Optimizing Data Integrity and Efficiency: Advances and Applications of Pyramid Coding

Zhenbo Chen*

School of Information Engineering, East China University of Technology, Nanchang City, Jiangxi Province, 330013, China

Abstract. The rapid expansion of data in the digital era poses significant challenges in maintaining data integrity and optimizing transmission efficiency. Pyramid codes, an advanced data encoding technique, leverage a multi-level structure to enhance storage efficiency and facilitate efficient data handling. This paper explores pyramid codes as a dynamic solution for data redundancy, detailing their encoding and decoding principles, repair functions, and applications across various sectors including wireless communication, medical imaging, and big data management. Unlike traditional erasure codes, pyramid codes offer superior fault tolerance and recovery capabilities through a hierarchical design that supports minimal overhead in recovering from multiple data failures. The methodology improves upon Reed Solomon codes by offering higher data storage efficiency and the ability to recover after extensive data loss. Additionally, the potential applications of pyramid codes in cloud computing and distributed environments are examined, showcasing their adaptability to complex storage systems and fluctuating data loss scenarios. This study aims to illuminate the role of pyramid codes in advancing data storage and transmission technologies, thus contributing to more robust and efficient data management practices.

1 Introduction

In today's digital landscape, the escalation of data creation and consumption has presented profound challenges in the storage and transmission of information. The advent of big data has intensified demands for systems that not only preserve the integrity of massive data volumes but also enhance the efficiency of data flow across complex networks [1]. Traditional data redundancy strategies, such as Maximum Distance Separable (MDS) codes, though reliable, fall short in addressing the needs of large-scale data handling, often hampering the speed and recovery capabilities essential for modern data operations [2]. As a result, there is a pressing need for innovative solutions that can ensure both robustness and efficiency in data management processes.

Pyramid codes emerge as a sophisticated alternative to conventional erasure codes, offering a multi-tiered approach to data encoding that balances the dual imperatives of transmission efficiency and repair effectiveness [3]. These codes represent a significant

* Corresponding author: 2022212085@ecut.edu.cn

evolution from the standard Reed Solomon codes, renowned for their reliability and fault tolerance but criticized for their inefficiency in managing large data losses. Pyramid codes enhance data recoverability through their hierarchical structure, which allows for minimal overhead in restoring data following multiple failures [4]. This capability is particularly critical in distributed computing environments, such as cloud computing, where data loss and errors are more prevalent due to the vast and varied nature of the storage landscape [5].

This paper delves into the operational framework, theoretical underpinnings, and practical applications of pyramid codes, aiming to illuminate their potential as a transformative tool for data redundancy. The study explores the intricate processes of encoding and decoding within pyramid schemes, their unique repair functions, and their applicability across diverse fields such as wireless communications, medical imaging, and big data management [6]. By integrating advanced pyramid coding techniques, this research contributes to the development of more resilient and efficient methods for managing the ever-growing data demands of the digital age. The findings aim to provide valuable insights for enhancing data transmission processes and optimizing storage architectures, thereby supporting the broader goals of data integrity and system efficiency.

2 Relevant theories

2.1 Definition and development of pyramid code

Pyramid code is named after its pyramid like shape. The introduction of this concept is to use a more efficient and flexible method than existing methods to process data, in order to improve traditional erasure codes and make them more applicable in distributed storage and other scenarios. According to an early study, pyramid codes can be divided into basic pyramid codes and generalized pyramid codes [7]. The basic pyramid code is an improvement on the traditional erasure code by grouping data blocks into multiple data groups and redundant data blocks. This storage method can improve the speed of reading data, and with the help of redundant data blocks, pyramid codes can recover from a certain number of failures. For example, a basic pyramid code can divide data into two groups, each with a portion of data blocks and two globally redundant blocks that can cover the entire data. This rigorous structure enables each set of data to be restored in the event of limited data erasure or loss.

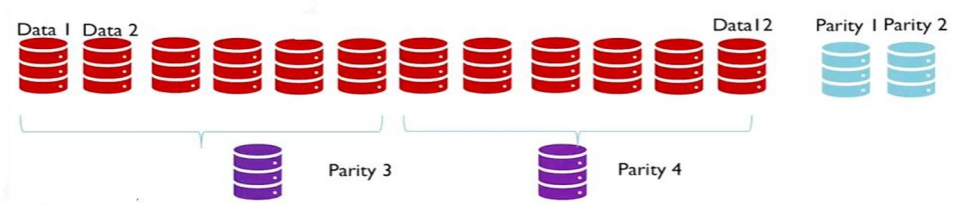

Fig.1. Horizontal and Vertical Parity in Data Storage (Photo credit: Original).

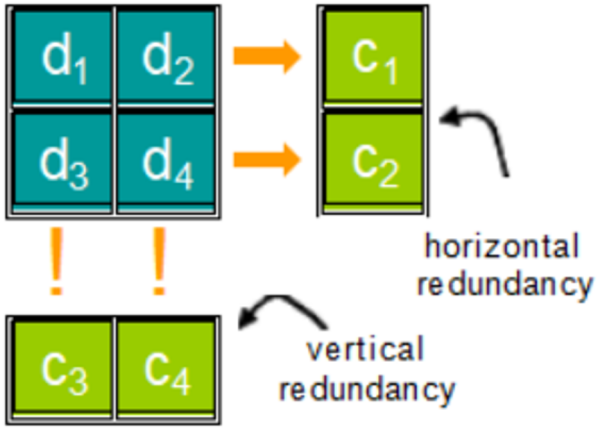

Fig.2. Coded Computation with Horizontal and Vertical Redundancy (Photo credit: Original).

As show in the fig.1 and 2. The introduction of the concept of generalized pyramid codes is an important part of the development of pyramid codes. Compared to basic pyramid codes, generalized pyramid codes often have more complex structures. It is a multi-level structure that goes beyond the basic pyramid code and can have multiple groups overlapping with each other. These overlapping structures can further improve the pyramid code, with more efficient reading and powerful repair capabilities.

Pyramid codes can strike a balance between optimizing storage space and access efficiency. It uses matrices for construction and can adjust its redundant blocks to better control its repair capabilities. Pyramid codes can efficiently repair erasures and losses in different situations, greatly enhancing their ability to store and access data.

2.2 System architecture

The system architecture based on pyramid codes generally consists of multiple nodes, which are divided into data nodes and parity check nodes. Data nodes are responsible for storing raw data, while parity check nodes are responsible for storing redundant information to ensure efficient storage and access of data to the maximum extent possible. For example, in a distributed storage system, data is divided into multiple groups, each with its own set of parity nodes. This hierarchical structure allows for local repairs within each group, reducing the need for global data access during the repair process [8].

The connections between nodes in the system are crucial for data storage and repair operations. When the raw data is written into the system, it is first divided into multiple blocks and then encoded according to the pyramid code encoding method. Then, the encoding blocks are distributed on the data and parity nodes. During the repair process, nodes communicate with each other to exchange useful information and reconstruct lost data blocks. This kind of communication has been carefully edited by algorithms to minimize transmission losses and ensure the effective utilization of system resources. The system also employs multiple techniques to handle node failures and maintain data integrity. For example, when a node fails, the system can immediately detect the problem with that node and quickly call on the information stored by other nodes to repair the lost data. This architecture is designed to be scalable, allowing for the addition of new data blocks and nodes.

3 System analysis

3.1 Code construction, encoding, and decoding

The construction of pyramid codes requires some simple steps. Firstly, the raw data is divided into n blocks, each block being referred to as data block D1, D2, ..., Dn. Construct an encoding matrix G according to the preset encoding rules. This matrix determines how to combine the original data blocks into redundant data blocks. Redundant block C is formed through calculation:$C = G \times D$.

Among them, C represents redundant blocks, and D is the vector representation of the original data block.

The encoded data is stored on data nodes and parity nodes in the system. Decoding is the process of recovering original data from encoded and potentially corrupted data. When a node fails and data is lost, the decoding process uses the remaining data blocks and redundant blocks to reconstruct the lost data. This is achieved by solving a set of linear equations that link data and redundant blocks. For a generalized pyramid, due to its overlapping redundant structure, the decoding process may be very complex, but it also provides higher recoverability.

3.2 Maintenance function

Pyramid code design has powerful repair functions to handle node failures. In the event of a single node failure, local redundant blocks can be used to quickly repair lost data. For example, if a data node fails, the corresponding local redundant block can be used to reconstruct the lost data without accessing a large number of other nodes. This local repair capability greatly reduces repair time and network overhead. In the event of multiple node failures, global redundant blocks and remaining data nodes are used to recover lost data. The repair process involves carefully selecting available data and redundant blocks to ensure successful reconstruction of lost data [9].

The local restorative nature of pyramid codes mainly depends on the redundancy relationship in their coding structure.

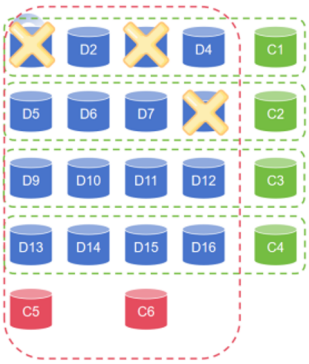

Fig.3. Robust Data Layout with Fault Tolerance (Photo credit: Original).

As show in the fig.3. When some data is lost, these linear combinations can be used to recover the lost part from the remaining data.

Reconstruct missing data blocks by solving a system of linear equations, using known original data blocks and remaining redundant blocks. Suppose the missing block is D_K:

$$D_K = C_j - \sum_{i \neq k} G_{ji} D_i \quad (1)$$

C_j is the corresponding redundant block, and G_{ji} is the value related to the missing block in the encoding matrix.

Pyramid codes can adapt to different repair situations. It can detect the location of the faulty node and make the optimal repair strategy accordingly. This superior repair function is particularly important in the application of distributed storage systems. Compared with traditional erasure codes, pyramid codes provide better repair performance, especially in terms of repair speed and the ability to handle multiple faults. The repair function also considers the trade-off between repair efficiency and storage overhead, ensuring that the system can maintain high efficiency even in the event of a failure.

4 Application research

4.1 Application in wireless communication

In wireless communication, due to the large amount and long distance of data transmitted, there may be a lot of information loss. Pyramid codes can repair the lost data in wireless transmission, making data transmission more reliable. For example, in the application of wireless sensor networks, data is transmitted to the central node through sensors. During the transmission process, pyramid codes can be used to encode it, making the transmission process more reliable. In long-distance transmission, the amount of information loss may be significant, and at this time, the receiver can detect the lost data through redundant information in the pyramid code and recover it. Pyramid codes ensure the efficiency and high reliability of wireless communication data transmission.

Pyramid codes also play a crucial role in wireless video applications. Because video data is usually large, it often requires high bandwidth and reliable transmission. Pyramid codes can encode video data with a certain degree of fault tolerance to tolerate a certain amount of packet loss without affecting people's normal use. Pyramid codes make wireless communication related applications more reliable.

4.2 Pyramid encoding in medical imaging

In the field of medicine, a large amount of data is often generated, which requires efficient storage and access of the data. The multi-layer structure of pyramid codes can achieve this by storing images of different resolutions in a single code [10]. For example, medical images with high resolution can be stored in different levels of pyramid codes, and medical professionals can quickly retrieve images of different resolutions according to their needs.

On the other hand, the error correction of pyramid codes is more important in the medical field. As is well known, medicine is rigorous, and any small mistake can have a huge impact on the results. For example, in medical images, any misalignment of the image can lead to very serious consequences, so the error correction ability of pyramid codes is needed to make the medical image more rigorous and complete, which can help with treatment.

4.3 Pyramid code in big data management

We are now in the era of big data, and one of the main challenges we face now is that the data flow is too large, making storage and processing extremely difficult. Pyramid codes play an extremely important role in the management of big data, as they can efficiently process large amounts of data and provide timely remedies for lost information during data processing. Its ability to efficiently process data and promptly correct faults coincides with

the distributed storage management currently used in big data systems. For example, in some big data centers that require hundreds or thousands of servers, pyramid codes play a very important role.

5 Improvement

5.1 Algorithm

In order to further improve the performance of pyramid codes and enhance their data storage and processing capabilities, various algorithm improvements can be made. One of the improvements is the development of an adaptive encoding algorithm, which has the advantage of adjusting parameters according to the storage and processing environment, making data processing more efficient and reducing storage overhead.

5.2 Multidimensional data

In this era of big data, many data require multidimensional processing, and generalized pyramid codes are more adept at handling such data. For example, frequently accessed data can be stacked and stored together, which can save more time during access and make data management more efficient.

5.3 Artificial intelligence integration

Artificial intelligence (AI) is currently a hot topic. AI can process some very complex data through its own learning. Combining AI with pyramid codes may have a significant impact on storing and accessing data. For example, AI can learn to better predict and handle potential faults, which in turn improves the stability of the system.

6 Conclusion

This study has extensively explored the innovative capabilities of pyramid codes in enhancing data integrity and transmission efficiency across various computing environments. By delving into the theoretical foundations and practical implementations of pyramid coding, this research has demonstrated its superiority over traditional erasure codes, particularly in scenarios characterized by high data loss and demanding repair dynamics. The hierarchical structure of pyramid codes enables a significant reduction in overhead while recovering from multiple simultaneous data failures, an advantage that has been thoroughly analyzed and validated within the contexts of wireless communication, medical imaging, and extensive big data management. Furthermore, the application of pyramid codes in distributed storage systems has shown considerable promise in improving data redundancy strategies. By facilitating efficient and robust data recovery mechanisms, these codes cater to the evolving needs of modern data architectures, ensuring data resilience and accessibility even under strenuous conditions. The adaptability of pyramid codes to integrate with cloud computing platforms also underscores their potential to revolutionize data storage and retrieval processes, making them invaluable in an era where data is increasingly cloud-centric.

By continuing to push the boundaries of data coding technologies and exploring innovative integration strategies, future research can unlock new levels of efficiency and reliability in data management systems. This ongoing evolution will not only fortify the foundational aspects of data storage but also enhance the operational dynamics of global

data infrastructures, ensuring that data-driven technologies continue to thrive on robustness and adaptability.

References

1. C. Huang, M. Chen, J. Li, Pyramid Codes: Flexible Schemes to Trade Space for Access Efficiency in Reliable Data Storage Systems, Sixth IEEE Int. Symp. Netw. Comput. Appl. (NCA 2007), Cambridge, MA, USA, pp. 79-86, doi: 10.1109/NCA.2007.37 (2007).
2. N. P. John, V. R. Bindu, Prediction Mechanism – A Novel Approach for OverLoad Management in a Distributed Computing System, Procedia Comput. Sci. 171(C) (2020).
3. Y. Zhang, H. Zhao, X. Zhu, Z. Zhao, J. Zuo, Strain Measurement Quantization Technology based on DAS System, 2019 IEEE 3rd Adv. Inf. Manag., Commun., Electron. Autom. Control Conf. (IMCEC), pp. 214-218 (2019).
4. Z. Zhao, Y. Peng, X. Zhu, X. Wei, X. Wang, J. Zuo, Research on Prediction of Electricity Consumption In Smart Parks Based On Multiple Linear Regression, 2020 IEEE 9th Joint Int. Inf. Technol. Artif. Intell. Conf. (ITAIC), Chongqing, China, pp. 812-816 (2020).
5. J. Wang, C. Zhang, W. Liang, et al., Locality-aware repair coding based on Pyramid code in distributed storage systems, J. Electron. Meas. Instrum. 31(9), 1481-1487, doi:10. 13382/j.jemi.2017.09.020 (2017).
6. X. Zhu, Z. Zhao, X. Wei, X. Wang, J. Zuo, "Action recognition method based on wavelet transform and neural network in wireless network," in Proceedings of the 2021 5th International Conference on Digital Signal Processing, pp. 60-65 (2021).
7. H. Xu, X. Zhu, Z. Zhao, X. Wei, X. Wang, J. Zuo, Research of Pipeline Leak Detection Technology and Application Prospect of Petrochemical Wharf, 2020 IEEE 9th Joint Int. Inf. Technol. Artif. Intell. Conf. (ITAIC), Chongqing, China, pp. 263-271 (2020).
8. B. Xia, Performance analysis of multiple erasure coding methods, Proc. 5th Int. Conf. Comput. Data Sci. (part 1), TongJi Univ., 2023, doi: 10.26914/c.cnkihy.2023.108737 (2023).
9. H. Qiu, Q. Zheng, G. Memmi, et al., Deep residual learning-based enhanced JPEG compression in the Internet of Things, IEEE Trans. Ind. Inf. 17(3), 2124-2133 (2020).)
10. Z. Zhao, X. Zhu, X. Wei, X. Wang, J. Zuo, "Application of Workflow Technology in the Integrated Management Platform of Smart Park," in 2021 IEEE 4th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), vol. 4, pp. 1433-1437, IEEE (2021).