ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 3, March 2018

Efficient Cauchy Coding Schema for Distributed Storage System

Priyanka R kene[1], Prof. R. A. Kulkarni[2]

Student, Department of Computer Engineering, PICT, Pune, India[1]

Asst. Professor, Department of Computer Engineering, PICT, Pune, India [2]

ABSTRACT: In the current era of cloud computing, data stored in the cloud is being generated at a tremendous speed, and thus the cloud storage system has become one of the key components in cloud computing.Now a days cloud is usually use for storage system because of characteristics like broad network access , resource pooling , on demand model , measured service etc. Due to some unfavorable conditions like power surge, faulty spindle motor, mechanical damage or firmware defects users data may loss. The storage system should provide mechanism to protect user data. For this purpose replication is used. Though it is easy to store replicated data to tolerate a certain amount of data losses, it suffers from a very low storage efficiency. With data protection as main concernand also to reduce storage overhead cloud file system are transitioning from replication to erasure codes. Erasure code technique is used to protect data present over distributed system. Conventional erasure coding techniques, such as reed-solomon codes, are able to achieve a much lower storage cost with the same level of tolerance against disk failures. However, it incurs much higher repair costs, not to mention an even higher access latency. In this sense, designing new coding techniques for cloud storage systems has gained a significant amount of attention in both academia and the industry. So this paper introduce a new coding schema which improve efficiency of reed solomon coding technique which uses cauchy matrix to achieve fault tolerance.

KEYWORDS*:* distributed storage system, erasure code

I. INTRODUCTION

The data generated and stored in the modern world is increasing exponentially with time. With more and more people getting connected to the internet through various platforms like the social media, blogs and e-shopping, a huge amount of data is pushed to the storage of these platforms. Social media, e-shopping sites and several similar enterprises cannot afford to lose even a single bit of this information since it is critical to their business−they do a lot of analysis and studies on this data and base their strategies on the gathered information. With the arrival of smartphones, this pile of data achieves newer heights. Smartphones help people to get more connected, and hence every snapshot of every instance of celebration in their life, each video/audio clip they liked, a funny message that relieved them from their stress, a vital piece of information they wish to send across a group of people, all get pushed to these platforms.

The amount of data generated and stored on a daily basis is currently of the order of exa, peta bytes and is increasing further.so For large storage system, disk failures are common problem and that must be tolerated to protect data loss. Previously to protect data 'replication' of disk is widely used. Because of this huge size of data, replicating the data for ensuring availability has become impractical. The standard replication factor is 3 (Replica factor of 3 ensures that it can withstand two simultaneous failures.), implying a 200% overhead for storage. In addition to this, replication creates the problem of consistency among the replicas, since the active replication among the replicas are not effective due to various network related problems

Now a days data increases exponentially so there is high probability of loss and hence there is need of system which support data recovery option within time. Erasure coding method also used later on to achieve fault tolerance. But it produce limited level of tolerance which is no more sufficient in real application data. For distributed storage system chances of disk failure increases as storage system grows. So to handle more loss 'Reed Solomon' can be used.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 3, March 2018

In this paper, 'Cauchy Reed Solomon ' technique is used as the improvement on 'Reed Solomon' technique. This uses the XOR operation overcoming the Galois Filed Multiplication to make it less complex while encoding and storing the erasure as well as while recovering data from loss. This 'CRS' code converts 'k' data blocks into 'm' coding blocks and using new approach one can handle 'm' disk failure also.

To achieve this mechanism used : first to find 'Cauchy matrix' of data blocks and the best match pair is found after listing down all possible combination of data matrix and scheduler used to encode. But presently there is no such method which will find best match as data has different priority depend on user also redundancy of each data required is different. So option remains is to list down all and choose optimal also can make conclusion on basis of experimental result.

ERASURE CODING

Although erasure coding is related to RAID systems, factors like adaptability of fault tolerance to various scenarios, lower storage footprint for fault tolerance compared to mirroring and the ability to scale in a distributed environment makes erasurecodes superior to RAID systems . As a promising technology for cloud storage, erasure coding is a hot topic of research. A quick scan through the concept of erasure coding, follows. Given a data of size *M*, erasure coding is organized in the following manner. The data *M* is broken down into *k* equalsized information blocks. The *k* information blocks are then coded into *n* blocks of equal size. The *n* blocks contain *k*information blocks and $n - k$ parity blocks. These*n* blocks are written into *n* storage nodes. The storage nodes holding data are called *data nodes* and those having parity bits are referred to as *parity nodes.* In the case of failure of any $n - k$nodes, the lost data can be reconstructed using the remaining *k*nodes. By downloading the required blocks from k live nodes, new nodes or newcomernodes reconstruct the data of failed nodes.

The reconstruction mechanism acts as the technique for ensuring fault tolerance and hence data durability in erasure coded storage systems. When file request by a client is received by the master server, it reads the fragments of the file by referring to the meta data, combines the file contents and send the file to the client. But if any of the storage nodes holding the fragments has failed, then the master initiates data reconstruction to serve the client request. Such cases are called *degraded reads* as opposed to 'normal reads'. Since the coding facilitates the recreation of the data in an erased(failed) node, it is called Erasure Coding. Erasure coding makes use of the error correcting property of MDS codes. A linear code which meet the bound ($d <= n - k + 1$) is called a Maximum Distance Separable (MDS) code.

II. MOTIVATION

Data has different replication factor depending on user priority of data so to find best match as no combination of matrix and scheduler performs the best for all redundancy configurations and also
To find improved coding scheme for different combinations of matrix and schedule, Since there is a large gap in the number of XOR operations And find the schema for each combination which will finish operation in minimum time and less operations.

III. REVIEW OF LITERATURE

In this paper, focus is on to recover the randomly spread as in large storage system data is never placed at one contiguous place. So using proposed method one can find random and dissimilar data situated on one disk The proposed method is 'direct and Recursive' in nature and useful for spread erasure to handle errors on disk.[2]
In this paper , concept of 'RAID i.e. parity disk' is efficiently used. The parity disk is measured using ' XOR '. To recover the data some subset of data block is used to generate parity. In fact to recover data listing down all recovery equations to recover the lost disk data the two new XOR code are introduced.[3]

In this paper , the two new heuristics are derived which are useful for minimizing XOR operation which will ultimately make mechanism less complex and process in less time To 'Encode and Decode' this new schedulers are used on 'XOR operation' . so use if these efficiently decreases time required by other ones.[4]

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 3, March 2018

In this paper , focus is on 'Encoding' time required for XOR operation because it is mostly used operation compare to 'Decoding'. So author describes that it is crucial to protect data from loss due to failure in large system . A new algorithm is proposed for this purpose . They focus on minimizing the no. of XOR operation required and on cache behavior of CPU.[5]

In this paper, a new approach for cloud system is proposed i.e. A new cryptographic system which results in high availability and integrity of data stored in cloud. It can be said that it is extension to the RAID for distributed system like cloud. It ensures file availability against strong, mobile adversary. But it is limited for only static file also it does not provide security to file block level , it can be attacked. Different variants to HAIL are also proposed for distributed file system availability. [6]

In this paper, a new algorithm is proposed for the recovery of data if there is a loss .they proposed a 'Rotated Reed Solomon' algorithm which recover from minimal number of code words that are usually required for other XOR based Erasure codes. Hence it also improves the I/O performance. It achieves degraded read than other codes and also maintains the reliability and performance properties of RS codes.[7]

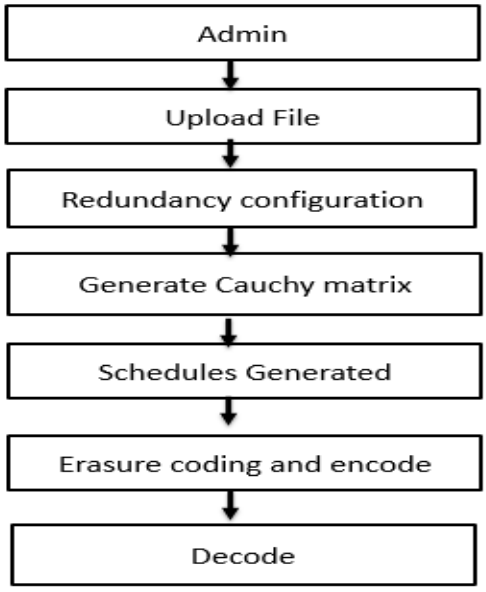
The paper is about the study of how important the RS code for any storage system . In older days, the only space optimal code use as 'Erasure codes' for small storage system . Then variant to 'Reed Solomon' i.e. 'Cauchy RS' come in 1990's. The algorithm is about optimizing the generation of Cauchy Matrix as it impact on Encoding Performance. The algorithm is about optimizing the generation of Cauchy Matrix. As it impact on Encoding Performance. The algorithm gives good Cauchy matrices for large systems also.[8]

In this paper, A new lowest density MDS code is proposed which is multi erasure array code. This erasure are found with minimal redundancy and minimum update penalty. A new property is define which is of type structure property . A simple multi erasure array code performs good. Further exploration is to develop efficient encoding and decoding algorithm for these code. [9]

In this paper , It is a new coding scheme which is more reliable than the previous coding schema to achieve fault tolerance. Beside that limitation of this schema is they can only use up to three level of failure of component . This also works using 'XOR' operations. Hence this work can be extended to achieve more faults/loss.[10]

IV. SYSTEM DESCRIPTION

It is a new approach using 'Cauchy Reed Solomon' technique to achieve fault tolerance. In this idea that listing down all the possible combination of data- matrix representation and different scheduler and choosing the best pair from them is applied. As there is no such rule/method is defined for finding the best combination. Workflow of this is shown :



ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 3, March 2018

As there is different need of 'redundancy configuration' depend on user data hence no conclusion can be made . Beside on basis of experimental result one can make the decision for same type of 'matrix calculation and redundancy configuration'.

As shown in fig. , Here three main entities are enlisted alongwith their interconnection. The Cloudserver reside between the data user and data owner.The data owner and data user have access permissions to upload and retrieve the data on cloud server.when user will upload a file it should provide the login credential's for authentication purpose.After Successful login user can upload file.Then file encoding and preprocessing of that file created.References(indexes) of first file created and the file is stored.when user again upload a second file then its contents are compared with previous uploaded files references.The duplicate data is removed and references are passed to save memory and seeking time.this process results into CPU utilization time.

Algorithm:

1] Stemmer Algorithm

It is use in normalization process mainly in IR system. It basically removes the 'stems' from words without changing meaning and reduces it to the root word.

Input: - Words. For ex: - studying

Process:-

Step 1: remove the plurals and -ed or -ing suffixes

Step 2: convert end letter y to i

Step 3:eliminate same suffixes to one: -ization, -ational, etc.

Step 4handle all suffixes, -full, -ness etc.

Step 5also removes -ant, -ence, etc. Step 6: eliminate final letter 'e'

Output:- Study

2] XOR Algorithm

The simplest implementation of Erasure coding is XOR operation. Lets assume X and Y and Z are data cell then parity cell is xor of these three data cells x Ł y Ł z so in XOR operation only one parity bit is generated and if anyone bit is lost it can be recovered by the remaining data cells and a parity bit. It is very limited since it produces 1 parity bit so XOR operation can tolerate only 1 failure with n group size. In Xor operation fault tolerance 1 and storage efficiency is n-1/n when group size is n.

3] Deduplication:

Remove the deduplication of files from cloud storage for purpose of save the storage space and improve the performance of cloud.

V. CONCLUSION

In this paper , After studying on different reasons of data loss and regenerating data using erasure code the parameter i.e. count of 'XOR operation' is found to improve the efficiency. Previous study has some limitation. Hence to improve fault tolerance for real applications data by finding new approach to recalculate the loss data using 'Cauchy Reed Solomon' technique.At long last, we promptly recognize that diminishing XORs is by all account not the only approach to enhance the execution of a deletion code. Other code properties, similar to the measure of information required for recuperation and corrupted peruses may constrain execution more than the CPU overhead.

REFERENCES

1. Guangyan Zhang, Guiyong Wu, Shupeng Wang, JiwuShu, Member, IEEE,WeiminZheng, and Keqin Li, Fellow, IEEE Caco: An EfficientCauchy Coding Approach For Cloud Storage Systems IEEE TRANSACTIONS ON COMPUTERS, Vol. 65, No. 2, February 2016
2. J. L. Hafner, V. Deenadhayalan, K. K. Rao, and J. A. Tomlin, "Matrix methods for lost data reconstruction in erasure codes," in Proc. 4th Conf. USENIX Conf. File Storage Technol.-Volume 4, Berkeley, CA, USA, 2005, pp. 14–14.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 3, March 2018

3. Kevin M. Greenan ,Xiaozhou Li , Flat XOR-based erasure codes in storage systems: Constructions, efficient recovery, and tradeoffs, ParaScale ,HP Labs IEEE 2010.
4. J. S. Plank, C. D. Schuman, and B. D. Robison, Heuristics for optimizing matrix-based erasure codes for fault-tolerant storage systems, in Proc. 42Nd Annu. IEEE/IFIP Int. Conf. Dependable Syst. Netw, Washington, DC, USA, 2012, pp. 112.
5. JianqiangLuo, LihaoXu, James S. Plank, An Efficient XOR-Scheduling Algorithm for Erasure Codes Encoding, Dependable Syst. Netw., 2009, pp. 504513.
6. K. D. Bowers, A. Juels, and A. Oprea, "Hail: A high-availability and integrity layer for cloud storage," in Proc. 16th ACM Conf. Comput. Commun. Security, New York, NY, USA, 2009, pp. 187–198.
7. O. Khan, R. Burns, J. Plank, W. Pierce, and C. Huang, "Rethinking erasure codes for cloud file systems: Minimizing I/O for recovery and degraded reads," in Proc. 10TH USENIX Conf. File Storage Technol., 2012, pp. 251–264
8. J. S. Plank and L. Xu, "Optimizing Cauchy Reed-Solomon codes for fault-tolerant network storage applications," in Proc. 5thIEEE Int. Symp. Netw. Comput. Appl., Washington, DC, USA, 2006, pp. 173–180.
9. C. Huang and L. Xu, Star: An efficient coding scheme for correcting triple storage node failures, in Proc. 4th Conf. USENIX Conf.File Storage Technol.-Volume 4, Berkeley, CA, USA, 2005, pp. 1515
10. S.Lin, G. Wang, D. Stones, J. Liu, and X. Liu, "T-code: 3-erasure longest lowest-density mds codes," IEEE J. Sel. Areas Commun., vol. 28, no. 2, pp. 289–296, Feb. 2010.