ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 6, June 2017

Optimizing Cauchy Technique to Improve CPU Performance in Cloud Storage System Using N-gram Method

Nilesh V. Wanare[1] , Prof. Kanchan Varpe[2]

M.E., Dept. of Computer, RMD Sinhgad School of Engineering, Pune, India[1]

Assistant Professor, Dept. of Computer, RMD Sinhgad School of Engineering, Pune, India[2]

ABSTRACT: In great storage systems, it is essential to protect data from loss due to failures. In case of cloud Storage Systems, data resides on various servers so it becomes very tedious task to manage and retrieve the data from servers. To resolve the problems like device failures, fault tolerance and CPU optimization some data storage mechanisms introduced to support the cloud data management.This proposed system concentrates on enhancing the CPU utilization and avoiding the data duplication by using the coding schemes. One of the Coding scheme CaCo uses Cauchy matrix heuristics to produce a matrix set. For each matrix set, it performs XOR schedule heuristics to generate series of schedules. At last it selects the shortest one from all produced schedules. Proposed system defined with N-gram model which is used for approximate string matching.This paper Focuses on enhancingthe CPU utilization and avoiding the data duplication by using the coding schemes.It uses stemming algorithm to generate the root words of strings and indices of root words are compared with the next file contents.

KEYWORDS: Cloud Storage, Fault Tolerance, Reed-SolomonCodes, Cauchy Matrix, XOR Scheduling

I. INTRODUCTION

Distributed storage is developed of various reasonable and untrustworthy parts, which prompts a decline in the general interim between disappointments (MTBF). As capacity frameworks develop in scale and are sent over more extensive systems, part disappointments have been more basic, also, prerequisites for adaptation to internal failure have been further expanded. Along these lines, the disappointment assurance offered by the standard levels has been no more adequate by and large, also, capacity planners are thinking about how to endure bigger quantities of disappointments . For instance,Google's cloud capacity, Windows Azure Storage Ocean Store, Disk Reduce, HAIL, and others all endure at any rate three disappointment. An overview of the CaCo approach is given and describeshow CaCo accelerates selection with parallel computing. An implementation of CaCo in a cloud storage and present the evaluation results on a real system. For different combinations of matrix and schedule, there is a large gap in the number of XOR operations. No one combination performs the best for all

redundancy configurations. With the current state of the art, from the $\binom{2^w}{k+m}\binom{k+m}{k}$Cauchy matrices, there is nomethod discovered to determine which one can produce thebest schedule. Giving a Cauchy matrix, different schedulesgenerated by various heuristics lead to a great disparity oncoding performance. For redundancypattern, itis with very low chance that one coding scheme selectedby rules of thumb performs the best.

Previous coding schemes for cloud data storage does notsupport the best redundancy configuration.This paper helpsto transfer the data over cloud with minimum number of operations using cauchy encoding technique. It manages data recovery, data duplication and results into CPU Optimization.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 6, June 2017

II. REVIEW OF LITERATURE

The XORs is not the only way to improve the performance of an erasure code. Other code properties, like the amount of data required for recovery and degraded reads, may limit performance more than the CPU overhead. This paper also describes the algorithms for XOR codes that can recover the failed disk. It also analyzes the efficiency of XOR codes and storage properties of XOR codes such as replication andRAID 6[1]. CRS coding involves only XOR operations, the number of XORs required by a CRS code directly upon the performance of encoding or decoding.. This paper studies the longest-density MDS codes which are multi-erasure array code with efficient redundancy and minimum update penalty.

We prove some basic structure properties for longest lowest density MDS codes[2].It is crucial to protect the data from loss due to failure in larger storage system, Erasure codes provides protection to data by encoding and deoding.This study focuses on improving encoding operations on XOR codes. It uses XOR

scheduling technique on XOR codes and conducts the performance evaluation on erasure codes. Enhancing the execution of encoding, the more regular operation. It does as such by planning the operations of XOR based codes to upgrade their utilization of store memory[3]. The later system addresses some failures on lost disk using erasure codes. It also observes the problems like recovery of lost data due to scattered or uncorrelated erasures and recovery of partial data from a single lost disk. This study provides methodology for scattered erasure to handle errors on disk. It defines that lost data isuncover able or it is declared uncover able and formula is provided for the reconstruction that depends only on readable sectors[4].The Cloud record frameworks are transitioning from replication to deletion codes to reduce the storage overhead. This study present an algorithm that finds the optimal numberof codeword symbols required for recovery of any XOR-based erasure code and produces recovery schedules that use a minimum amount of data. The Reed-Solomon format is new group that perform corrupted reads more efficiently than all known codes, but otherwise inherit the consistency and performance properties of Reed-Solomon codes[5].The proposed system defines the WAS architecture and global namespace as well as its resource provisioning, load balancing and replication system.In WAS data is stored durably using both local and geographic copies to ease disaster recovery. Recently WAS storage comes in the form of files, Tables and Queues (message delivery) [6].This system provides mechanisms for fault-tolerance over large distributed storage systems. Recently, erasure codes arebeing used for the replication, since they provide the same fault-tolerance for reduced storage overhead. However, their performance is unclear in a geographically diverse distributed storage system. This study compares the performance of triple replication with the erasure coding (Reed-Solomon codes) used in Apache Hadoop implementation of a distributed file system[7].This paper defines Cauchy Reed-Solomon codes to P2P streaming systems which are different from classical Reed-Solomon codes. It also focuses on concern about how to apply the coding into these systems and present the interaction among the peers and effective buffer management. Peer-to-Peer systems are widely used to share resources. In these case, data availability is the primary concern because peer dynamics is a prevalent phenomenon. In recent years, erasure codes, i.e. Forward Error Correction (FEC), prevent data loss including in transmissions and in storage systems. From a fault-tolerance point of view, the use of FEC reduces meantime of failures by many orders of magnitude, compared to replication systems with similar storage and bandwidth requirements[8].The proposed system states that efficiency of an erasure code using a bit-matrix is directly mapped to the number of XOR operations required for the encoding scheme. So the problem in the field of erasure coding is how to prepare the XOR operations for given matrix so that the smallest number of XOR operations are required. The system stated an algorithm for finding the optimum solution and analyzes the performance of two known heuristics on a set of encoding matrices[9].The system proposed that to minimize the payment cost of customers while guarantee their SLOs by using the worldwide distributed data centers belonging to different CSPs with different resource unit prices. Many Cloud Service Providers provide data storage services with data enter distributed worldwide. cloud customers of globally distributed applications (e.g., online social networks) face twochallenges: i) how to allocate data to worldwide data enters to satisfy application SLO requirements including both data retrieval latency and availability, and ii) how to allocate data and reserve resources in datacenters belonging to different CSPs to minimize the payment cost.This system first model the cost minimization problem under SLO constraints using integer Programming.Then it introduce our heuristic solution, including a dominant-cost based data allocation algorithm and an optimal resource reservation algorithm[10].To tolerate more failures than RAID for storage in cloud Systems used Reedsolomanfor fault tolerance.The , study proposed efficient Cauchy coding approach for data storage in the cloud. This approach used to reduce the redundancy and improve the accuracy by avoiding the faults over the previous

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 6, June 2017

scheme. The CaCo has the ability to identify an optimal coding scheme.Due to CaCos n behaviour of parallel processing, we improve considerably the selection process performance with rich computational resources in the cloud[11].

III. PROPOSED SYSTEM ARCHITECTURE

To improve storage system of cloud by avoiding data duplication and to enhance CPU performance with fault tolerance by using Cauchy coding encoding technique with extended n-gram method.In Proposed System CaCo, a new scheme that monitor all existing matrix and map heuristics, and thus results into optimal coding scheme within a given redundancy configuration. The selection process of CaCo has an adequate complexity and can be accelerate with parallel computing. In proposed system, N gram model is used to improve system performance, reduced CPU overhead, achieving fault tolerance.

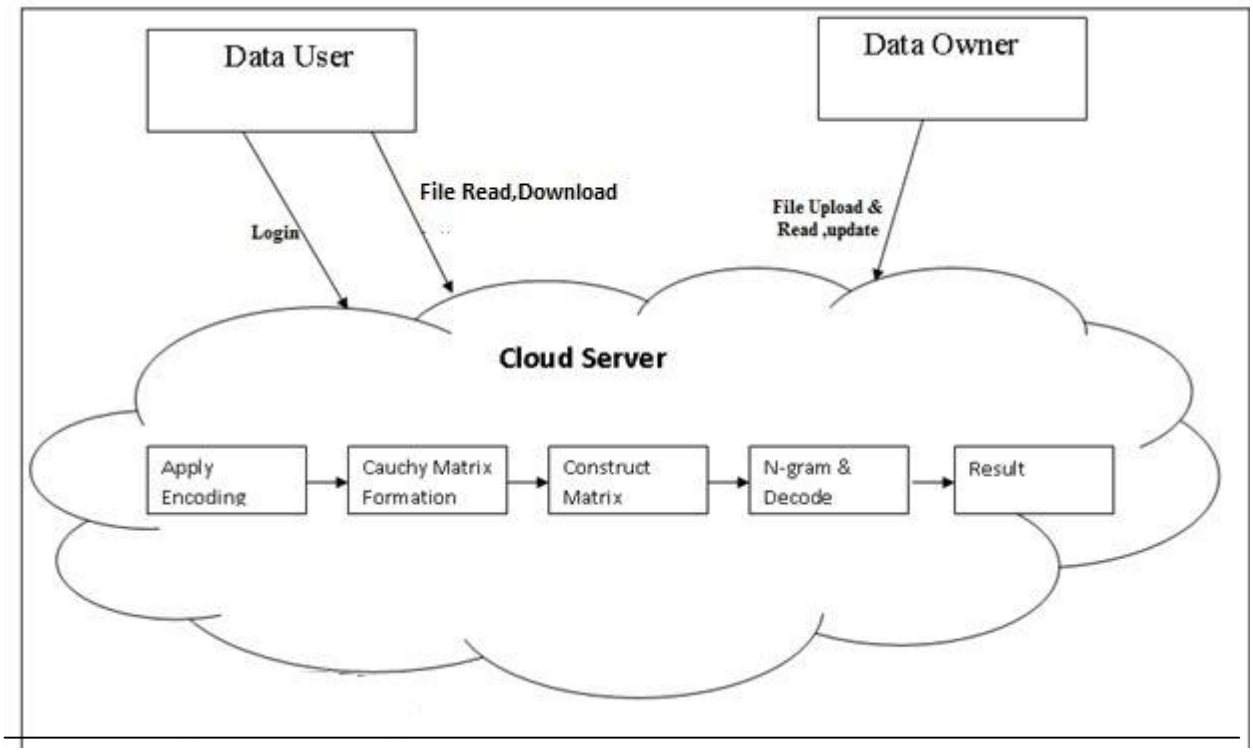


Fig 1.System Architecture

The system architecture shows the flow of proposed system .Here three main entities are enlisted along with their interconnection.The Cloud server reside between the data userand 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 credentials for authentication purpose. After Successful login user can read file and download. Then file encoding and n-gram creationof that file created. References (indexes) of first file created and the file is stored. When user again uploads 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.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 6, June 2017

A. Advantages of proposed system

1. CaCo N gram model helps to find the existing files on the server and it helps in reducing the data duplications, CPU overhead,fault tolerance Cloud storage systems always use different redundancy configurations.
2. Due to its parallel processing, maximum utilisation of computing resoures is possible.
3. Cloud storage systems always use different redundancy configurations depending on the desired balance between performance and fault tolerance.

IV. SYSTEM ANALYSIS

In this section different factors relevant to system behavior is defined.This factors defines the characteristics of the system.

A. Proposed Algorithm

It defines the alogorithm steps for computing the N-gramof a given string.It finally genetates the root word of a givenstring.

Stemming Algorithm:

Information Retrieval (IR),Stemming is very useful toolwhich is supported by almost all indexing and searchsystems.The algorithm proposed by stemming is a mechanismthat decrease words with the same stem to a common form,it removes the derivational and inflectional suffixes fromeach word . For example the words study, studies, studied,studying, student or studious are decrease to the root word
study. The information retrieval contains grouping of wordsinto common form .The development of stemming algorithmsfor freetext retrieval purpose is evidenced by the work ofmany researchers. A stemming algorithm reduces the words”fishing”, ”fished”, and ”fisher” to the root word, ”fish”Themain purpose of stemming is to decrease different syntacticalforms / word forms of a word such as its noun, adjective,verb, adverb etc. to root form.

Steps: Input:- Words. For ex:- Semantically
Process:-
Step 1: Gets rid of plurals and -ed or -ing suffixes
Step 2: Turns terminal y to i when there is another vowel in
the stem
Step 3: Maps double suffixes to single ones: -ization, -ational,
etc.
Step 4: Deals with suffixes, -full, -ness etc.
Step 5: Takes off -ant, -ence, etc.
Step 6: Removes a final e
Output:- ”Semant” which is root word finally generates
.

B. Module Description

1. User Module: It is the user Authentication with thecloud server. Each user has a rights to read and download the files. User can see the existing files provided by the data owner.

2. Data Owner Module: It is responsible to registration,login and upload the files. Data owner sets the permissionson the files so that it should retain the data integrity.

3. Server Module: It provides services to anonymousauthorized users. It interacts with the User during theauthentication process.It handles the data flow between theuser and data owner.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

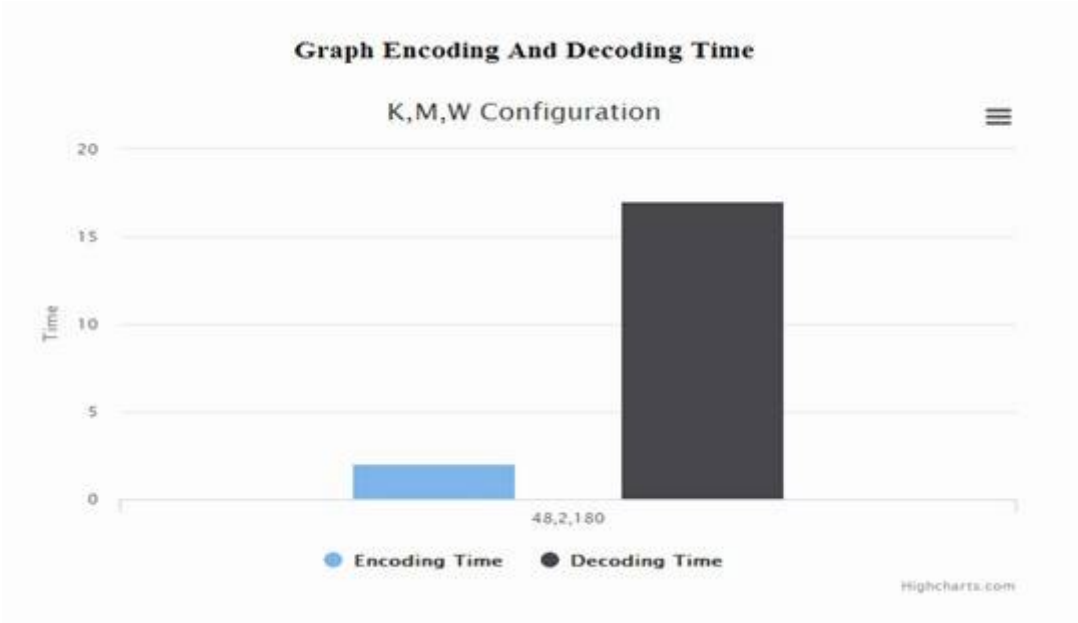
International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 6, June 2017

C. Result



Result Table:-

K,M,W configuration	Encoding Time	Decoding Time
48,2,180	2.0mm	17.0mm

Table 1:Redundancy Configuration for 48,2,18.

V. CONCLUSION

Cloud storage systems always use different redundancyconfigurations depending on the desired balance between performance and fault tolerance.CaCo N gram model helps tofind the existing files on the server and it helps in reducingthe data duplications, CPU overhead, fault tolerance. Due toits parallel processing, maximum utilisation of computing resources is possible. It can recover the corrupted datablocks and code blocks. N gram model is used mainly forapproximate string matching and string operations. In thisway CaCo can be extended to improve the CPU performanceand achieves the fault tolerance.

VI. FUTURE WORK

The system can be improved with advanced cloud storagetechniques to achieve the file fast processing on cloud storage

ACKNOWLEDGMENT

First,i would like to thanks all referies and reviewers for helping me in my research work.I take this chance to expressmy appreciation to my guide Prof.KanchanVarpe and head of department, Prof. V. M. Lomte, Department of Computer Engineering, RMDSSOE, for their kind cooperation and guidanceduring the entire research work. I would also like tothank our, Principal and Management for providing lab and
other facilities.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 6, June 2017

REFERENCES

[1] Kevin M. Greenan ,Xiaozhou Li , "Flat XOR-based erasure codes in storage systems:Constructions, efficient recovery, and tradeoffs", ParaScale ,HP Labs IEEE 2010.
[2] Sheng Lin, Gang Wang, Member, IEEE, Douglas S. Stones, Xiaoguang Liu and Jing Liu, "T-Code: 3-Erasure Longest Lowest-Density MDS Codes", VOL. 28, NO. 2, FEBRUARY 2010.
[3] JianqiangLuo, LihaoXu, James S. Plank," An Efficient XOR-Scheduling Algorithm for Erasure Codes Encoding", Dependable Syst. Netw., 2009, pp. 504–513.
[4] James Lee Hafner, VeeraDeenadhayalan, and KK Rao," Matrix Methods for Lost Data Reconstruction in Erasure Codes", IBM Almaden Research Center.
[5] Osama Khan, Randal Burns," Rethinking Erasure Codes for Cloud File Systems: Minimizing I/O for Recovery and Degraded Reads", Cheng Huang Microsoft Research
[6] B. Calder, J. Wang, A. Ogus, N. Nilakantan, A. Skjolsvold, S.McKelvie, Y. Xu, S. Srivastav, J. Wu, H. Simitci, J. Haridas, C.Uddaraju, H. Khatri, A. Edwards, V. Bedekar, S. Mainali, R.Abbasi, A. Agarwal, M. F. u. Haq, M. I. u. Haq, D. Bhardwaj, S.Dayanand, A. Adusumilli, M. McNett, S. Sankaran, K.Manivannan, and L. Rigas, "Windows Azure storage: A highly available cloud storage service with strong consistency", inProc. 23rd ACMSymp. Oper. Syst. Principles, New York, NY,USA, 2011, pp. 143–157.
[7] Lakshmi J Mohan_, Renji Luke Haroldy, Pablo Ignacio Serrano Caneleoz,
UdayaParampallix and Aaron Harwood, "Benchmarking the performance of Hadoop triple replication and erasure coding on a nation-wide distributed cloud," 978-1-4799-1911-6/15/$31.00@2015 .IEEE.
[8] G. Xu, F. Wang, H. Zhang, and J. Li, "Redundant data composition of peers in p2p streaming systems using Cauchy Reed-Solomon codes," in Proc. 6th Int. Conf. Fuzzy Syst. Knowl. Discovery-Volume 2, Piscataway, NJ, USA, 2009, pp. 499–503.
[9] 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. 1–12.
[10] Guoxin Liu and HaiyingShen," Minimum-cost Cloud Storage Service Across Multiple Cloud Providers", Clemson, SC 29631, USA
[11] Guangyan Zhang, Guiyong Wu, Shupeng Wang, JiwuShu, WeiminZheng, and Keqin Li" CaCo: An Efficient Cauchy Coding Approach for Cloud Storage Systems". IEEE TRANSACTIONS ON COMPUTERS, VOL. 65, NO. 2, FEBRUARY 2016