METHODS TO IMPROVE DATA REPLICATION PROCESS 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, Maharashtra, India.

[2] Associate Professor, Department of Information Technology, Government College of Engineering, Amravati, Maharashtra, India.

Abstract – Online social networks make it more suitable for people to search and communicate with other people based on their shared interests, ideas, association with different groups, etc. Common social networks such as Facebook and Twitter have millions of users spread all around the world sharing interconnected data. Users always demand for low latency access for themselves and also for their friends e.g. videos, pictures etc. In this paper, we proposed various data replication methods in online social network that enforces data availability to prevent data transfer and also exploits data replications to develop overall system performance. Among these methods we have proposed a method of remote replication which is efficient in the data replication. In recent years online social network model are having datacenters that are worldwide distributed which helps to decrease the service latency which causes higher inter-datacenter communication load.

Key Words: Online social networks, Distributed storage, Data replication, Load balancing, methods.

1. INTRODUCTION

Nowadays many data-intensive applications and high performance applications are emerging. A traditional storage facility is not able to provide the complete data storage. The main problem in such systems is of distributing and locating method of various data objects. This kind of method deals with a problem of distributing and locating data, which has given an identifier file, such as how to find data location between thousands of storage devices. This will be done according to availability, scalability, user experience, and resource utilization of the system. Facebook is a most common and most visited online social network. It has more than a billion monthly active users who are interconnected through network. Recently, Facebook have gained major popularity in online social network. Firstly users create their personal profile and then they can send friend request to various users which leads to communication between many users situated at different places in the network. In this paper, we propose various methods for data replication in online social network. But the main is to focus on remote data replication technique and to provide more efficient and secure data to the user. The replication of data is on the basis of user interaction and their interest. Some important facts are found in the study that; peoples spend their most of the time on online social networking sites, most probably people shares only those files that they are interested in and users also don't like to provide faulty files to their socially close friends. As if user provide faulty file to their friend then it will lead to degrade their reputation in their social communities.

Replication means taking a complete copy of user data to another system, usually at a separate geographical recovery location. This means you have a backup of your data you can use this stored file if the primary copy of file fails. This replicated copy can also be used for other things like off-site backup, load testing. The replication can happen in all sorts of different places and the replication is configured and managed by the controllers of the storage providers. It can work at the application level, where replication is configured and set up between your host and storage system and virtualises your storage. That brings benefits in that the make and model of storage array at either end can be different so you're not confined to using the same model at both ends. The last place it can happen is at the host server level, and that gives you greater flexibility in allowing replication between different types of storage at both ends. In that model the software or the operating system controls the replication between two different servers at two different sites and the replication is over IP links.

2. DATA REPLICATION TECHNIQUES

Learn how remote replication fits into a disaster recovery plan, as well as the various data replication methods available, such as host-based, array-based and network-based replication.

2.1 Remote replication

Remote replication is a well-established practice that can replace traditional backup and enhance disaster recovery planning. But, remote replication can be deployed in a variety of ways, so what data replication methods are available

Remote replication copies data to a secondary site as part of a disaster recovery plan; it traditionally involved backing up application data, but it is now possible to replicate entire virtual machines too. This can be useful to maintain server images with the latest configuration, including operating system and application security patches that are all set to be made live in case of a serious outage at the primary site.

2.2 Host-based replication

Host-based replication is specific to a particular server. This can be a good way to improve service-level agreements at key points in an organisation. This replication provided faster recovery options than array-based replication could offer. Host-based replication probably gives user the best chance of integration with their application and is probably going to turn out to be the lowest-cost option.

2.3 Array-based replication

Array-based replication tools, such as EMC's SRDF and NetApp's SnapMirror, do have their advantages. This replication process provides how to replicate a whole storage array at once, which can make the replication process easier to manage. The downside of this data replication method is that these tools are often vendor-specific, which reduces customer choice when buying equipment. Array-based replication gives good performance because it removes processing and networking load from the servers and is probably the simplest to set up.

2.4 Network-based replication

The network-based replication generally uses an application level that sits at the edge of the network. These tools, such as EMC's Recovery Point, have the advantage of being able to manage heterogeneous arrays and servers. The other advantage of this data replication method is that it makes it easier to arrange replication policies that take multiple arrays and servers into account.

3. PROPOSED APPROACH

Remote replication is the process of copying data to a server at a remote location for data protection or disaster recovery purposes. Remote replication may be either synchronous or asynchronous. Synchronous replication writes data to the primary and secondary server at the same time. With asynchronous replication, there is a delay before the data gets written to the secondary site. Because asynchronous replication is designed to work over longer distances and requires less bandwidth, it is often a better option for disaster recovery. However, asynchronous replication risks a loss of data during a system outage because data at the target device isn't synchronized with the source data. Replication occurs in one of three places: in the storage array, at the server or in the network. Most enterprise data storage vendors include replication software on their high-end and mid-range storage arrays. Host-based replication software runs on standard servers, making it the cheapest and easiest type of replication to manage for many, but it taxes server processing. Replication on the network requires an additional device, either an intelligent switch or an appliance. The advantage of synchronous replication is that it eliminates the risk of accidental data loss. The downside is that it requires low-latency communication because the secondary server must confirm that each packet has been received without error. As far as the secondary site is from the primary, the harder is the connection establishment. Asynchronous replication will theoretically work over any distance, but user risk is of running behind the primary server grows as the distance between sites increases. Asynchronous replication also risks data loss in the event that the primary site goes down in between replication sessions. There are some workarounds, include multi-hop replication, in which an intermediary SAN within a viable distance is written to using synchronous replication, followed by asynchronous replication to a secondary site much farther away.

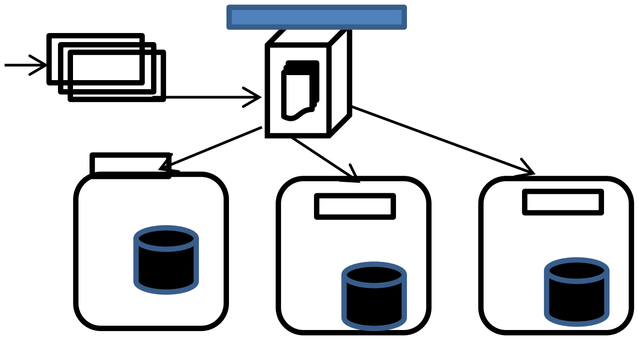

Fig. 1 System Architecture

Another alternative is to enhanced physical protection at the local server. Fig. 1 shows the system model of a geographically distributed datacenters which are interconnected through network. All the data from various servers are submitted in a centralized global scheduler and then this data is distributed among various datacenters for processing. The global scheduler maintains a first in first out technique for all submitted data. Various data concentrated applications produce and store data locally on nearest server. On the other hand, these data may be replicated on other datacenters in the network. For the same model we assume that each data has given the initial data location and the replication policy. In this paper, we are considering various data replication methods to replicate and store data at various distributed servers. A local task of data is maintained by each of the datacenter in the network for the tasks that are programmed to a particular datacenter. Due to the costly data transfer delay in geographically distributed datacenters, a task can only be planned on a datacenter which has that requirement of data. Temporarily, the completion of a particular task is scheduled by the latest completion of all its tasks in the system. As in the network if any one of the task is completed and stored in the datacenters then that task will remain in datacenter until it will be called. The global scheduler must be having all the tasks from various local schedulers in each datacenter. The above described model is related to various cloud platforms. For example, study shows that Amazon Web Service (AWS) provides datacenters in multiple geographically distributed manners where datacenters are collected by their nearest zone. Each zone of this datacenters provides users with some specific local storage and also computes various nodes available from any other location of datacenters.

5. RELATED WORK

Moise W. Convolbo, Jerry Chou, Shihyu Lu and Yeh Ching Chung have proposed a novel data replication aware scheduling algorithm which can influence the existing replicas to minimize the average completion time of job submitted to a geo-distributed system[4].
Guoxin 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]. Viswanath *et al.* found that social links can grow stronger or weaker over time, which supports *SD*3's strategy of periodically checking the necessity of replicas. Also shows that different atomized user data has different visit/update rates, which supports the atomized user data replication in the basic design of data replication in datacenters [5]. The consistency is maintenance of replicas over geographically distributed datacenters or within a datacenter is done by using user interaction in social network. These works focuses on providing replicas on the geographically nearest server [2]. 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. Online social networks are categorized by the existence of different communities based on user communication, with a high rate of interaction between communities and low rate of interactions outside [2]. For very large online social network, the network communities are become untight, which leads to idea behind the result of data replication in various datacenters to create replicas based on different user communication and update rates rather than static friend communication.

6. CONCLUSIONS

In this paper, we have concluded that the process of replication on the data storage and to replicate a particular data from various datacenters. We have also proposed various data replication techniques to improve the performance of the model. As we have proposed approach for cloud applications to replicate their data across datacenters to prevent users from their data loss and also to provide service availability. Our method is consists giving priority to datacenters is based on the data stored to nearest server. Then we attempt various data replication method and compare them to approve our remote replication method.

References

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

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

[4] Moise W. Convolbo, Jerry Chou, Shihyu Lu and Yeh Ching Chung, "DRASH: A Data Replication-Aware Scheduler in Geo-distributed Data Centers," IEEE 8th International Conference on Cloud Computing Technology and Science, pp. 302-309, 2016.

[5] Hourieh Khalajzadeh, Dong Yuan, John Grundy, Yun Yang, "Improving Cloud-based Online Social Network Data Placement and Replication," IEEE 9th International Conference on Cloud Computing, pp. 678-685, 2016.

[6] C. L. Abad, Y. Lu, and R. H. Campbell, "Dare: Adaptive data replication for efficient cluster scheduling," in *2011 IEEE International Conference on Cluster Computing*, Sept 2011, pp. 159–168.

[7] M. Zarina, F. Ahmad, A. N. bin Mohd Rose, M. Nordin, and M. M. Deris, "Job scheduling for dynamic data replication strategy in heterogeneous federation data grid systems," in *Informatics and Applications (ICIA),2013 Second International Conference on*, Sept 2013, pp. 203–206.

[8] C.-C. Hung, L, Golubchik, and M. Yu, "Scheduling jobs across geodistributed datacenters," in *Proceedings of the Sixth ACM Symposium on Cloud Computing*, ser. SoCC '15. New York, NY, USA: ACM, 2015, pp. 111–124.