Survey paper on replication strategies in cloud computing

Aditya Anand[1], Karan Rajiv Redkar[2], Dr. R. Sumathi[3], Prof. R. M. Savithramma[4]

[1,2]4th Year Students, [3]Professor, [4]Assistant Professor
Department of Computer Science & Engineering
Siddaganga Institute of Technology, Tumakuru, India

Abstract - Cloud computing is one of the most recent domains in the field of computer science. It is growing exponentially day by day because of its features like virtualization, universal access, low cost and fault tolerance. Replication is one of the methods by which the cloud data center is able to provide effective fault tolerance. Replication is a method in which a replica of the original file or a part of the whole file is created. It is a crucial and effective strategy because it enables user to access the files even when the original file is damaged or even destroyed. Hence, it increases accessibility to the data stored in cloud center. In this paper we present the study of various techniques for replication.

1. INTRODUCTION

Cloud computing is a type of Internet-based computing that provides shared resources and data to computers and other devices on demand. It is a model for enabling ubiquitous, on-demand access to a shared pool of configurable computing resources. It can be rapidly provisioned and released with minimal management effort. Cloud computing allows companies to avoid up-front infrastructure costs. As well, it enables organizations to focus on their core businesses instead of spending time and money on computing infrastructure.

Data Replication is the process of storing data in more than one site or node (Data center). This is necessary for improving the availability of data. Data replication ensures availability, accessibility and fault tolerance. In data replication the whole data can be replicated or a fragment of the file (which is capable of reconstructing the whole file) is replicated. If the data retrieval from the original location fails, the data can be retrieved from the secondary location. This is how availability of the data is ensured through replication. If a query is to be processed, it could be processed at both locations leading to faster query execution. Data replication ensures that there is less data movement across the network . The more replicas of, a relation are there, the greater are the chances that the required data is found where the transaction is executing. Hence, data replication reduces movement of data among sites and. increases speed of processing.

But when a data is replicated, all the copies of the data must be updated together otherwise the functioning would not be effective (consistency of the data is lost). There are few clampdowns to replication such as, replication process obviously would require more space than normal and is little expensive. But even with these constraints, replication is a widely used strategy. In case of replication, it proves that, the advantages outweigh the disadvantages.

Virtualization is a key feature which gives an unsurpassed advantage to the cloud. Virtualized environment allows multiple applications to be run on the same platform at the same time and overcomes the physical limitations of any type of platform. User request are served by creating a **Virtual Machine (VM)**. Under some circumstances, VM needs to be moved from one data center to the other, this process is known as Virtual Machine Migration (VMM). VMM methods are divided into two types [1] : hot (live) migration and cold (non-live) migration. The status of the VM loses and user can notice the service interruption in cold migration. VM keeps running while migrating and does not lose its status; user does not feel any interruption in service in hot (live) migration.

There are two major approaches in live migration : post-copy memory migration and pre-copy memory migration. In post-copy VMM process, first VM execution is suspended at the source, copies minimal processor state to the target node, resumes the VM execution at destination, and begins fetching memory pages over the network from the source. In pre-copy approach, pages of memory are iteratively copied from the source machine to the destination machine without stopping the execution of the VM being migrated.

2. RELATED WORK

Autonomic data replication [1]

Local Machine: An Application is executed on the local machine running on windows 7 32-bit platform. This application is developed using java Eclipse API. The data generated by this application is sent to HDFS which stores it on multiple locations. While performing experiments, we kept replication factor to be 2. Application access the file system using the HDFS client, that exports the HDFS file system interface.

Master VM: It runs over Ubuntu 12.04 platform. Hadoop 1.0.4 is set up on this VM. NameNode and the DataNode runs on this VM. Data is received by the NameNode and replicated to the DataNodes depending on the replication factor. To increase reliability and availability the replication factor can be increased. NameNode keeps track of which DataNode is live so that when one DataNode is down data can be fetched from the other one.

Slave VM: It runs over the same Ubuntu platform. Same Hadoop 1.0.4 is set up on this VM also and it runs the second Data node. This node receives data from the master VM. Whenever Data node on the master VM fails then Name node automatically fetches data from this data node.

Optimized data replication for small files[2]
There is a limitation on file size are to be replicated or divided into several files and then save into the clouds. If the file is large then it is divided into several small files and then saved. But when it comes to small files, the files are simple replicated to all the clouds. This reduces thrisk of losing the data. A metadata file is maintained which keeps the record of all the files which are replicated and stored in the cloud. Before downloading the complete file, the information is taken from the metadata, all the small files are combined and then downloading begins.

Job scheduling in data replication[3]
If many users want to access same data, the waiting time for the users increase which will reduce the performance. To counter this problem data is replicated into many clouds so that users can access the same data from different clouds. The problem of waiting time is removed and hence the performance is increased.

Dynamic selection of nodes replication[4]
Replication of the files is done by 2 strategies. Those are static and dynamic. In static, the number of replicas and their location is set prior to uploading the file. Whereas in dynamic, the replicas are made keeping in mind the load conditions and the environment. In static, the user specifies the number of replicas. GFS algorithm is used for replication. Initially, all the data is stored in chunk server. Further, the user specifies which data is to be replicated and where it is to be replicated.

Efficient data replication in data centers [5]
The power consumed by the data center is tremendous. 45% of power is spent on cooling process, 15% is spent on power distribution, and 40% is left for computation servers. This issue of power consumption can be removed by using a strategy : the components which are idle (not in use) are put into sleep. By doing so, a lot of power is saved. Data replication process is carried out during this idle period, which results in power efficient data replication in data centers.

Replication in wide area distributed networks[6]
There are two types of replication protocols : Synchronous and asynchronous. Synchronous protocol updates all the replicas as soon as they are uploaded. They totally depend on scheduling. But in asynchronous, only a part of replicas are updated and the other replicas are updated after some time. This protocol is also known as lazy replication and synchronous protocol is known as eager replication.

3. CONCLUSION

The paper summarises the survey of replication techniques in various environments. Every strategy has presented its own terms for evaluation. Each technique gave an oration to some of the issues and tries to rectify them. This survey may help in future improvisation in replication, particularly in the field of cloud computing.

REFERENCES

[1] Autonomic Data Replication in Cloud Environment, Dhananjaya Gupt, Mrs. Anju Bala,Computer Science and Engineering, Thapar University, Patiala, India
[2] Optimized Data Replication for Small Files in Cloud Storage Systems,Xiong Fu, Wenjie Liu, Yeliang Cang, Xiaojie Gong and Song Deng
[3] Job scheduling and data replication on data grids,Ruay-Shiungchang',Jih-Sheng Chang,Shin-Yi Lin
[4] DATA REPLICATION STRATEGIES IN WIDE AREA DISTRIBUTED SYSTEMS,Sushant Goel