International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 11, November-2014

A Review on Cloud Storage Performance to Improve File Accessing Efficiency

Vidhya N. Gavali
Computer Engineering
Dnyanganga College of Engineering and Research, Pune.

Rahul P. More
Computer Engineering
Dnyanganga College of Engineering and Research, Pune.

Suvarna D. Potdukhe
Computer Engineering
Dnyanganga College of Engineering and Research, Pune.

***Abstract*— Cloud storage is a type of service in which data is placed, maintained and backed up at remote site. Cloud Storage provides various features like usability, accessibility, disaster recovery and saving of cost. In cloud storage multiple users and devices are present that result in cloud network manager can not give guarantee for optimal status of each storage node. Also some files are re-uploaded into server that decreases the bandwidth and increases the server workload. So, to remove the redundant or duplicate copies De-Duplication technique is used in cloud storage. To optimize the transmission node performance the Index Name Server (INS) architecture is used. Beside from this file compression, chunk matching, real time feedback control and load balancing techniques are also discussed. Using these techniques the cloud storage performance increases also the storage workload is reduced.**

Keywords— Cloud storage; De-Duplication; Index Name Server (INS);

I. INTRODUCTION

In cloud storage, data is placed, managed and also backed up at remote site. The data is available to the user over a network through internet from any location. There are various cloud storage providers some of popular providers are Google Drive (provides up to 5 GB free space), Microsoft Sky Drive (provides up to 7GB free space) and Amazon.

Data is stored in the space provided by third party companies and storage space is integrated and distributed through centralized management. So that multiple users can access data at a time [1].There are two different terms compute and storage. To decrease the bandwidth and networking cost ,it is beneficial to use computation of data instead of bringing data to computation (e. g if we want to send a file ,then instead of sending a file to individual only send a link to receiver). In cloud model, compute to storage has been applied using restricted programming like key-valued pair, MapReduce [2].

For cloud storage two protocols are used i.e Network Attached storage (NAS) and Storage Area Network (SAN). However, because of large number of users and devices in the cloud network, the efficiency of storage node can not be managed by cloud network manager that will decrease the effectiveness of cloud means network traffic and complexity of hardware is also increased [3] .

To increase the performance of cloud storage there are various methods have been used like file chunking and data compression. Also the file is re-uploaded to the server which affects the network bandwidth as well as workload of server, result will wastage of resources. The multiple users are present in cloud storage, so there is possibility that they can access the same file, operate the same function. Because of this, the cloud network manager does not give at each time optimal status of storage node [2].

To overcome these problems, we are going to discuss various methods suggested by different authors. To remove redundant data De-duplication technique is used. The Index Name Server Architecture (INS) is used for Wireless Area Network which is similar to Domain Name Architecture [3].

II. CURRENT TECHNIQUES IN USE

Instead of storing data on single server, cloud storage refers to use third party provider. The cloud storage interface is installed based on client requirement to different storage nodes. So operating in cloud storage is similar to local storage operating device. Using various network devices cloud storage changes the Application Programming Interface (API) of cloud storage (like Simple Object Access Protocol(SOAP), Representational State Transfer (REST)) [8]. For achieving good result many of the researchers have suggested to use De-duplication and Feedback control schemes.

Ohsaki et al. [9] The term job manager, an independent program generated according to different demand and Feedback control system is used from Resource Management Mechanism which is used to distribute resources and to manage Quality of Service. The job manager is generated according to a different demand. Thus as increase in data, multiple job managers might be generated, which increases the server workload.

Dezhi Han et al. [10] Introduced the Quality of Service controller to regulate bandwidth of accesses it also monitors the bandwidth for load distribution to distribute load when storage system is overloaded. However, heterogeneous domains are not considered as well as metadata server is a not a specific server, this affect on performance of cloud storage.

Jianzong Wang et al. [11], for adjusting and balancing the load according to different service levels introduced the Service Level Objectives (SLOs). It monitors the delay and input/output commands and achieving distributed storage.

Then it is applied to virtualized storage devices. But there is need to check input/output performance optimization and performance also testified whenever applying to WAN cloud network.

To check how changes in configuration of controller affect the system, the utility function and predicator function is used which monitors and distribute the resources of the system [7]. It gives better reading or writing efficiency and load balancing efficiency using De-duplication mechanism. But, block size and duplication ratio is fixed. Thus there is more need to perform experiment to test its better efficiency in WAN cloud network. To overcome above problem a hybrid approach is used to increase the performance of cloud storage. The hybrid approach uses Index Name Server architecture which integrates De-Duplication and Feedback control system to optimize storage node performance.

III. TECHNIQUES TO IMPROVE FILE ACCESS EFFICIENCY IN CLOUD STORAGE.

Several Techniques to improve the file access efficiency in cloud storage.

- Secure De-Duplication
- Index Name Server
- Feedback Control System

These Techniques are very useful to increase the cloud storage performance in case of file accessing.

A. Secure De-Duplication Technique

Server workload increases due to duplicate copies of repeating data. To remove this redundant data, De-duplication is an essential data compression technique. De-duplication not only removes duplicate copies but also provides secure access to file, reduces cost of storage and also saves the bandwidth. In previous De-duplication techniques are unable to provide this features.Fig.1 shows the secure de-duplication architecture.

De-duplication architecture contains three entities:

1. User
2. Private Cloud
3. Public Cloud

The user wants to upload data on storage –cloud service providers and later want to access data from public cloud. The private cloud is to provide secure file access to the user. And the public cloud uses the storage –cloud service provider to provide data storage service. It has an important role to reduce cost of storage by removing redundant copies. Also it saves the bandwidth by keeping only unique data.

In Fig.1, a group of clients i. e employees of company are placed at highest privilege level, they use storage server providers to store data with De-duplication. Data backup and disaster recovery are main features of cloud storage, De-duplication is used with these features, it checks the contents of two files, if contents are same then it stores only one copy of them.

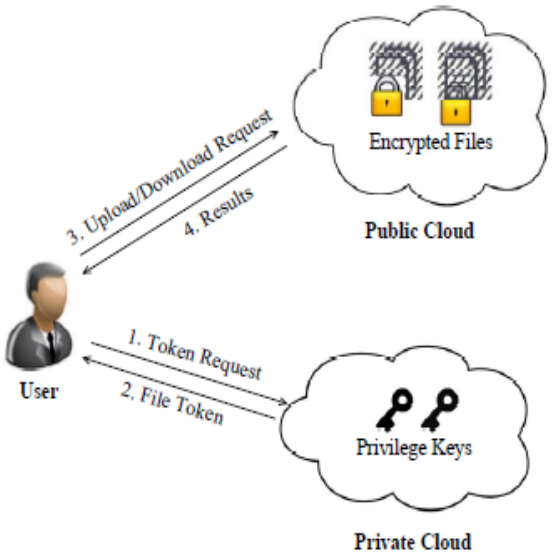

Fig. 1: Secure De-duplication Architecture [12]

Privilege levels provide the secure access to a file. There are various types of privilege levels (for example, time based, job position based, role based privilege levels). The users can send a token request to private cloud, they get the privilege key, considers only a file level De-duplication for simplicity means provides the file level De-duplication. If file is already presents then user get result back.

In secure De-duplication architecture both private and public cloud is used. This hybrid cloud approach is much popular today. The public cloud for example, Amazon S3 is used to keep storage data. The Private cloud, it is provided by third party and also implements hardware security features and remote execution trusted by users.

Tin-Yu Wu et al. [3] Proposes a de-duplication techniques at client side, yet reported literature have aimed to delete duplicate data at client side. They use a de-duplication scanning process to improve the system performance and also decrease the bandwidth taken for data transmission.MD5 function is used in De-duplication which divides the files into chunk and generates a unique 128-bit hash code. MD5 is a hash function used in hash algorithm to convert variable length into a unique fixed size digital signature. The signature is a collection of random characters and numbers. For each file chunk, there is a unique signature.

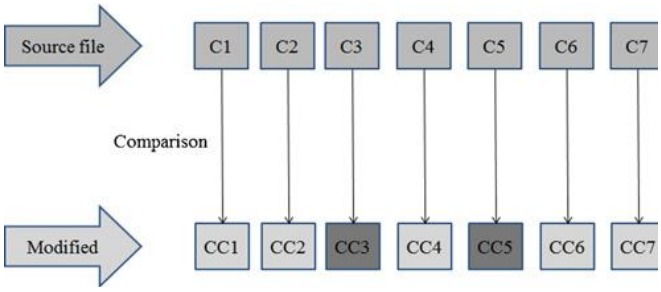

Fig.2: File Chunk Matching and Comparison [3]

B. Index Name Server

Index Name Server (INS) architecture is like to Domain Name system (DNS) and it manages data similar to P2P system. Although INS works similar to DNS, there are some functions of Index Name Server works like:

1. For transmission fulfill the user requirement.
2. To switch between storage node and signature of that storage node.
3. To confirm load balancing of storage nodes.

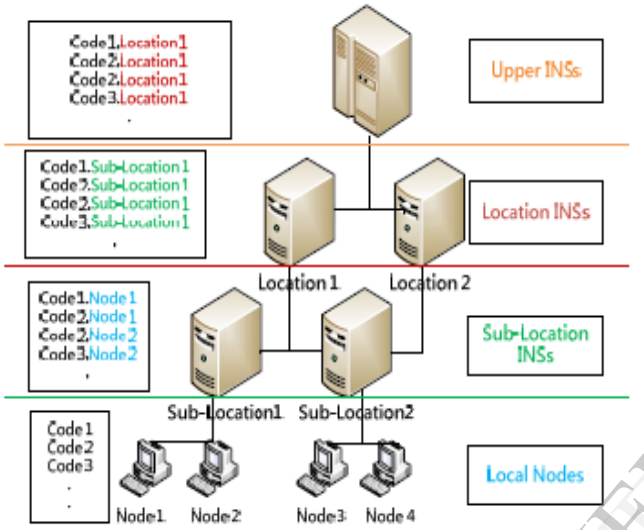

Fig. 3: Hierarchical Index Name Server Architecture [2]

INS has a hierarchical architecture which performs the one to many matches. Upper layer provides the services to lower layer. Every Index Name Server has its own database. According to database INS have stack structure similar to domain Name System. The database consists of storage nodes and signature to optimize the performance of the transmission node.

Fig. 3 describes the hierarchical architecture of Index Name Server. This INS architecture maintains client server relationship with each other to obtain the signature and storage nodes of all data chunk .Every storage node give its condition and data for record.INS record the information of location instead of keeping record of file chunk. The INS finds out maximum throughput storage node for back up. Also focus on computing and transmitting data [2].

To reduce burden on INSs and for load balancing the hierarchical INS architecture is used because when very few INS used to monitor the file system in wireless area network of cloud, workload of INSs will increase.

Fig. 4 shows the Index Name Server flow chart.

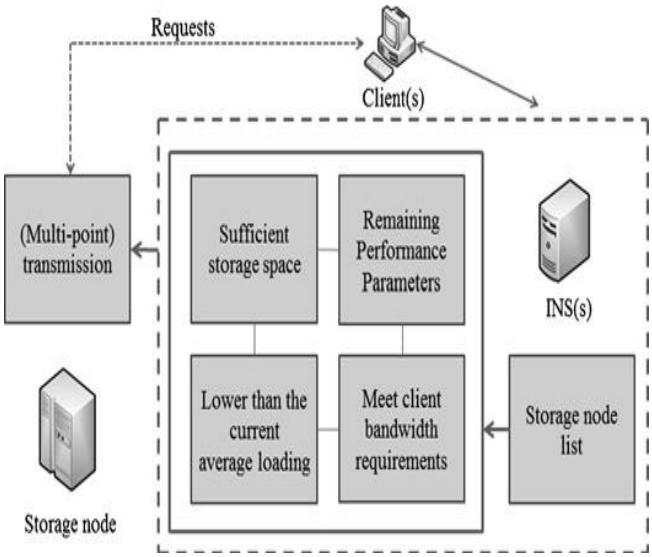

Fig.4: Index Name Server Flow Chart [3]

C. Feedback Control System

The role of storage node in cloud storage is very important. The various performance parameters affect on network. The parameter metric of having unit files / s is used to define the best performance of each storage node. These parameters are bandwidth, CPU/RAM performance and read/write capability of storage hardware [2].

The R (k) is a client-side parameter which is used to calculate bandwidth of storage node. In feedback system, system modifies the transmission parameter according to the result of previous transmission to increase the system performance. According to that it assigns the best storage nodes to the client but in case of some external interrupt for example network delay, the transmission value is not equal to actual bandwidth. In this case, there is possibility of selecting inaccurate storage node and leads to wastage of resources.

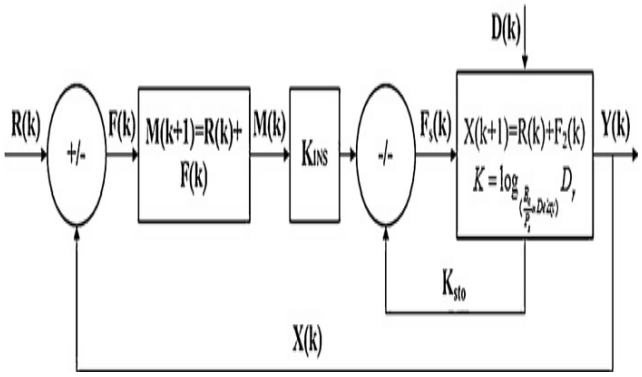

Fig. 5: Feedback Mechanism for INS Controlling Process [3]

To avoid this problem, feedback mechanism is used. Fig. 5 describes the feedback mechanism for INS controlling process. The following terms are used as:

1. *R* (*k*): The initial expected value
2. *F* (*k*) : The output feedback
3. *M* (*k*) : The modified feedback
4. *F*s (*k*) : The modified internal function of the storage node
5. *D* (*k*) : The external interference factor (random variable)
6. *X* (*k*) : The result within the storage node
7. *Y* (k) : The actual result
8. *K*INS : The optimal node determined by the INS based on the feedback

Index name server uses the client side parameter R (k) which computes the bandwidth which is used by client for client for transmission. The system adjusts the parameter according to the modified feedback M (k) and allocates the suitable storage node [2].

The performance parameters of storage nodes:

The parameters of storage node play a significant role to achieve better efficiency of storage node. The efficiency of storage node based on maximum bandwidth available is [3]:

$$B_s = B_c \ / \ [N\text{download} + (N\text{upload} - Fu)] \times (1 - Fd)] \quad (1)$$

1. B_S = bandwidth provided by storage node.
2. B_C= Bandwidth used by client
3. F_u = No. of duplicate copies determined by INS database.
4. F_d = Network delay time after transmission.
5. Ndownload and Nupload = are the number of files that the client will download and upload resp.

Then, transmission completion time for every second is [3]:

$$D_t = B_c / P_s \quad (2)$$

If Delay is occurred then probability per second is [3]:

$$P_{Delay} = (1 / D_t)^K \quad (3)$$

Finally getting actual data position [3]:

$$K = \log (B_C / P_S * \text{Delay})^D \quad (4)$$

Where, K - stream number, Ps –Packet size. B_c- client side bandwidth. In this way the performance of storage node is increased by controlling delay time.

There are various techniques are proposed to improve file accessing efficiency of storage node. But using hybrid approach the performance of cloud storage increased much more. Thus today many of research is going to use two or more techniques together like Index Name Server used with De-duplication and feedback control system.

V. CONCLUSION

Data De-duplication provides the data security by using privilege levels and it also eliminates the duplicate copies to reduce the workload of server. The Index Name Server is useful network architecture to optimize the storage node performance and to improve the file access efficiency in cloud storage. Feedback Control System plays a significant role; it adjusts the transmission parameter by observing previous transmission to get optimal cloud storage performance. The aim of Today's research is to use the hybrid approach to improve the file access efficiency in cloud storage. Researchers might use two or more techniques together to obtain more optimal result.

REFERENCES

[1] T.-Y. Wu,W.-T. Lee,and C.F.Lin, "Cloud storage performance enhancement by real-time feedback control and de-duplication," in Proc. Wireless Telecommun. Symp., Apr. 2012, pp. 1–5.
[2] Y.-M. Huo,H.-Y. Wang, L.-A. Hu, and H.-G. Yang,"A cloud storage Architecture model for data-intensive applications" in Proc. Int. Conf. Compute. Manage. May 2011, pp. 1–4.
[3] Tin-Yu Wu, Member, IEEE, Jeng-Shyang Pan ,"improving Accessing Efficiency of Cloud Storage Using De-Duplication and Feedback Schemes", IEEE SYSTEMS JOURNAL, VOL. 8, NO. 1, MARCH 2014 ,pp.208-218.
[4] H. He and L. Wang, " P&P: A combined push-pull model for resource monitoring in cloud computing environment" ,in *Proc. IEEE 3rd Int. Conf. Cloud Comput., Jul. 2010, pp. 260–267*
[5] T.-Y. Wu, W.-T. Lee, Y.-S. Lin, Y.-S. Lin, H.-L. Chan, and J.-S. Huang, "Dynamic load balancing mechanism based on cloud storage",in *Proc. Compute. Com. Appl. Conf., Jan. 2012, pp. 102–106*
[6] J. Dinerstein, S. Dinerstein, P. K. Egbert, and S. W. Clyde, "Learning based fusion for data de-duplication", in *Proc. 7th Int.*
[7] L. B. Costa and M. Ripeanu, "Towards automating the configuration of a distributed storage system" ,in Proc. 11th IEEE/ACM Int. Conf. Grid Comput., Oct. 2010, pp. 201–208.
[8] Direct hosting of SMB over TCP/IP. Microsoft http://support.microsoft.com/kb/204279
[9] Ohsaki, H.; Watanabe, S.; Imase, M.; "On dynamic resource management mechanism using control theoretic approach for wide-area grid computing", in Proc. Control Applications, 2005, pp. 891-897.
[10] Dezhi Han; Fu Feng; "Research on Self-Adaptive Distributed Storage System", in Proc. Wireless Communications, Networking and Mobile Computing (WiCOM '08.), 2008 .
[11] Jianzong Wang; Varman, P.; Changsheng Xie; "Avoiding performance fluctuation in cloud storage", in Proc. High Performance Computing (HiPC), 2010.
[12] Jin Li, Yan Kit Li, Xiaofeng Chen," A Hybrid Cloud Approach for Secure Authorized De-duplication ", IEEE Transactions on Parallel and Distributed Systems,2014.
[13] Xin Sun, Kan Li, Yushu Liu, "An Efficient Replica Location Method in Hierarchical P2P Networks", in Proc. Computer and Information Science(ICIS 2009), 2009, pp. 769-774
[14] J. Yuan and S. Yu. "Secure and constant cost public cloud storage auditing with deduplication",IACR Cryptology ePrint Archive, 2013:149, 2013.