placed datacenters or within a datacenter in the network. This work mainly focuses on providing replicas on the geographically nearest server.

After this study, Guoxin Liu, Haiying Shen [2016] uses the algorithm for Automatic user data replication in datacenters. This work focus on data replica to be stored into database on local server which may increase the local database load as replica increases.

Y. Zhou, T. Z. J. Fu, D. M. Chiu [2013] and M. S. Ardekani, D. B. Terry [2014], also shares the adaptive replication techniques with some works in P2P systems and in clouds, which energetically used to the number and location of data replicas. These works focus on load balancing, while AD3 focuses on saving network load, availability of document, security.
Y. Wu, C. Wu, B. Li, L. Zhang, Z. Li, and F. C. Lau [2012] utilize the geo distributed cloud concept to support large scale social networking activities. AD3 must always focus on OSNs' all datacenters to replicate user documents and distributed worldwide datacenters this data should be in the form of XML document, each datacenter should contain replicas but not all replicas.

Viswanath et al. [2010] showed that social links can grow on large scale or can reduce its importance over time, which helps to achieveAD3's strategy of regularly improving the need of replicas. AD3 focuses on different visit/update rates of user interaction to supports the atomized user data replication.

3. Automatic Data Replication Using Distributed Datacenters

In this section, we authenticate the profit of the new OSN model for distributed datacenters. The main aim to developAD3 for OSNs is reduce the network load with low service latency. If at any instance an id is created randomly in the OSN is occurred and the user of that id is widely accessible, then we crawled that user's data in OSN. We observe all the users' IDs and the time stamps of events on their profile not including crawling event contents. All the datasets in online social network are completely protected and made private. In this section we design a model for AD3, to achieve our goal of load balancing concept in the distributed datacenters by using replica formation concept. This leads to increase document security and availability.

3.1Automatic User Data Replication

In the existing system only inter-network communication load balancing is discussed. There is no provision is given to increase the availability and security of documents. Data Replica will be stored into database on local server which may increase the local database load as replica increases. Replica data will be stored on local server as it is, which may cause attack on replica data. Replica update will increases network congestion. To overcome all this disadvantages we proposed model of AD3 in the online social network for distributed datacenters. Replication is the process of creating copy of documents; this document should be in the form of XML document on a local data server this decreases the load of database which automatically increases performance of model.

Here we produce data replica created through different user interactions is used in the online social networks. In study it has been seen that the interactions between various users decreases with respect to their age and time and also each user in the network has different update rates and visit rates according to their interest. This shows that each user relationships do not essentially have high data visit or update rates between the friends in social networking sites. These rates depend upon the friends and overtime of their updates in the sites. According to study about90% of all friend pairs in networks have a regular interaction rate of friends visit below 0.4, with the regular interaction rate of the remaining10% is ranges from 0.4 to 1.8[1], so the data update rate is completely depends on users activities. From above information it is clear that not all the users have same data visit rate and updates rates in the networks. Therefore users with low visit rates in communities are leads to generate replicas with low intensity. But the replicas are created for these types of users also this leads to waste of storage space in the datacenters. Thus, we are considering only those visit rates of a user's data that are communicating regularly in the network data replication. As various users have different updates and visit rates, thus we have to differentiate various users update rates for this we use equation to calculate variance by using $\sigma^2 = \sum(x - \mu)^2 / (n - 1)$, where x is the interaction rate of users friendship and μ is the average and n is the interaction rates number[1]. From this equation it is shown that about 10% of friend interactions in networking sites are having high variation which ranges from[0.444 to 29.66], these leads that update rates of various users can vary according to time in the network. From all this study it is clear that visit or update rates of user's data replica should be checked always with respect to time and updates.

Some points to be considered are:

3.1.1 Documents replication:

In order to increase the security and availability of user's documents, we proposed a document replication scheme concept. In this scheme, every document will be stored on a particular application server and the replica of the document will be stored on the other servers. This provides a document access to users anywhere.

3.1.2 Document Encryption:

In AD3, the replicated documents will be stored on server in encrypted format using AES algorithm. For this system will generate unique encryption key for every document.

3.1.3 Document Decryption:

After encrypting a document user needs to decrypt it while accessing it again. At the time of document decryption, users have to specify secrete key which is generated while encryption. This same key is used for both encryption and decryption. If secrete key is verified, document will be decrypted. In case of any decryption error, system will fetch replica of the document and deliver it to user.

3.1.4 Location Based Server Redirection:

In AD3, to categorize the users we propose location based server redirection. Using this technique, user will be redirected on geographically nearest server to increase efficiency of system. Frequently accessible Friend's data will be stored as a replica on user's server in XML document in encrypted format using AES. The replica document will be updated in case of any changes occurred.

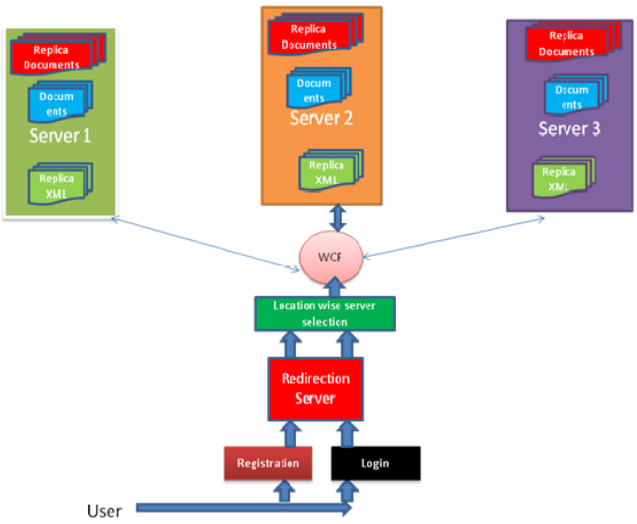

Figure 1: Proposed Architecture of AD3

The main aim of proposed system architecture is to reduce network load with respect to datacenter load and provide security to documents which increases performance as time required to fetch data from XML is less than that of database. Figure 1 shows a complete architecture of AD3 model. Mainly data replica is provided to increase security as well as availability of documents in case of any attack. Data replica will be stored on the various datacenters according to their geographically nearest server which is shown in the figure. This helps user to get his data within fraction of seconds.In case of any damage, system will automatically recover the file from stored replica documents in distributed datacenters. All the connected datacenters can have all copies of replica in the network. Firstly, user has to register in OSN network, and then according to its location they are redirected to a particular datacenter. After registration user can communicate with their friends in the network. This reduces workload in the network. Therefore AD3 architecture is proposed to overcome this entire load in the network. Data will be provided in the form of encrypted and decrypted data to increase the security of data. Replica of data is stored in the form of XML document to improve load balancing.

3.2 Algorithm Design

The communication happens when different users reads or writes their friend's data into other datacenters. This generates replicas which can be reduced by creating local replicas. This also leads to data update load on datacenters. In this work we are decreasing the network load with respect to service latency by using the concept of automatic data replication. Now for this work we are introducing an algorithm of automatic data replication which will cause low network load in the datacenters. For this purpose we are using some measure from [12] for network load. It also shows estimated cost in data communication.

Table 1: Notations Table

S/s	Total number of datacenter/datacenter s
$V_{out}(s)$	Number of users at datacenter s
$R_{out}(s)$	Number of replicated users at $V_{out}(s)$
i/d_i	User i/d is a data of user i
s(i)	User i's master datacenter
Ur_i	update rate
$Vr_{c,i}$	visit rate
$A^v_{j,i}$	jth message size with respect to visit at i
A^v_i/Au_j	average visit/update message size
$A^u_{s,i}/Cu_{s,i}$	Saved visit/consumed load
$G_{s,i}$	Gain in the network load
$Di_{s,s(i)}$	distance between datacenters s and s(i)
$Ls_{s,i}$	saved visit network load
$Lu_{s,i}$	update network load

Now consider the jth visit from any one of the user in network at datacenters towards user iin datacenter s(i) will be measured by $A^v_{j,i} \times Di_{s,s(i)}$MBkm(Mega-Byte-kilometers).

After some time, we calculate the whole network load of users datacenter communications which will be the saved network load of users replicas (denoted by Ls). This time period is denoted by T. Now equation of Ls becomes:

$$Ls=\sum_{c\in C}\sum_{i\in R_{out}(s)} Vr_{s,i}A^v_i \times Di_{s,s(i)} \times T \ldots\ldots\ldots(1)$$

Nowadays in OSNs, users are will be interested in friends' recent updates such as posts in the News Feed. Now, inAD3 we are focusing on user i's recent data updates generate replica, which may have high $Vr_{s,i}$. If in any case j ∈Uout(c), then Ls will reach the maximum value. This also leads generate extra update load which is then denoted by Lu. Related to Ls in Eq. (1), Lu is then calculated by the summation of network load. Thus,

$$Lu=\sum_{c\in C}\sum_{i\in R_{out}(s)} Ur_i \times A^u_i \times Di_{s,s(i)}) \times T \ldots\ldots\ldots..(2)$$

Now our aim is to overcome network load by increasing the gain (denoted by G) of replication data:

$$G_{tot}= Ls- Lu \ldots\ldots(3)$$

Now each datacenter try to achieve low network load for it and for achieving this aim it choose replicas with low threshold value which is denoted as δ_{mx}. Now each datacenter s must keep track of all its visited user replicas from master datacenter I and calculates its gain on datacenter:

$$G_{s,i}= Ls_{s,i}-Lu_{s,i} \ldots\ldots\ldots\ldots..(4)$$

If $G_{s,i}>\delta_{mx}$, datacenters replicates user i. As we know that various friends have different interaction rate. Then we remove $G_{s,I}$ with low rate after calculating all the $G_{s,I}$ of replica. Removing of this low rate replica does not mean it is deleted but it is stopped for receiving updates. Replica lost is occurring in only one condition if datacenter don't have enough storage space. AD3 introduces a threshold δ_{mn} value to avoid frequent creation of replica which will be less than

δ_{mx} value. If $G_{s,\ i}<\delta_{mn}$, then datacenters removes user i's replica. This result in,

$R_{out}(s) \leftarrow \{i|i\in U_{out}(s)\wedge((G_{s,\ i}>\delta_{mx}\wedge \neg\ j \in R_{out}(s))\vee(G_{s,\ i}>\delta_{mn}\wedge i\in R_{out}(s)))\}$……(5)

As in Eq. (5), it is shown that AD3 becomes the system of replicating all previously data [12] with a long supply time, and this will be happen due to if we set negative δ_{mx} and δ_{mn}. AD3 uses the gain value and δ_{mx} and δ_{mn} thresholds to achieve an ideal balance as shown in Eq. (5) which shows the weights for their objective. We use service latency constraints, saved network load, user data replication overhead, replica management overhead factors for determining all this equations. Now we have to decide a T which is a period for users visit and update rates, it needs to be selected very carefully. The automatic user data replication algorithm has O(N) time complexity.

Algorithm for Atomized user data replication

Input: Set of visited users during previous period, V(s) Current slave replicas set, $R_{out}(s)$
Output: $R_{out}(s)$
For each $i\in R_{out}(s)$**do**
If $i\in V_s$ **then**
$G_{s,\ i}\leftarrow (\sum_j A^v_{j,i} \times Di_{s,s(i)} - \sum_j A^u_{j,i} \times Di_{s,s(i)}) \times T$
else
$G_{s,\ i}\leftarrow 0$
end if
if $G_{s,\ i}<\delta_{mn,\ c(i)}$**then**
remove local replica of i
delete i from $R_{out}(s)$
notifys(i)
end if
end for
for each $i\in V_s\wedge i\notin R_{out}(s)$**do**
$G_{s,\ j}\leftarrow (Vr(s,\ i) \times A^v_i \times Di_{s,s(i)} - Ur_i \times A^u_i \times Di_{s,\ s(i)}) \times T$
If $G_{s,\ i}\geq \delta_{mx,\ s(i)}$**then**
create a local replica of i
addiinto $R_{out}(s)$
notifys(i)
end if
end for

Algorithm 1 shows the overall procedure for automatic data replication of user. Whenever any datacenter updates its replica of user i, it will be always notifies toi's master datacenter. Every master datacenter should need to be maintaining all the records from its slave datacenters for data updates of user.

3.3 Replica Deactivation

In OSNs, the time period of two successive visits of the identical user replica may be long or short, so between this times period there may happens many updates in profiles. Now the replicas are not formed immediately, the replica will be created after next visit of the corresponding user according to time period calculated by using equation of variance. By using this strategy we can overcome the network load in network. For simplifying this we propose a concept of replica deactivation. We use this scheme to maintain the storage space of datacenters. By deactivating replica, network load and storage space is minimized. The replica deactivation will be depends on the users interaction and its update rates. For deactivating replica we set some time period limit T, which is use for automatic deactivation of replica.

Recall, if at any time users ask for its deactivated replica, then replica datacenter should ask for its missed updates before replying to user request, this will produce some service delay. The replica deactivation scheme can reduce load for network and also improves the storage capacity of datacenters.

4. Conclusion

Until now OSN model with distributed datacenters results in improved service latencies for users but challenge is that this model are not able to reducing inter-datacenter network load. Thus, we propose the Automatic Data replication concept in Distributed Datacenters (AD3) to lower interdatacenter network load while achieving low service latency in OSN. This also provides documents in order to achieve security as well as availability of documents in case of any attack and manages location wise datacenter selection to enhance the security of documents stored in social network using AES algorithm. Some friends may not have regular interactions and some distant friends may have regular interactions. On basis of this AD3 provides replica activation and deactivation mechanism. In AD3, rather than relying on inactive friendship, each datacenter refers to the actual user interactions between friends and it jointly considers the update load and visit load in determining replication in order to achieve low interdatacenter communications. Thus by applying all this features, we summarize a concept of AD3. In all this provide document security, availability by reducing network load.

5. Future Scope

In our future work, we will examine how to decide parameters in design to get different achievements on service latency and network load.

References

[1] Guoxin Liu, Haiying Shen, Senior Member IEEE, Harrison Chandler,"Automatic Data Replication for Online Social Networks with Distributed Datacenters, " IEEE Transactions on Parallel and Distributed Systems, 2016.
[2] Facebook statistics. https://newsroom.fb.com/Key-Facts.
[3] A. Thomson, D. J. Abadi, "Calvin FS: Consistent WAN Replication and Scalable Metadata Management for Distributed File Systems," in Proc. of FAST, 2015.
[4] Y. Zhou, T. Z. J. Fu, D. M. Chiu, "On replication algorithm in P2P VoD," IEEE/ACM Transactions on Networking, Vol. 21, No. 1, February 2013.
[5] M. S. Ardekani, D. B. Terry, "A Self-Configurable Geo-Replicated Cloud Storage System," in IEEE/ACM Transactions on Networking, 2014.

[6] Y. Wu, C. Wu, B. Li, L. Zhang, Z. Li, and F. C. Lau, "Scaling Social Media Applications Into Geo-Distributed Clouds, " in Proc. of INFOCOM, 2012.
[7] B. Viswanath, A. Mislove, M. Cha, K. P. Gummadi, "On the evolution of user interaction in facebook," in Proc. of WOSN, 2009.
[8] **"**Facebook." Available: http://www.facebook.com/
[9] "Socialbakers." http://www.socialbakers.com/facebookstatistics/
[10] H. Shen and G. Liu. A geographically-aware poll-based distributed file consistency maintenance method for P2P systems. TPDS, 2012.
[11] F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida. Characterizing user behavior in online social networks. In Proc. of ACM IMC, 2009.
[12] M. Wittie, V. Pejovic, L. B. Deek, K. C. Almeroth, and B. Y. Zhao, "Exploiting locality of interest in online social networks, " in Proc. of ACM CoNEXT, 2010.
[13] Z. Li and H. Shen, "Social-p2p: An online social network based P2P file sharing system," in Proc. of ICNP, 2012.
[14] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks, " in Proc. of IMC, 2007.
[15] A. Nazir, S. Raza, and C. Chuah, "Unveiling facebook: A measurement study of social network based applications, " in Proc. of IMC, 2008