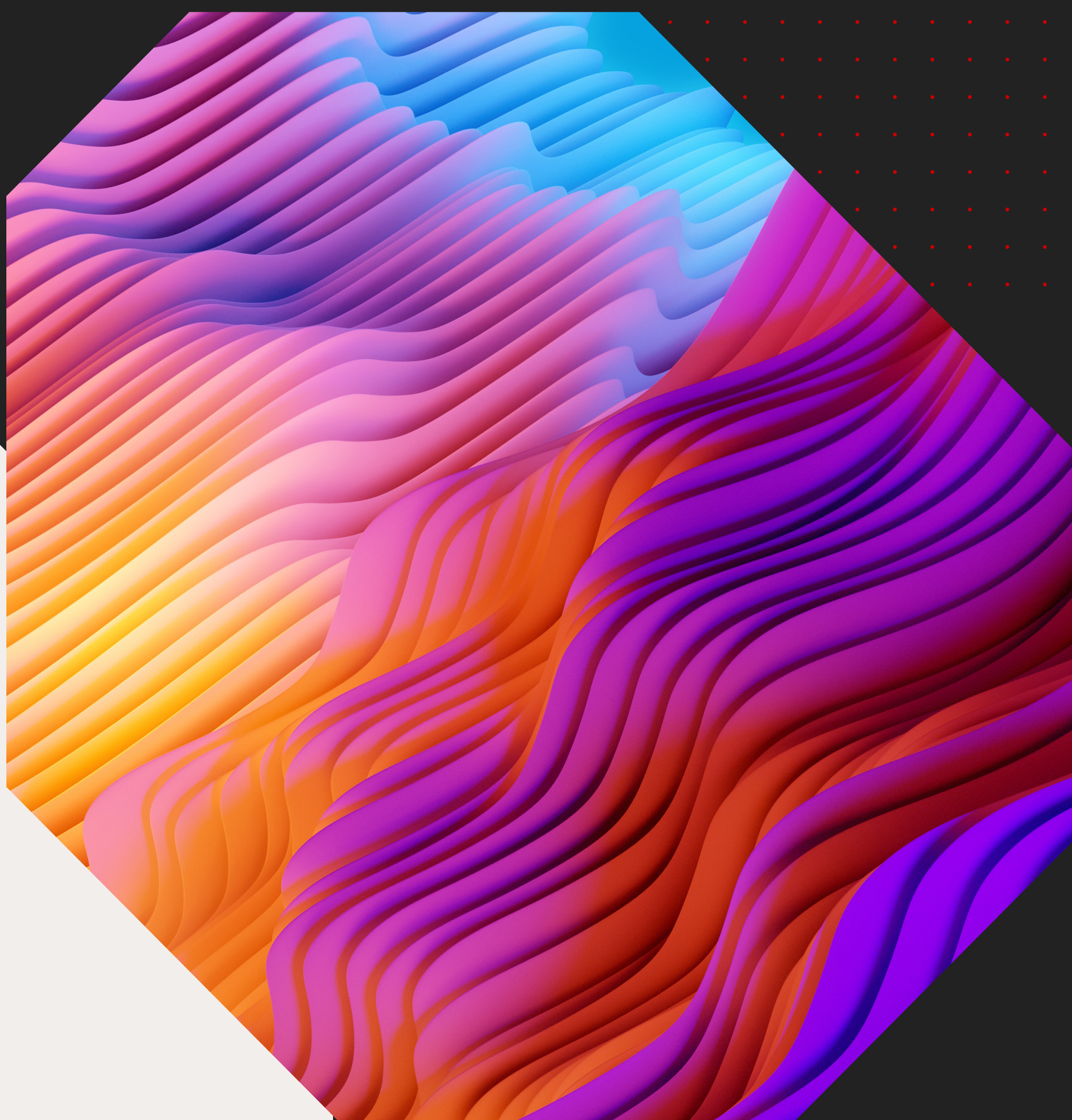
White Paper

Hitachi Polyphase Erasure Coding

Hitachi Vantara Combines High Data Capacity Efficiency and High Performance With Patented Erasure Coding Technology



Hitachi Vantara

White Paper

Table of Contents

Introduction

Storage admins and cloud architects are constantly challenged to provide cost-effective, high-performance and scalable data protection in the face of ever-increasing data volumes and demanding application SLAs. Traditional methods like RAID and mirroring offer robust protection but often come at the cost of reduced storage utilization or increased latency. Hitachi Vantara delivers patented Hitachi Polyphase Erasure Coding (HPEC) to address these challenges.

HPEC provides a unique approach to erasure coding, delivering a combination of high data capacity efficiency, exceptional read performance, and flexible data protection options. This paper introduces Polyphase Erasure Coding, exploring its architecture, key features and the benefits it offers compared to traditional data protection methods. This solution enables users to optimize their storage infrastructure for diverse workloads, achieving both reliability and performance without compromising capacity.

Data Protection Technologies in Enterprise Storage

For enterprise storage, maintaining high reliability is important for uninterrupted business operations, often exceeding the requirements of conventional HDDs and SSDs. While traditional block storage uses RAID technology to enhance data protection by bundling multiple disks, offering benefits like fault tolerance and improved read performance depending on the RAID level (e.g., RAID-1 mirroring, RAID-5/RAID-6 parity), it also presents drawbacks.

RAID's drawbacks include complex rebuilds, limited scalability and an inability to protect against node-level failures, which are common in distributed systems and cloud architectures. Additionally, RAID can be inefficient in terms of capacity utilization and often requires specialized hardware.

These drawbacks include the overhead of parity calculations (impacting write performance in some configurations), the complexity of rebuild processes after a disk failure (potentially leading to further risk during rebuild), and the fact that RAID primarily protects against disk-level failures, not broader system outages or logical corruption.

In contrast, software-defined storage (SDS) builds systems on general-purpose x86 servers, necessitating server-level redundancy mechanisms. For instance, SDS often employs techniques like data replication across multiple storage nodes, such as mirroring, to ensure data availability — even in the face of entire node failures. This approach offers greater resilience against hardware failures beyond individual disks but introduces its own set of considerations regarding network bandwidth utilization and consistency management across distributed copies.

Erasure Coding Features

While replicating data across multiple storage nodes, as seen in mirroring, boosts reliability, it comes at a cost: reduced data capacity efficiency proportional to the number of copies. For instance, a simple mirror effectively halves the usable storage space. Addressing this inherent trade-off between redundancy and capacity utilization is where erasure coding (EC) comes into play. Erasure coding is a data protection technique that offers a more storage-efficient alternative to mirroring. By mathematically dividing data into blocks and calculating parity information, EC can reconstruct lost data from a subset of the stored blocks. Unlike mirroring, which simply duplicates data, EC uses algorithms like Reed-Solomon coding to achieve higher levels of fault tolerance with less storage overhead.

EC divides data into multiple data blocks and generates the redundant data necessary to restore the data blocks, called parity blocks, from the multiple data blocks. Data reliability is ensured by distributing data blocks and parity blocks across multiple storage nodes.

For example, 2D+1P divides data into two data blocks, a and b, as shown in Figure 1 below, and then generates one parity block by taking the exclusive OR (XOR) of the a and b data blocks. By placing the a, b and parity blocks in different storage nodes, even if one of the three storage nodes fails, it is possible to recover by using the two surviving blocks out of the two data blocks and parity block. In terms of capacity, since 1.5 data blocks are generated from the original data, it is theoretically possible to achieve a data capacity efficiency of approximately 66%[1], which is equivalent to one redundancy copy method.

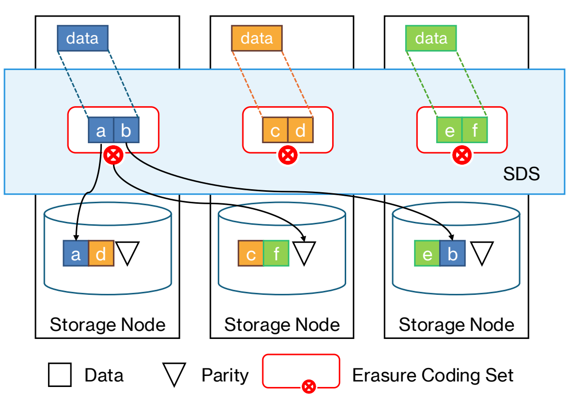


Figure 1. EC divides data into multiple data blocks and generates the redundant data necessary to restore the data blocks.

Hitachi Polyphase Erasure Coding

Data Placement for Fast Reads

In normal EC, the original data is distributed across multiple nodes, so in both read and write cases, access across the network between storage nodes occurs in the I/O path. In contrast, the patented Hitachi Polyphase Erasure Coding[2] allocates data locally and distributes parity blocks so that I/O across the network between storage nodes does not occur in read operations.

HPEC optimizes read performance by prioritizing local drive access. Accessing data locally within a storage node avoids the latency and overhead associated with network communication between nodes. Local drive access can provide significantly lower latency and higher IOPS compared to network-based reads, as it eliminates network congestion and protocol overhead. This design choice enables HPEC to achieve read performance comparable to mirroring in many scenarios.

For example, in Figure 2, the a and b data blocks are placed together in a local drive, but the parity blocks required for the a and b data blocks are distributed in a storage node separate from the a and b data blocks. As a result, local read is realized while maintaining the same redundancy and data capacity efficiency as normal EC.

This method is especially useful for workloads that frequently reference data, such as web applications and analysis systems.

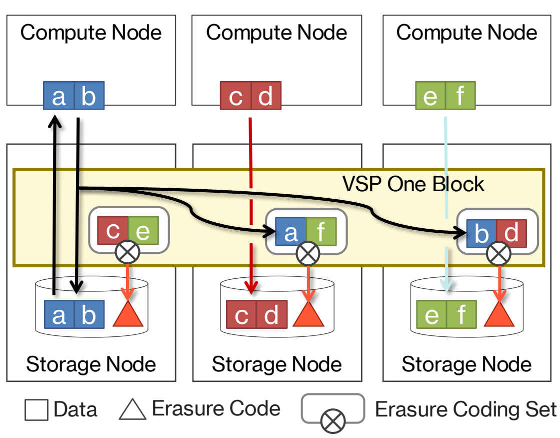


Figure 2. Parity blocks distributed in a storage node separate from the a and b data blocks allow local read while maintaining the same redundancy and data capacity efficiency as normal EC.

1 In reality, the value is lower than the theoretical value because data other than user data, such as metadata, also consumes space.

2 U.S. patent: US 10,185,624, US10,496,479

Response Time Equivalent to Mirroring by Writing Logs

A common trade-off with erasure coding is reduced write performance (throughput and response time) compared to mirroring due to the computational overhead of parity generation. However, with Polyphase Erasure Coding, Hitachi Vantara overcomes this limitation of latency. By employing a strategy where write operations are acknowledged upon the persistent logging of the data across nodes before the synchronous calculation and distribution of parity, HPEC delivers write response times on par with mirroring.

Figure 3 depicts the behavior when data block a is updated to a'. The log of updating data block a to a' is copied to multiple storage nodes according to the redundancy level (process 1), and the response is returned to the compute node as soon as the log redundancy is completed (process 2).

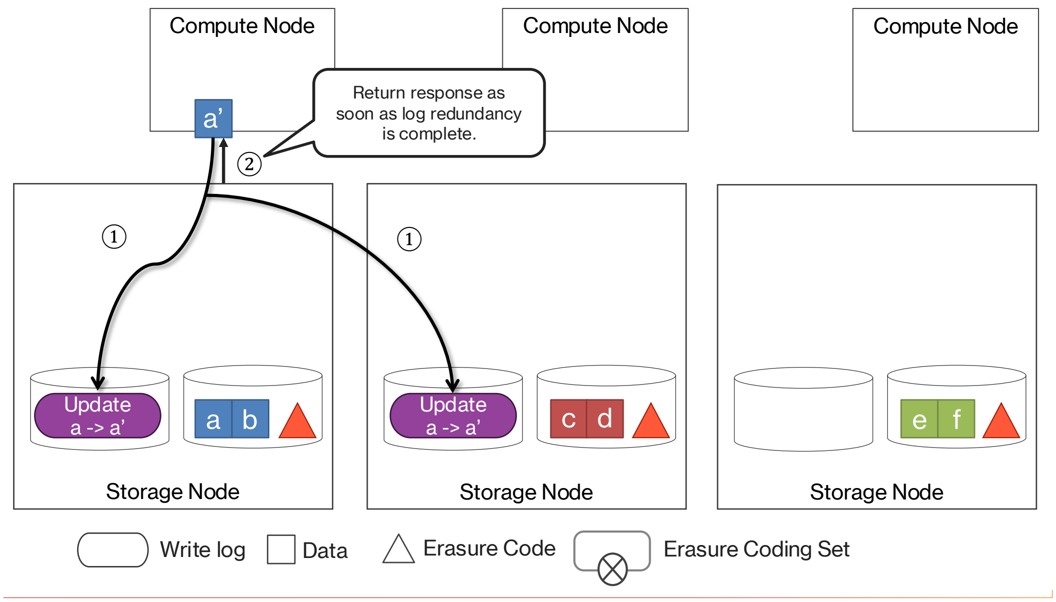


Figure 3. Polyphase Erasure Coding delivers write response times on par with mirroring.

Real data and parity updates are performed asynchronously with I/O after the response to the compute node.

In Figure 4, the log is used to update the data block a and the parity block consisting of data block a and data block f to consist of data block a' and data block f.

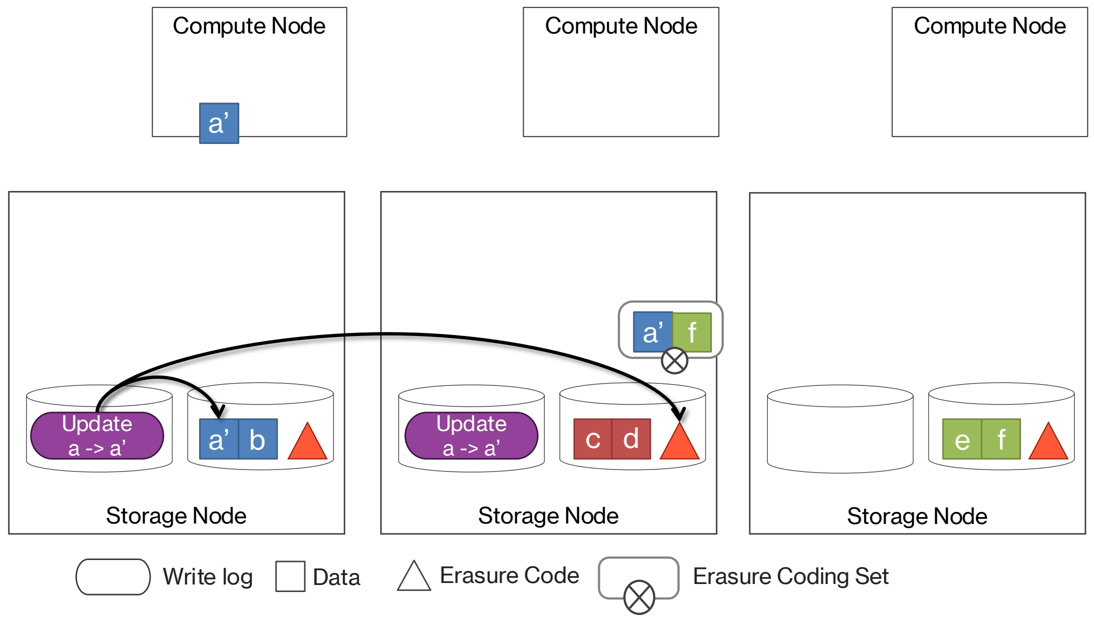


Figure 4. Real data and parity updates are performed asynchronously.

Rebuild Behavior

In the event of data corruption, data can be recovered using data and parity as in normal EC.

Figure 5 depicts the case where data block c is missing. Data block c can be recovered by taking the XOR of the parity block consisting of data block c and data block e.

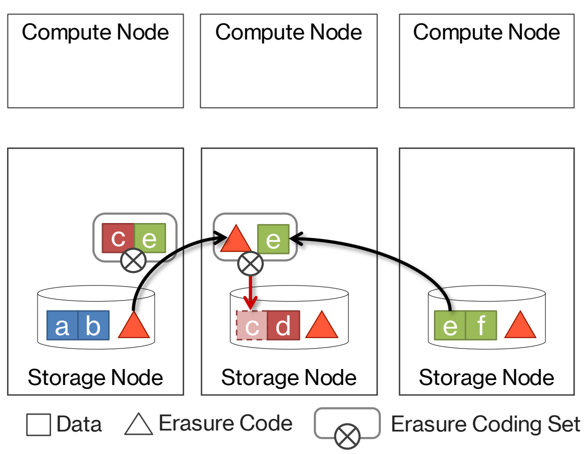


Figure 5. Recovery when data block c is missing.

Redundancy Method Provided by Hitachi Vantara

Hitachi Vantara offers mirroring and HPEC 1-parity and 2-parity EC. Table 1 examines the features of each redundancy method.

Table 1. Hitachi Polyphase Erasure Coding and mirroring: 1-parity and 2-parity.

Redundancy Technology	Feature	Fault Tolerance	Data Capacity Efficiency	Performance	Use Case/ Workload Recommendations
Mirroring	Performance oriented	Single-point fault	40-48% of	High speed	High-performance databases, applications requiring very low latency
Polyphase Erasure Coding	Emphasis on data capacity efficiency	Single-point fault	60-75% of	Read is equivalent to mirroring	General-purpose storage, large-scale data
HPEC 4D+1P				Write response is equivalent to mirroring, but IOPS is less than mirroring	Repositories, applications where capacity optimization is critical
Polyphase Erasure Coding: HPEC 4D+2P	Fault tolerance	Fixed two-point fault	55-65% of	Read is equivalent to mirroring Write response is equivalent to mirroring, but IOPS is less than 4D+1P	Mission-critical applications, workloads with stringent availability requirements, environments where multiple concurrent failures are a concern

Mirroring Architecture

Mirroring simply copies data across storage nodes, which has better performance but lower data capacity efficiency than other methods.

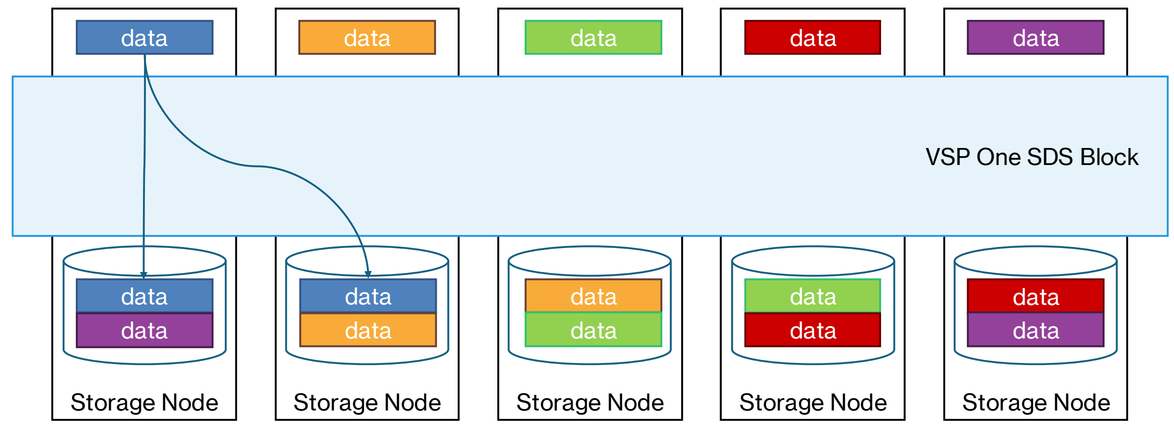


Figure 6. Example of Mirroring — copying data across storage nodes

HPEC 4D+1P Architecture

HPEC 4D+1P generates one parity block for every four data blocks, resulting in high data capacity efficiency. Performance is inferior to mirroring because of the parity calculation, but read performance is equivalent to mirroring by taking advantage of HPEC's features.

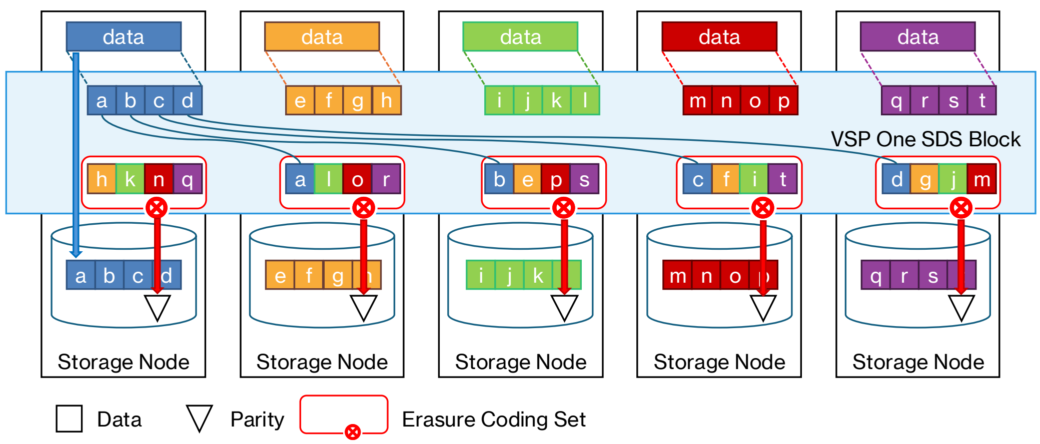


Figure 7. In Polyphase Erasure Coding 4D+2P, parity calculation is performed in two parts.

HPEC 4D+2P Architecture

In the first of two parts of parity calculation, a parity block is generated from a block of data blocks a, b, c, d. In the second part of parity calculation, two parities are generated from the parity generated in the first parity calculation and the data blocks stored in other storage nodes. The second parity calculation generates two parities from the parity generated in the first calculation and the data blocks stored in other storage nodes to realize 4D+2P, where the data blocks are local and the parity blocks are distributed.

Since two parity blocks are generated for every four data blocks, data capacity efficiency is between Mirroring and HPEC 4D+1P, as shown in Figure 8. Throughput is inferior to HPEC 4D+1P due to multiple parity calculations, but read performance is equivalent to Mirroring/HPEC 4D+1P.

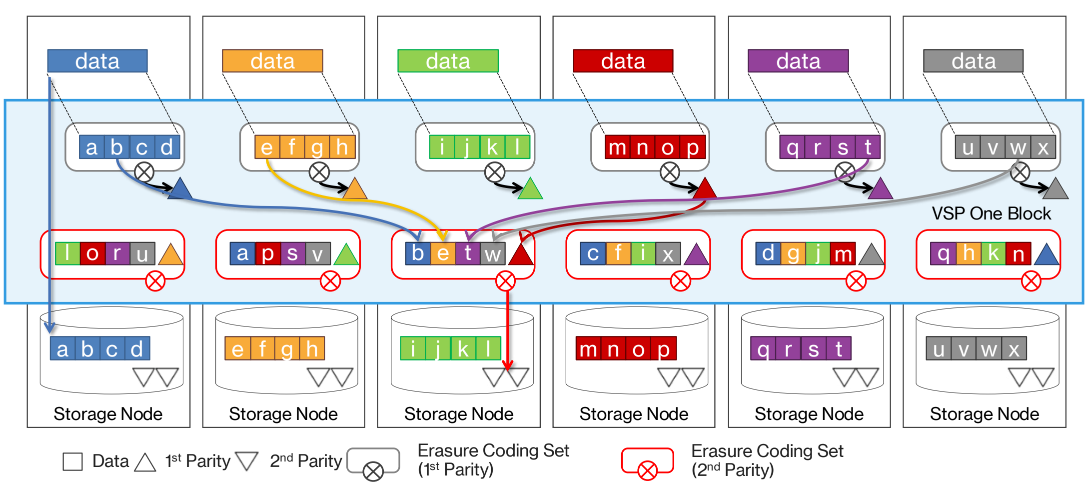


Figure 8. Two parity blocks are generated for every four data blocks, so data capacity efficiency is between mirroring and Polyphase Erasure Coding 4D+1P.

Summary

Hitachi Vantara addresses the balance between data capacity efficiency and high performance through its implementation of Hitachi Polyphase Erasure Coding (HPEC), a patented erasure coding scheme.

HPEC is unique in that if offers flexible reliability options, extending beyond single parity to include dual parity for enhanced fault tolerance, alongside traditional mirroring for workloads prioritizing performance. This versatility enables users to strategically select the optimal data protection method — whether mirroring, HPEC 4D+1P, or HPEC 4D+2P — tailored to the specific demands of their diverse application workloads.

Get more information about Hitachi Polyphase Erasure Coding and Hitachi Software Defined Storage.

Learn more →

About Hitachi Vantara

Hitachi Vantara is transforming the way data fuels innovation. A wholly owned subsidiary of Hitachi, Ltd., we're the data foundation the world's leading innovators rely on. Through data storage, infrastructure systems, cloud management and digital expertise, we build the foundation for sustainable business growth.

Hitachi Vantara

Corporate Headquarters
2535 Augustine Drive
Santa Clara, CA 95054 USA
hitachivantara.com | community.hitachivantara.com

Contact Information
USA: 1-800-446-0744
Global: 1-858-547-4526
hitachivantara.com/contact

HV-BTD-WP-HPEC-17Sept25-A