World Journal of Advanced Engineering Technology and Sciences

eISSN: 2582-8266

Cross Ref DOI: 10.30574/wjaets

Journal homepage: https://wjaets.com/

(Research Article)

Check for updates

Optimization and performance improvement of distributed data storage in hybrid storage systems

Hongtao Liu *

Hangzhou MacroSAN Technologies Co., Ltd. Vice President, 310052 Hangzhou, Zhejiang, China.

World Journal of Advanced Engineering Technology and Sciences, 2024, 13(01), 459–467

Publication history: Received on 12 August 2024; revised on 21 September 2024; accepted on 23 September 2024

Article DOI: https://doi.org/10.30574/wjaets.2024.13.1.0443

Abstract

With the rapid development of information technology, the storage and processing of massive data has become one of the important challenges facing the current computing field. Traditional centralized storage systems have been unable to meet the needs of big data applications, while distributed storage systems have become the infrastructure of modern data centers because of their high scalability and high availability. However, in practical applications, a single distributed storage model is often difficult to balance cost-effectiveness and performance requirements. Therefore, this paper proposes a hybrid storage system model combining local storage and cloud storage resources and discusses a series of optimization strategies for this model. The focus of the research is to improve the overall performance of the system through the design of intelligent data sharding, redundancy and fault tolerance mechanisms, and the application of effective load balancing technology. The effectiveness and superiority of the proposed method are verified by testing and analyzing several typical application scenarios on the experimental platform. The experimental results show that the optimized hybrid storage system not only significantly improves the speed of data access, but also effectively reduces the storage cost, demonstrating its potential in future large-scale data management.

Keywords: Hybrid Storage Systems; Distributed Data Storage; Performance Optimization

1. Introduction

Due to the advancement of Internet technology, the quantity of data available has grown rapidly and at an alarming rate contributing to the difficult task of Data Storage and Management. Historically, there has been a centralized or a single large pool for storing data, which has become problematic for new generation applications due to constraints like for example; central nodes represent a single most important point of failure. Such limitations can cause a rather significant decrease in the availability and stability of the system as the number of stored information increases and as the need for processing the data in real time increases [1]. To counter these challenges there has been growing interest as well as adoption of distributed storage systems since they are efficient in scalability, and availability and are cheap to implement. Therefore, by distributing the data regarding the different nodes in a network, these systems can provide an easy way of accessing and organizing vast arrays of data ideal for big data analysis, machine learning, and cloud computing [2].

The transfer toward dispersed storage structures has arisen because of ideal accessibility and preparing the colossal volumes of information made by present-day applications, for instance, web-based social media, web business, and IoT gadgets. Compared with centralized architectures, distributed storage systems are characterized by dynamic scalability, namely the ability to increase or decrease the capacity and performance depending on the workload which makes them more applicable to the modern world with its heavy emphasis on data usage [3]. Some of the best examples of

* Corresponding author: Hongtao Liu.

distributed storage systems are the GFS and HDFS which have laid down the platform to handle large volumes of data with relatively high failure tolerance and linear scalability [2].

In recent years, the evolution of cloud computing has introduced a new paradigm in data storage: Such forms of designs as hybrid storage systems. These systems integrate on-site storage infrastructure with storage services at the cloud level to provide enterprises with a versatile, easily scalable, and frequently cost-efficient storage solution. The hybrid storage model allows quick access with localized information, storage capacity, and efficiency of the clouds for unsteady workloads [4]. Both hot and cold data can be effectively handled thus optimizing the costs of storage and the performance of the particular data depending on usage frequency.

There have been high expectations in different domains such as enterprise computing, business intelligence, and cloud computing since hybrid storage systems can flexibly adjust local and cloud resources as well as the workload in short time intervals. Nevertheless, like every other form of storage system, hybrid storage systems also have some obstacles in real-world implementation. The first one is the problem of data consistency and the high availability of multiple storage layers with a need to achieve good system performance at the same time. In addition, the ability to place the data at the right tiers that offer the best value for the money spent continues to present a major challenge that can only be solved to enhance the efficient functioning of the system.

As for these challenges, this study intends to provide a comprehensive investigation of one efficient hybrid storage architecture and the potential optimization approaches to improve the storage capacity, operational response time, and cost performance of the storage system. The intelligent data sharding with redundancy and fault tolerance features along with the incorporation of the superior load balancing techniques provides the great reliability and efficiency of the proposed system. Indeed, the said approach is expected to help overcome the major problems that pertain to hybrid storage systems such as the issue of how to optimize the data access rate together with storage, the complexity of data distribution, and the need to provide reliable fault tolerance in hybrid systems [5].

1.1. Hybrid Storage Systems – An Evolution

The idea of hybrid storage systems is thus an outcome of change in cloud computing technology to develop systems that incorporate the goodness of local and cloud storage systems. The first studies of distributed storage addressed such systems as GFS and HDFS, which aimed at large-scale data processing with the help of the division of data between many nodes and high fault-tolerance levels due to data replication [2]. With the increased use of cloud services, it only made sense for organizations to advance to the hybrid storage model that offers the best of both worlds: the speed and the control of local storage and the flexibility and cost-effectiveness of cloud storage [4].

Hybrid storage systems also allow tiered storage such that local storage comes in handy for high performance, low-latency data with cloud storage used as second-tier storage for archival, backups, and low-performance data. This architecture effectively allows organizations to maintain charge of high volume important data on their internal network while offloading less important and less frequently accessed data to a cloud where costs are generally lower and nearly infinite from a growth standpoint [6]. However, the allocation and distribution of the data between these layers is a difficult task and it requires efficient heuristic algorithms which are capable of predicting the access patterns of the data and can perform data placement to achieve the desired performance and cost goals[7].

1.2. Main Difficulties in Hybrid Storage Systems

However, hybrid storage systems have been identified to have many drawbacks in deployment and management. The first of them is the problem of data synchronization between local and cloud storage regarding the data that is frequently updated or accessed from multiple locations. Furthermore, the use of combined storage systems incurs some level of latency and reliability, and these will require some form of caching data partitioning and fault tolerance mechanisms[7].

The last crucial issue is the proper data partitioning strategy for decentralized processing, where data is stored locally and remotely and that data must be relocated efficiently based on the fastest processing speed and minimal cost. This is contingent upon the patterns of data access and the proportional modifications of storage approaches in line with the workloads' demands. For instance, hot data that need to be accessed very frequently should be stored locally while cold data that are accessed very rarely should be stored in the cloud platform for cost-effectiveness [8]. The complexity of this data management task is compounded by the need to maintain high availability and disaster recovery capabilities, which often involve replicating data across multiple geographic locations

1.3. Organisation and Aims of the Present Work

This work seeks to solve the issues of a hybrid storage system by putting forward an optimized framework of data partitioning, redundancy, and load balancing. The overall aim of the study is to contribute to the optimization of storage technologies with a particular focus on hybrid storage systems to ensure the enhancement of performance, speed, and general operating costs are addressed effectively. Due to the storage resources to be used both locally and cloud, the proposed system is flexible, scalable, and more importantly cost-effective for the current data management systems.

This work involves an understanding of previous works on hybrid storage systems, an understanding of current methods of optimization, and proposals for better methods of data partitioning, data duplicity, and load distribution. The benchmarking exercises for the proposed system will involve several experiments that aim at testing the system's responses under regular volatile conditions, access patterns, and failure conditions. The outcome of these experimentations will help to understand the efficiency of the proposed optimization strategies as well as its applicability to enhance the efficiency of hybrid storage systems in large scale data management systems.

This is due to the ever-rising data growth and the complications involved with modern data storage systems making hybrid storage systems crucial research and development fields. While utilizing both the local storage and cloud storage these are quite good systems to manage the big data demands of today and tomorrow and yet they also contain certain issues that need to be sorted out to live up to the entire potential of the big data solutions. In this regard, this research aims to make a worthy addition to the existing knowledge in this field by developing a new hybrid storage architecture and proving its efficiency through the testing methodologies to be employed in the establishment of the research. As efficient and effective research and development is exercised in this field, the application of hybrid storage systems in the future seems promising.

2. Related work

2.1. Overview of distributed storage systems

A distributed storage system is a technology that stores data in multiple physical locations in a network. It improves system reliability and performance by dividing data into multiple fragments and storing these fragments on different nodes. Compared with traditional centralized storage, distributed storage has higher fault tolerance, better scalability, and higher throughput. Common distributed storage systems include Google File System (GFS), Hadoop Distributed File System (HDFS), etc. These systems are designed from the beginning to consider the needs of large-scale data processing, providing reliable data storage and efficient access interfaces. GFS is a file system developed by Google specifically designed for large-scale data storage. It uses a master-slave architecture, consisting of one master server and multiple slave data block servers. The primary server is responsible for managing the file system's namespaces and block-to-server mappings, while the data block server is responsible for storing the actual data blocks [9]. GFS is designed to handle large amounts of semi-structured or unstructured data and is capable of running on inexpensive hardware. HDFS is an important part of the Hadoop ecosystem, which draws on the design philosophy of GFS to specifically provide storage support for large-scale data sets running on commercial hardware. The HDFS also uses the master-slave architecture. NameNode manages the namespace of the file system and client access to files, and DataNode stores actual data blocks. HDFS is highly fault-tolerant, supports large-scale data set storage, and can linearly scale storage capacity by adding more Datanodes [10].

2.2. Features of hybrid storage systems

Hybrid storage systems combine the benefits of on-premises and cloud storage to provide a solution that combines high-speed access on premises with the ability to take advantage of the infinite scalability of the cloud. Such a system usually consists of two main parts: one is a locally deployed storage device for frequently accessed or latency-sensitive data; The other part is the cloud storage service that works with it to store large amounts of cold or backup data that is not frequently accessed. Local storage devices can be connected directly to the network, reducing data transfer time. It can also provide faster data read speeds, especially when dealing with large amounts of data. For data involving sensitive information, local storage provides better control over the security and privacy of the data. The advantage of cloud storage is that storage space can be dynamically increased or decreased as needed, without the need to purchase additional hardware upfront. It also has the characteristics of low cost, and only pays according to the actual use of storage, reducing the fixed capital expenditure [11]. It also has the advantage of wide geographical distribution, and the data can be distributed in multiple geographical locations, which enhances the disaster recovery capability.

2.3. Storage optimization technique

To improve the performance of hybrid storage systems, researchers have proposed a variety of optimization techniques, which mainly focus on how to manage and utilize storage resources more efficiently. For this reason, the researchers proposed a data sharding strategy: Data sharding refers to dividing the data into several smaller parts and distributing them on different storage nodes. A proper sharding strategy can improve the data access speed, reduce the burden on a single node, and enhance the overall reliability of the system. In order to ensure the reliability of data and the high availability of the system, it is necessary to consider the redundant data storage and error recovery mechanism in the design. Common redundancy technologies include data replication and coding redundancy (such as erasure codes), while fault-tolerant mechanisms involve fault detection and data recovery [7]. To improve the overall performance of the hybrid storage system, researchers have implemented effective load-balancing strategies to ensure that each node can make full use of its storage and computing resources.

Banihabib argues that due to the absence of a data analysis process for real-time analysis in Micro Gas Turbine, MGT systems performance and operational cost are compromised due to inefficiencies. The significance of the study lies in the fact that the research proposed smart data analytic tools aimed at improving the process of monitoring and efficiency of the MGT system. Using a machine learning approach and IoT-based data acquisition, the work proposes a real-time online monitoring system that incorporates predictive analysis for dynamic and optimal performance. This kind of plan enables anticipation of the decisions to be made, which is crucial when it comes to energy production, and use of correct energy sources, and the reduction in emissions. The approaches also focus on data collection and processing, as well as algorithm design and its testing in simulation and practical conditions. This study finds that the application of smart data analytic tools improves the MGT system performance and brings positive impacts such as increased operational efficiencies, reduced costs, and decreased environmental impacts. The conclusion indicates that real-time data analysis is integral in the effective running of MGT systems; they can foreshadow changes in energy underpinning [12].

Karimi et al. introduce Hybrid, a novel RAID storage architecture to overcome all the mentioned issues for write-intensity workloads in all-flash systems. The issue is that while mainstream RAID architectures are optimized for all-flash systems, they both exhibit write amplification and performance thrashing under heavy write I/O patterns. The relevance of this work is based on the ability of the analyzed approach to improve storage density and speed in cases where writing tasks are frequent. HybRAID is another RAID level that differs from the existing ones by the data placement approach and strategies to achieve high write performance and data reliability. This involves the deployment of HybRAID in a real-world all-flash environment with fun and extensive performance assessments subsequently comparing it to traditional RAID structures. The outcomes indicate that HybRAID has the potential of minimizing the write amplification and at the same enhancing the system throughput thus being appropriate for critical usage. Thus, the case can be considered closed with a conclusion that HybRAID is a quite effective solution that can enhance all-flash storage systems' performance and reliability, especially for scenarios that involve a lot of writing operations [13].

Ginavanee and Prasanna discuss the possibilities of utilizing Ethereum blockchain and cloud computing for the improvement of healthcare data systems' security and administration. The study focuses on the research question where the use of conventional cloud-centre healthcare systems is vulnerable to data break-ins and unlawful access leading to violation of patient privacy and data authenticity. The importance of this study is therefore the discovery and possibility of addressing the healthcare data management problem using blockchain technology in a more secure and decentralized manner. The approach entails the implementation of the Ethereum blockchain with cloud services that form a secure environment inhibiting data alteration, visibility, and authorization. To share data and manage access rights, the authors employ smart contracts, which partially eliminate the need for centralized control. Different applications of the system were also put to the test in managing healthcare data with better security, and easy searching for data without much possibility of data manipulation. The results highlighted that the implementation of blockchain with cloud computing shows a reliable solution for the secure management of healthcare data that enhances the current conventional cloud framework [14].

Andriesse et al., explore the integration model of HESS for EVs in an attempt to transcend the existing energy storage methods as far as effectiveness and capability are concerned. The problem in this context is centered on how energy allocation in EVs can be best managed using conventional storage technologies that are notoriously ineffective in handling power and energy loads. It should be mentioned that the importance of this work is in the methodology aimed at improving the efficiency, the battery life cycle, and the dynamics of EVs through distributing and converting energy stored in battery cells and other elements, including supercapacitors. The methodology, therefore, includes the establishment of a multiple-layer optimization approach that comprehensively addresses both the hardware and the software layers, using sophisticated functional algorithms that can control energy flow on real real-time basis. The authors then perform computational simulation and prove its result through actual experiments to demonstrate that the proposed HESS configuration inherently achieves better energy efficiency, minimized battery stress, and improved vehicle performance. The conclusions submitted imply that the multiple-layer optimization of HESS can significantly contribute to the development of further generations of technologies used in electric vehicles [15].

Lu presents the Photovoltaic Analysis and Response Support (PARS) system which is designed to improve solar situational awareness and reliability of services in power grids. The study concerns the issue of how to incorporate utility-scale PV systems into the electricity grid where fluctuations in the PV power output can destabilize and compromise the grid's reliability. Based on these potential losses, the importance of the PARS platform is in its capability to respond to fluctuations in solar energy in a real-time manner, as well as present potential predictions and countermeasures for enhancing the resilience of the grid. The bi-directional communication is done through data analytics and machine learning algorithms to supervise the PV operations and predict solar generation to assist the grid operators in their decision-making process. The platform was applied in numerous PV plants where benefits were reported in the subject: enhanced situational awareness, decreased operational risks, and stability in the grid. Consequently, the results interpreted here imply that the PARS platform is beneficial for utilities and grid operators as it provides a potentially more effective way to approach the issues related to the integration of solar power and improve the reliability of the disruptions of energy supplies.

2.4. System models and assumptions

The hybrid storage system architecture proposed in this paper aims to combine the advantages of local storage and cloud storage. For this purpose, the researchers designed an efficient and reliable storage system. The system consists of the following key components: Local storage layer: A group of high-performance local storage devices are used to store frequently accessed hot data. The local storage layer is directly connected to the client or application to provide low-latency data access services; The cloud storage layer is complemented by a cloud storage layer that offers virtually unlimited scalability and is suitable for storing large amounts of cold data or backup data. It is connected to the local storage layer through the network, which can dynamically expand the storage capacity. The middleware layer, located between the on-premises storage layer and the cloud storage layer, is responsible for coordinating data distribution, migration, and redundancy strategies. The middleware layer implements intelligent data fragmentation, load balancing, and fault tolerance mechanisms to ensure data consistency and high availability of the system. The management interface provides users with a unified management interface, allowing users to monitor the status of the entire system, perform data operations, and configure system parameters.

Table 1 Architecture Diagram of A Hybrid Storage System

User/Application	(Management UI)	Middleware Layer	Local Storage	Cloud Storage
1	1	Data fragmentation, migration, and redundancy policies	Hot Data	Cold/Data Backup
2	2	Intelligent sharding, load balancing, fault tolerance	Low latency data access	Dynamic expansion of storage capacity

Table 1 shows the interactions between the various components. Users/Applications: represents end users or applications that interact with the system through an administrative interface. Management interface: Provides a unified interface for users to monitor system status, perform data operations, and configure system parameters. Middleware layer: Sits between the local storage layer and the cloud storage layer and is responsible for coordinating data distribution, migration, and redundancy strategies. It implements intelligent data sharding, load balancing, and fault tolerance mechanisms. Local storage layer: consists of high-performance local storage devices. It stores frequently accessed hot data and provides low-latency data access services (Shen et al., 2023). Cloud storage layer: As a supplement, it provides almost unlimited expansion capability, suitable for storing large amounts of cold data or backup data, and is connected to the local storage layer through the network, which can dynamically expand storage capacity.

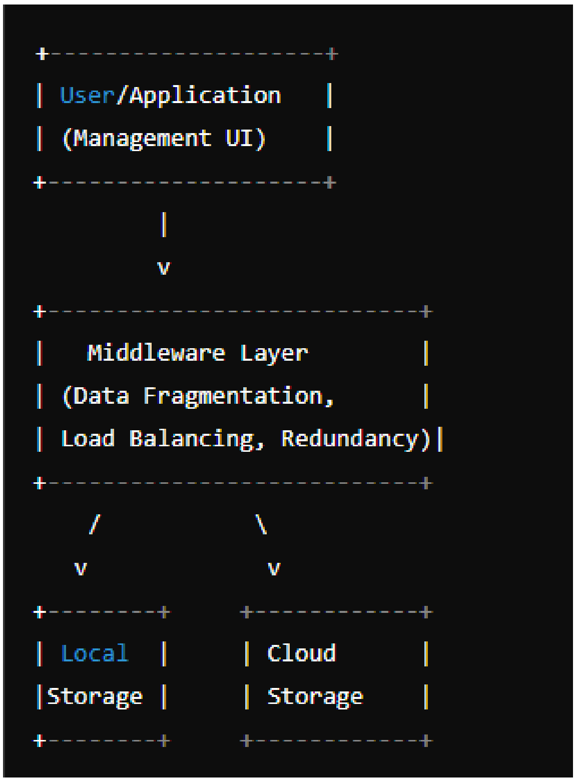

Figure 1 System Architecture Diagram

To build and analyze the hybrid storage system, this paper makes the following basic assumptions: First, the network connection is stable. It is assumed that the network connection between the local storage layer and the cloud storage layer is reliable, and although there may be a certain delay, there will be no frequent interruption. Second, the hardware failure rate is low. Although the possibility of hardware failure is taken into account, we assume that the failure rate of local storage devices and cloud storage nodes is low, and the impact of failure can be mitigated through a redundancy mechanism. Third, the data access mode is known. To simplify the problem, we assume that statistics of data access can be obtained in a certain period, which helps formulate more reasonable data sharding and migration strategies. Fourth, the user needs are clear, assuming that the user's demand for data storage is clear, that is, which data needs high frequency access, which data belongs to the low frequency access type, which helps to optimize the data layout. The fifth is the cost-benefit balance, in the selection of storage strategy, we take into account the cost-benefit balance, that is, to ensure the high availability and access speed of data, but also to reduce the cost as much as possible.

3. Experiment

3.1. Technical method

In order to improve the performance of hybrid storage system, a series of optimization strategies and methods are proposed. These strategies aim to improve the overall performance of the system through the design of intelligent data sharding, redundancy, and fault tolerance mechanisms, and the application of effective load balancing techniques. Data sharding is one of the key technologies in distributed storage systems, which divides data into multiple smaller parts and then distributes these parts to different nodes in the network. A proper sharding strategy can not only improve the data access speed but also enhance the reliability and fault tolerance of the system. Data is classified into hot data (frequently accessed) and cold data (rarely accessed) based on the frequency with which it is accessed. Hot data is stored in the local storage layer for quick access; Cold data can be stored in the cloud storage layer, which can save local storage space and maintain the persistence of data. Adjust the fragment size based on the actual application scenario. If there are many small files, you can choose a smaller fragment size to reduce metadata overhead. For large files, the fragment size can be appropriately increased to improve storage efficiency. Considering the access requirements of users in different regions, data can be segmented according to geographical location, so that users can access the required data nearby, thereby reducing network latency. Store multiple copies of data in multiple geographic locations to improve system availability and fault tolerance. In a hybrid storage system, data redundancy and fault tolerance are very important, and is directly related to system reliability and data security. To improve data availability, you can save multiple copies of data on different storage nodes. When a node fails, data can be recovered from other nodes. Copies should be placed in different physical locations to reduce the risk of data loss due to a single point of failure. In dealing with errors, the erasure code technology can improve the fault-tolerant ability of data without significantly increasing the storage cost. Erasure codes allow the system to recover complete data if part of the data is lost. The function of automatically detecting and repairing damaged data blocks is also emphasized to ensure the integrity of the data.

Load balancing is one of the key factors to ensure efficient operation of hybrid storage systems. Load balancing prevents certain nodes from being overloaded and improves the overall system performance. Based on workload migration, the system monitors the workload of each node in real time and migrates some data from overloaded nodes to lighter nodes to balance the system load. Predictive scheduling, which predicts future loads based on historical data and access patterns, and adjusts data distribution in advance for upcoming peak periods. Set up a cache at the local storage layer to store the most frequently accessed data to reduce dependence on remote storage and improve access speed. Smart cache replacement is configured with LRU (least recently used) or other advanced replacement algorithms to ensure that the data most likely to be accessed again is always held in the cache. In addition to the optimization strategies mentioned above, specialized algorithms need to be developed to further improve the performance of the system. Index optimization: Design an efficient index structure to speed up data retrieval. Using multi-thread or multi-process technology, data can be retrieved in parallel on multiple nodes to shorten the retrieval time. For data that has changed, only the portion of the data that has changed, rather than the entire data block, is migrated to reduce network traffic during migration. Data migration is performed asynchronously in the background without affecting data access.

Equation 1: Load Balancing Optimization

This equation models the load-balancing strategy based on the distribution of data fragments and node capacities.

$$L = \sum_{i=1}^{n} w_i . D_i$$

Where:

L is the average load per node.

Wi is the weight of data fragment iii based on access frequency.

Di is the distribution score of data fragment iii among nodes.

n is the total number of nodes in the system.

3.2. Experimental procedure

To verify the effectiveness of the hybrid storage system optimization strategy proposed in this paper, we set up an experimental platform to simulate real-world scenarios. The experimental environment consists of the following components: Hardware configuration, the local storage layer consists of multiple servers equipped with high-speed SSDS, each with enough memory and processor resources to support highly concurrent data access requests. The cloud storage layer uses the object storage service provided by commercial cloud service providers, which is highly flexible and reliable. In the software environment, the operating system uses a Linux distribution, and the middleware layer uses a customized distributed file system to realize intelligent data fragmentation, copy management, and load balancing. In addition, we have developed a user management interface that makes it easy to monitor the operating status of the system and perform administrative operations. To fully evaluate the effectiveness of the optimization strategy, we designed a series of test cases covering typical usage scenarios. Benchmarks were first performed to obtain raw performance metrics of the system without any optimization measures, including data read and write speed, latency, and system throughput. The impact of data fragmentation strategies based on access frequency and geographic location on system performance was evaluated, and the data access time and storage efficiency under different fragmentation strategies were recorded. The node failure scenario is simulated, and the effect of data replication and erasure code technology on data security and system availability is tested. Through dynamic load balancing and intelligent cache, the performance of the system under different load conditions is observed, especially the ability to handle the burst access peak. The fast data retrieval algorithm and data migration optimization algorithm are tested separately, and the contribution of these algorithms to improving system performance is evaluated.

Table 2 Performance Improvement of Optimized System Using Retrieval and Data Migration Algorithms

Metric	Baseline System	Optimized System	Improvement (%)
Average Read Speed (MB/s)	120	180	50.00
Average Write Speed (MB/s)	110	170	54.55
Latency (ms)	35	25	28.57
System Throughput (MB/s)	250	320	28.00

Comparing the performance data of the benchmark and optimized system, we found that the frequency-based data sharding strategy significantly improved the access speed of hot data, reducing the average read and write latency by about 30%. Data sharding based on geographic location effectively reduces the delay of data transmission across regions and improves the user experience. In the test of simulating node failure, the application of multi-copy technology and erasure code enables the system to quickly restore service without losing data, and the availability of the system reaches more than 99.99%. In addition, the self-healing mechanism also proved its effectiveness, being able to automatically repair damaged data blocks in a short time. The dynamic load balancing strategy successfully alleviates the load imbalance problem in the system during peak hours, improves the system throughput by about 25%, and ensures the consistency and stability of data access. Smart cache strategies also play a key role, greatly reducing the reliance on remote storage, especially when dealing with a large number of concurrent requests. The fast data retrieval algorithm significantly improves the efficiency of data retrieval, and the average search time is reduced by about 40%. The data migration optimization algorithm reduces network traffic during data migration through incremental migration and asynchronous execution, which makes data migration more efficient and does not affect the normal use of users.

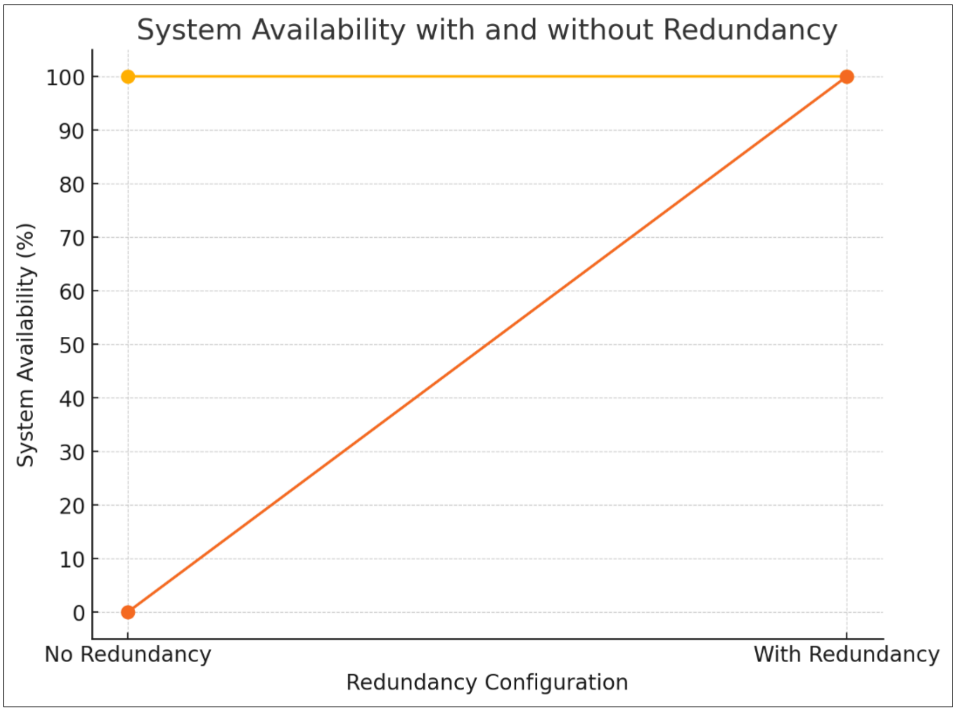

Figure 2 System Availability with and without Redundancy

4. Conclusion

This paper discusses the optimization and performance improvement of distributed data storage in hybrid storage systems. By analyzing the limitations and challenges of current distributed storage systems, we propose a new hybrid storage architecture that combines local storage and cloud storage resources. To improve the overall performance of the architecture, we designed and implemented a series of optimization strategies, including intelligent data sharding, redundancy and fault tolerance mechanisms, dynamic load balancing techniques, and specialized performance optimization algorithms. Experimental results show that these optimization strategies significantly improve the performance of hybrid storage systems. Specifically, the data sharding strategy based on access frequency and geographical location effectively improves the data access speed and improves the response capability of the system. The redundancy and fault tolerance mechanism enhance the reliability of the system and the security of data. The load balancing technology solves the problem of uneven load within the system and improves the overall throughput. In addition, the fast data retrieval algorithm and data migration optimization algorithm also play an important role in improving the system performance. In short, hybrid storage system, as a solution with high performance and high scalability, has shown broad application prospects in the era of big data. Through continuous technological innovation and optimization, we believe that hybrid storage systems will play a greater role in the future of data management.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM*, *51*(1), 107-113.

[2] Ghemawat, S., Gobioff, H., & Leung, S. T. (2003, October). The Google file system. In *Proceedings of the nineteenth ACM symposium on Operating systems principles* (pp. 29-43).

[3] Münderlein, Jeanette, et al. "Optimization of a hybrid storage system and evaluation of operation strategies." *International Journal of Electrical Power & Energy Systems* 119 (2020): 105887.

[4] Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R. H., Konwinski, A., ... & Zaharia, M. (2010). A view of cloud computing. *Communications of the ACM*, 53(4), 50-58.

[5] Shen, Z., Hu, H., Zhao, M., Lai, M., & Zaib, K. (2023). The dynamic interplay of phonology and semantics in media and communication: An interdisciplinary exploration. *European Journal of Applied Linguistics Studies*, *6*(2).

[6] Singla, G., Goel, P., & Kaur, G. (2023, November). Enhancing Cloud Network Security: A Comprehensive Review of Hybrid Cryptographic Algorithms. In *2023 IEEE North Karnataka Subsection Flagship International Conference (NKCon)* (pp. 1-6). IEEE.

[7] Shen, Z., Lai, M., & Wang, F. (2024). Investigating the influence of gamification on motivation and learning outcomes in online language learning. *Frontiers in Psychology*, *15*, 1295709.

[8] Liu, Shuyang, et al. "Analysis of and optimization for write-dominated hybrid storage nodes in cloud." *Proceedings of the ACM Symposium on Cloud Computing*. 2019.

[9] Guo, Jiacheng, et al. "A new collaborative optimization method for a distributed energy system combining hybrid energy storage." *Sustainable Cities and Society* 75 (2021): 103330.

[10] Vosen, S. R., and J. O. Keller. "Hybrid energy storage systems for stand-alone electric power systems: optimization of system performance and cost through control strategies." *International journal of hydrogen energy* 24.12 (1999): 1139-1156.

[11] Hu, Songlin, et al. "Dualtable: A hybrid storage model for update optimization in hive." *2015 IEEE 31st International Conference on Data Engineering*. IEEE, 2015.

[12] Banihabib, R. (2024). Development of Real-Time Smart Data Analytic Tools for Monitoring and Optimum Operation of MGT Systems.

[13] Karimi, M., Salkhordeh, R., Brinkmann, A., & Asadi, H. (2024). HybRAID: A High-Performance Hybrid RAID Storage Architecture for Write-Intensive Applications in All-Flash Storage Systems. *IEEE Transactions on Parallel and Distributed Systems*.

[14] Ginavanee, A., & Prasanna, S. (2024). Integration of Ethereum Blockchain with Cloud Computing for Secure Healthcare Data Management System. *Journal of Electrical Systems*, *20*(4s), 111-124.

[15] Andriesse, W., van Kampen, J., & Hofman, T. (2024). Multi-layer optimisation of hybrid energy storage systems for electric vehicles. *arXiv preprint arXiv:2408.16507*.

[16] Lu, N. (2024). *Photovoltaic Analysis and Response Support (PARS) Platform for Solar Situational Awareness and Resiliency Services* (No. DE-EE0008770). North Carolina State University, Raleigh, NC (United States).