ISSN 2319 - 6629

Volume 2, No.5, August - September 2013

International Journal of Wireless Communications and Networking Technologies

Available Online at http://warse.org/pdfs/2013/ijwcnt01252013.pdf

A Brief Analysis on Architecture and Reliability of Cloud Based Data Storage

Tejaswini S L Jayanthy
Tejaswini.jsl@gmail.com
Dept. of IT,GIT, G.U

Illuri V Jagadesh
illurivj@gmail.com
Dept. of IT,GIT,G.U

DesarajuPratyosh
prat.desaraju@gmail.com
Dept. of IT,GIT,G.U

AdicharlaSrujanSimha
simha.srujan@gmail.com
Wipro Technologies,Kochi

ABSTRACT

Cloud based data storage is one of the most recent approach of storing the data after a forty year domination by relational databases. A common architectural design of a cloud based distributive storage system is proposed in this paper. It also discusses about the general approach in considering a cloud service, a detailed description about various components and a few important technologies available and few problems with respect to the reliability are mentioned. It provides the services to the clients via the Internet; moreover, the advantage is that the clients need not know the minute and inner details or various mechanisms running within the upper layers. The proposed architecture is a structured and layered one which is cooperative as well.

Keywords: Distributive storage system, Metadata management, Virtualization, Load Balancing

1. INTRODUCTION

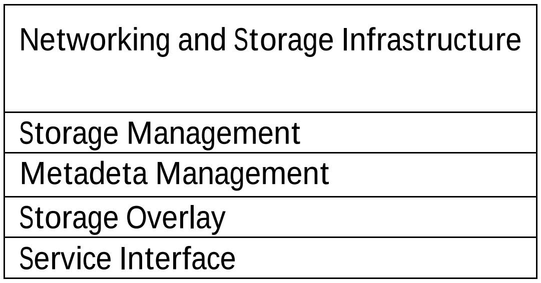
Cloud computing [1] can be addressed to by different people in different ways. Few such common characteristics which maximum interpretations say are scalability and completely reliable pooled resources used for computing, fully secured access to services from almost anywhere in the world, and transfer of data and services from various parts of the organization.

Though major breakthroughs have taken place in this area, it is still under development and can take quite some time in reaching its complete potential.

Cloud storage will provide data storage services at a very low cost and with much more reliability and security. It can be called as a cooperation storage service system with numerous devices, many domains used in various application. The paper proposed has various sections dealing with the general architecture of a cloud storage system and then its reliability. Various examples will be cited throughout the paper.

Architecture - Cloud Storage

The architecture [2] of a cloud storage system differs from service provider to service provider but when all the services are analyzed a model can be constructed by taking common entities into consideration. Cloud storage is a composition of numerous storage devices connected or combined by network, various distributed file systems and few storage Middleware services to provide a secured and reliable cloud storage facility. The structural model of cloud storage includes resource pool for storage, varied distributed file system, agreements with respect to service level, and interfaces, etc. Internationally, it is divided by logical and physical functions and their boundaries and relationships which provide compatibilities and may be interactions if required. Accordingly, the architecture design proposed here is :



Networking and Storage Infrastructure

Storage Management

Metadeta Management

Storage Overlay

Service Interface

Figure 1: Cloud storage model

In networking and storage infrastructure layer, distributed wired and various wireless networks, storage devices related networks are present.

In storage management layer, physical location based distributed storage resources are grouped by their domains and available logical entities, the data can be stored as blocks or as files in storage media available.

The metadata management layer combines the global data storage domain's metadata information and it collaborates it with different domains.

The storage overlay layer consists of the retrieval and redirecting operations which take place within the cloud. It can be considered as middle level layer which links the storage devices connected to the cloud with the virtually present storage networks. It also exposes the services to the respective and required interfaces.

In the last service interface layer, it provides clients with uniform interface to access, and a filter if required so that the inappropriate clients do not get a chance to enter into the system.

The policies which have to be considered for construction of any cloud storage service must have requirement analysis done initially after which capacity prediction of the cloud must be determined and then the performance planning, its deployment, post deployment verification, and after successful verification distribution must be done. Post distribution proper maintenance of the service must be provided. Lastly, if at all any updations are to the service to give better performance ability then updatability must be provided to the cloud service. The ultimate goal for any cloud storage provider must be to construct an available, cooperative, reliable, secure, scalable, concurrency data storage service at very economical price.

In order to construct a cloud storage system, a lot of components must be combined together. Like, virtualization, integration and clustering must be used to combine the distributed storage devices with its related management software so that an infinite unified storage pool is available for the users. QoS(Quality of Service) is the important factor which one must consider during the construction which affects directly the storage performances. This factor considers storage rates, bandwidth, reliability, volume size, responsibility, security, etc.

2. TECHNOLOGIES OF CLOUD SERVERS

The major technologies of majority of cloud storage servers include various components from networks, servers, clients. It must support automatic management of services, data integration, distributed collaboration, certification, audit, assignment, etc. Few servers and their control measures will be discussed further.

1. Deployment

During the deployment phases of cloud storage requirements analysis, optimization and evolution, storage resource redirection and few other factors are considered.

The scalability of cloud storage must be based on requirements and technology depending upon the application. The storage networks which are common are connected by middle level layer called middleware with the overlay layer.

The physical locations can be selected depending upon the data requirements of the specific applications. The ground rule is access overloads and duplications of data must be done most efficiently. During the deployment mode the approximate cost must be estimated and it must be optimized to the maximum level without compromising on the quality of the service. Feedbacks must be taken from the clients and server handlers so that the improvements can be made which indirectly increases the performance.

2. Virtualization of Cloud Storage

It is continually applied to operation system, network, server, etc. Storage virtualization is mapping physical storage to its logical storage during data accessing. Cloud storage virtualization will hide the exact storage physical positions and the modes used in storage and about the technologies those have been employed in storage.

3. Availability

The availability means the availability of the cloud storage throughout runtime and even for recovery in case of any failure. Availability must be high so that it ensures the application's Quality of Service. Few file systems which are commonly in use such as CIFS, NFS, and GFS, etc. are used to make sure that the cloud storage system has the highest availability thus ensuring the best QoS.

4. Data Organization of Cloud Storage

The data can be organized in the cloud storage in database mode, or at file or block level. The database having the data can be a business product, or an open source database. In order to make the retrieval of the data easy, all the data is organized in the form of records. But it can manage only some particular types of data. According to the application processing the file level can be changeable and flexible. All the databases or files are based on the block level and its mainly lower storing data format. Semantics are ignored in cases of pure block level and it is combined with other methods of storing. The most recently employed storage mode is the object oriented storage and it can be intelligent if added to a few independent operations.

5. Data Migration

It means moving the data stored in a storage system to another storage system in physically different places. Its aim is to maintain harmony and to keep a load balance in the storage system. Initially a storage capacity is assigned to specific purposes but sometimes the storage capacity is used for some proportion values having the threshold values and such situations the data must be shifted into other cloud storage servers if required. And all the required pointers must be especially the old ones need to be updated with the new addresses of the metadata.

6. Load Balance

It deals with keeping the storage space available for the future applications in different devices which are used to store the data in the cloud storage server or system. Storage responsibility and availability can be increased globally because of maintaining the load balance. One effective mechanism is data migration in order to maintain load balance, but this approach might bring a lot of workload overhead to both I/O process and network bandwidth, and another disadvantage as it doesn't relieve proper access to concurrence clients.

One specific case of data migration is data replication where in the original data is stored. This approach is an apt resolution to the single point fault in the distributed cloud storage server system, which generally keeps redundant and multiple copies of the same data but in various other and different storage devices.

The most ideal cloud storage system must always create replicas of the necessary data on its own depending upon the analysis of the client access frequency and the workload on the storage server. But without any reliable distributed storage system cloud computing cannot be trusted and can be prone to a lot of risks which might affect the clients.

3. RELIABILITY

Reliability [2] of any cloud storage system is basically the security of the data stored and I/O performance during the retrieval. Although a lot of research has been carried out in this area reliability is one question which hasn't been answered and is always a question for anyone working with a cloud storage system. With the new technologies coming in everyday security of the data has become a major concern for the clients. And because of these new technologies the traditional methods of securing the data have become outdated and they cannot be applied in the present scenarios.

When reliability is a situation of creating a new cloud storage system three major features have to be considered: 1) Availability 2) consistency 3) Partitioned-tolerance. Every distributed storage system provides the basic common ACID properties but in case of the cloud storage database these three properties mentioned above are a must for any cloud storage system to be considered as a reliable cloud storage system.

There are a lot of technologies existing in the current market which provide the above features but the problem is that any two features can be provided at a single time and not all three.

Following are the notable implementations of the CAP properties. Few such products are Google's Big Table, Amazon's Dynamo and Apache Cassandra.

- **Availability + Partitioned-tolerance system**

Dynamo is a highly available and one of the most efficient key-value data storage system which provides an "always available" feature, solving scalability and reliability at Amazon.com.

It is strongly believed that every minute change in the data can affect the customer's trust very significantly. Amazon.com is a large company dealing with ecommerce operations throughout the world. And in order to provide the best customer experience it sacrificed consistency feature and provides high sense of availability especially in case of the financial services. To attain both scalability and availability simultaneously, Dynamo uses a very consistent hashing technique to index all the data in a more flexible and object version to keep a loose and user friendly consistency. Duplication of data is synchronized in due course when the clients access their data or a particular data for conflict resolution. The policy of this feature is "last write is right" . For decentralization, it adopts a peer-to-peer system in a circular fashion, and it makes sure that every node is having the same set of responsibilities when compared to his peers.

Due to this P2P design, this Dynamo tolerates almost all kinds of system failure, has high availability and it also allows service owners move through the data as per their wish.

- **Consistency + Partitioned-tolerance system**

Bigtable by Google is a product designed to store petabytes data across thousands of commodity servers. It has a dynamic data layout and an proficient latency of big-scale access to data, and many more applications of Google are developed basing on this. The data model used for the development of the Big Table is an assortment of prospective uses such that varied users can reorganize their basic data model by making modifications like adding or deleting or modifying a column. It can also cooperate with a framework called the Map Reduce which is used for operating on large-scale parallel commands and computations which are developed at Google, which in turn enrich the capability of the users' database. The major problem with this product is that Google does not show and release the source code of the developers. Hence, there are two new and different open source implementations called the HBase and Hypertable, which almost combine almost every feature and component of the Big Table.

- **Partitioned-tolerance system + Availability/Consistency**

Cassandra is a one of the most efficient open source products in the market for cloud storage service which is admirably scalable, finally consistent, well distributed, and perfectly structured key-value implementer. This product combines the existing characteristics Dynamo and also the data [5]model of the Big Table. According to the usage, the features or the modes of the product can be changed. Users can specify a specific requirement in consistency level for every operation. ZERO consistency gives the least latency but the read or write operation takes place asynchronously. Therefore, no guarantee is given with respect to data consistency.

Similarly, Cassandra will become a consistent storage when the client chooses a consistency level called ALL. This elastic configuration makes Cassandra user-friendly software.

CONCLUSIONS

In this article, we addressed about the architecture of a basic cloud storage system and one of its major concern – reliability and security [3]. Cloud computing being a major research area leads to a productive development of the information technology and an elastic resource management for a lot of companies and clients all over the world. On the other hand, there is always a natural restriction on any distributed storage system. Hence the three major features have been discussed and even their successful implementations [6] were cited as examples. While constructing a new cloud based storage system the necessary details to be considered and the key technologies have also been discussed along with their importance and relevance. According to the study the data model chosen in order to implement the service is a major decision and must be taken after a consideration of a lot of parameters. With the increase n technologies, it is clearly understood that the old traditional models can no longer be implemented and hence the data isn't secured and reliable. Only more research in this field can bring in new methodologies with which a service can be made having all the possible features and enough security such that any client would opt for it without a second though.

REFERENCES

[1] Research on Cloud Storage Architecture and Key Technologies by WenyingZeng, Yuelong Zhao, KairiOu.

[2] Reliability and Security of Large Scale Data Storage in Cloud Computing by *C.W. Hsu*, *C.W. Wang*, *ShiuhpyngShieh.*

[3] Guidelines on Security and Privacy in Public Cloud Computing by Wayne Jansen Timothy Grance.

[4] Cloud Based Distributed Databases: The Future Ahead by ArpitaMathur, MridulMathur, PallaviUpadhyay.

[5] A. Ailamaki, D. DeWitt, M. Hill, and M. Skounakis. Weaving Relations for Cache Performance. In Proc. VLDB, 2001.

[6]High Throughput Data-Compression for Cloud StorageBogdan Nicolae University of Rennes 1 IRISA, Rennes, France