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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

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

Nilesh V.Wanare, Kanchan Varpe

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

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

ABSTRACT: In large storage systems, it is crucial to protect data from loss due to failures.In 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. Cloud storage uses distributed model to avoid fault tolerance.It makes the use the coding schemes for storage purpose and makes data secure.This paper Focuses 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. Then 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.CaCo approach helps in identifying the optimal coding scheme within given redundancy configuration. Also it provides parallel processing to improve the computing devices performance.

KEYWORDS: Cloud Storage, Fault Tolerance, Reed-Solomon Codes, 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 describes how 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 $\binom{2w}{k+m}\binom{k+m}{k}$ Cauchy matrices, there is no method discovered to determine which one can produce the best schedule. Giving a Cauchy matrix, different schedules generated by various heuristics lead to a great disparity on coding performance. For a given redundancy configuration, it is with very low probability that one coding scheme chosen by rules of thumb performs the best.

The goal of this paper is to find a best-performing coding scheme that performs data coding with the fewest XOR operations. It resulted into reduced CPU overhead, identifying fault tolerance.The system proposed the N gram technique as extention to previous system which helps in searching the file and fast processing in cloud storage.It also avoids the data duplication.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

II.RELATED WORK

In [1], author describes that 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 and RAID 6.

In [2], CRS coding involves only XOR operations, the number of XORs required by a CRS code directly upon the performance of encoding or decoding. Cache behavior and device latency, reducing XORs is a reliable and effective way to improve the performance. This paper studies the longest-density MDS codes which are multi-erasure array code with optimal redundancy and minimum update penalty. We prove some basic structure properties for longest lowest-density MDS codes.

In [3],author describes that 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 paper 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.
In [4], author defines 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 paper provides methodology for scattered erasure to handle errors on disk. It defines that lost data is uncover able or it is declared uncover able and formula is provided for the reconstruction that depends only on readable sectors.

In [5], Cloud record frameworks are transitioning from replication to deletion codes to reduce the storage overhead. This paper present an algorithm that finds the optimal number of codeword symbols required for any XOR-based erasure code and produces recovery schedules that use a minimum amount of data. It also defines Reed-Solomon codes of new class that perform degraded reads more effective than all known codes but inherit the reliability and performance properties of Reed-Solomon codes.
.
In [6], author describes the WAS architecture, global namespace as well as its resource provisioning, load balancing and replication system. In these architecture customers have access to their data at any time and only pay for what they use and store. In WAS data is stored durably both local and geographic replication to facilitate disaster recovery. Currently, WAS storage comes in the form of files, Tables and Queues (message delivery).

In [7], Proposed system provides mechanisms for fault-tolerance over large distributed storage systems. Recently, erasure codes are being used for the replication, since they provide the same fault-tolerance over reduced storage overhead. However their performance is uncertain 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.

In [8], author 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 applythe 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 FEC reduces mean time of failures by many orders of size compared to duplication systems with similar storage and bandwidth requirements.

In [9],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. Thus, a problem within the field of erasure coding is how to formulate the XOR operations for given matrix so that the fewest number of XOR operations are required. The

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

system stated an algorithm for finding the optimum solution and study the performance of two known heuristics on encoding matrices set.

In [10], author 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 two challenges: 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 paper 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

III. GOALS AND OBJECTIVE

The goal of this system is to find a best-performing coding scheme that performs data coding with the fewest XOR operations And find a Cauchy matrix, whose schedule is desired to be the Shortest.This system resolve the parameters like data duplication and decrease the CPU overhead. System study finds that with very low probability, one coding scheme chosen by rules of thumb, for a given redundancy configuration, performs best. This system propose CaCo, an efficient Cauchy coding approach for data storage in the cloud. This system helps to transfer the data over cloud with minimum number of operations using cachy encoding technique.It manages data recovery,data duplication and results into CPU Optimization.

IV.PROPOSED SYSTEM ALGORITHM

Stemming Algorithm:
Information Retrieval (IR),Stemming is very useful tool which is supported by almost all indexing and search systems.The algorithm proposed by stemming is a mechanism that decrease words with the same stem to a common form, it removes the derivational and inflectional suffixes from each 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 words into common form .The development of stemming algorithms for freetext retrieval purpose is evidenced by the work of many researchers. A stemming algorithm reduces the words "fishing", "fished", and "fisher" to the root word, "fish"The main purpose of stemming is to decrease different syntactical forms / word forms of a word such as its noun, adjective, verb, adverb etc. to root form.
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

V. PROPOSED SYSTEM ARCHITECTURE

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.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

Advantages:-

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

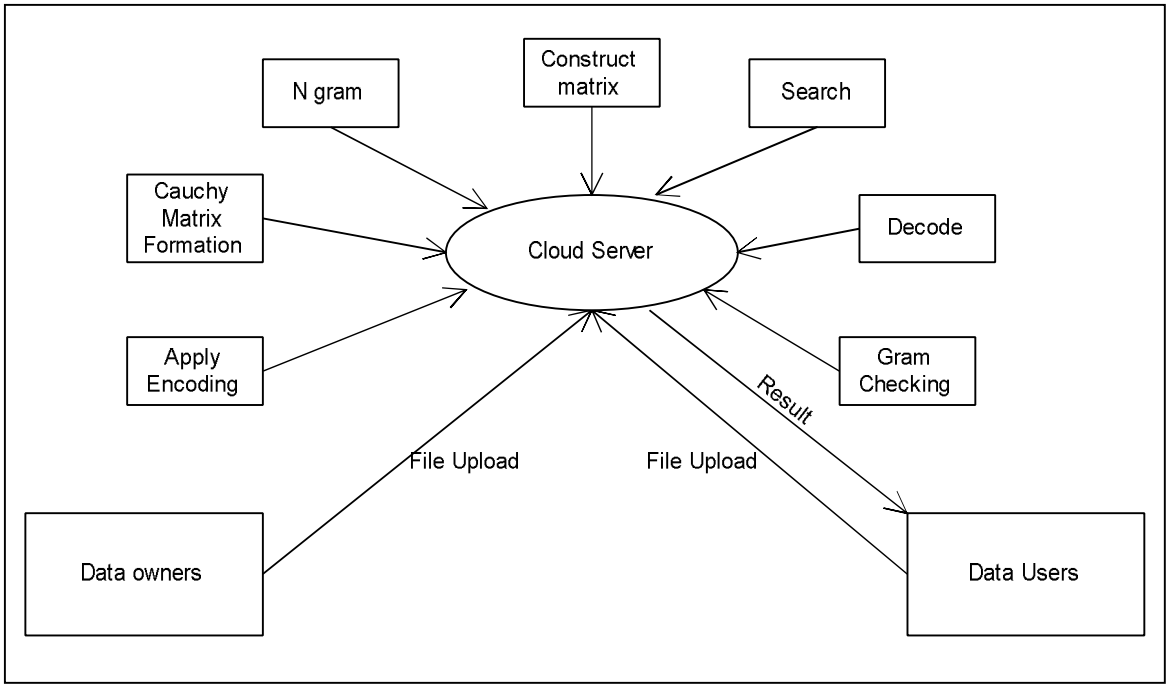

Fig. 1. System Architecture

The system architecture shows the flow of proposed system .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 credientials for authentication purpose.After Successsdul login user can upload file.Then file encoding and n-gram creation 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.

VI. EXPECTED RESULT

k, m, w	Matrix	schedule	XORs
4, 4, 3	Cauchy Good	Uber-XSet	31
4, 3, 4	Optimizing Cauchy	MW	36
7, 3, 4	Cauchy Good	MW-Matching	70
13, 3, 4	Cauchy Good	MW-SQ	137

Fig. 2. Combination of Cauchy Matrix and XOR Schedule for Different Redundancy Configurations

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

The Above Fig.Summarizes a list of the best combinations of a Cauchy matrix heuristic and an XOR schedule heuristic, selected by CaCo, for different redundancy configurations. Some conclusions that one may draw from the experimental results are as follows. Given a Cauchy matrix, changing the XOR schedule heuristic affects the coding performance significantly. Given a redundancy configuration of k , m ,w the locally optimal schedule with a different Cauchy matrix heuristic has an obviously different size.

VII.CONCLUSION

Distributed storage frameworks dependably utilize distinctive repetition setups (i.e., k; m; w), contingent upon the coveted equalization amongst execution and adaptation to internal failure. For various mixes of grids and calendars, there is an expansive crevice in the quantity of XOR operations, and no single mix performs best for all excess arrangements. In this paper, we propose CaCo, another methodology that fuses all current framework and timetable heuristics, also, consequently can recognize an ideal coding plan inside the ability of the present best in class for a given excess arrangement.

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] Jianqiang Luo, Lihao Xu, James S. Plank,"An Efficient XOR-Scheduling Algorithm for Erasure Codes Encoding", Dependable Syst. Netw., 2009, pp. 504–513.
[4] James Lee Hafner, Veera Deenadhayalan, 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 ACM Symp. Oper. Syst. Principles, New York, NY,USA, 2011, pp. 143–157.
[7] J. Edmonds, "Paths, trees, and flowers," Can. J. Math., vol. 17, no.3, pp. 449–467, 1965.
[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 Haiying Shen," Minimum-cost Cloud Storage Service Across Multiple Cloud Providers", Clemson, SC 29631, USA.