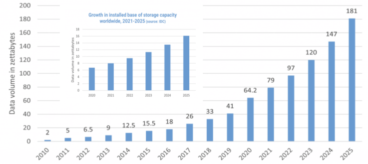
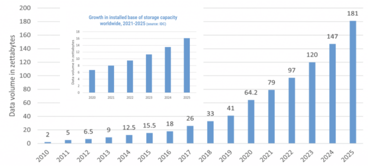
Energy Efficient Data Management Through Erasure Codes

Volume of data created and replicated is expected to reach 181 ZB in 2025[1]



Energy Efficient Data Management Through Erasure Codes

Volume of data created and replicated is expected to reach 181 ZB in 2025[1]

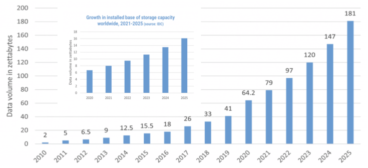


Data storage requires high availability and fault tolerance:

- Replication can provide strong fault tolerance to multiple failures but with high storage overhead.
- Erasure Code can offers is a more compact fault tolerance mechanism but with energy and network challenge due to encoding and data access during repair

Energy Efficient Data Management Through Erasure Codes

Volume of data created and replicated is expected to reach 181 ZB in 2025[1]



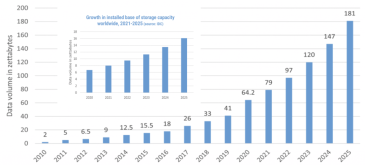
Data storage requires high availability and fault tolerance:

- Replication can provide strong fault tolerance to multiple failures but with high storage overhead.
- Erasure Code can offers is a more compact fault tolerance mechanism but with energy and network challenge due to encoding and data access during repair

How to coordinately optimize the performance and energy of data-intensive applications when adopting EC in cloud storage systems through innovative data placement and retrieval, and cost-effective code conversion and data repair.

Energy Efficient Data Management Through Erasure Codes

Volume of data created and replicated is expected to reach 181 ZB in 2025[1]



Data storage requires high availability and fault tolerance:

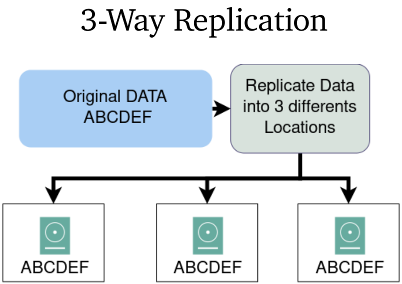
- Replication can provide strong fault tolerance to multiple failures but with high storage overhead.
- Erasure Code can offers is a more compact fault tolerance mechanism but with energy and network challenge due to encoding and data access during repair

How to coordinately optimize the performance and energy of data-intensive applications when adopting EC in cloud storage systems through innovative data placement and retrieval, and cost-effective code conversion and data repair.

- Provide an in-depth analysis of the energy footprint of EC in the CEPH distributed storage systems.
- Observation: CephFS read under EC is not very energy efficient. Next goal: Improve EC Ceph data retrieval

Source: IDC "Global DataSphere Forecast, 2021-2025"

Replication At Scale

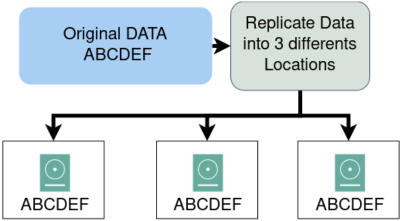


Replication

- Multiple copies of the same data are stored across different nodes (locations).
- If a node fails, data can be read and repaired from any other replica.
- Provides fast repair and ensures data availability.

Replication At Scale

3-Way Replication



Replication at Scale

WHY? In large-scale storage systems, failure is the norm - tolerance for multiple failures is essential

- In 2013, Facebook reported that : A cluster of 3,000 nodes experiences roughly 20 non-transient node failures every day.[1]

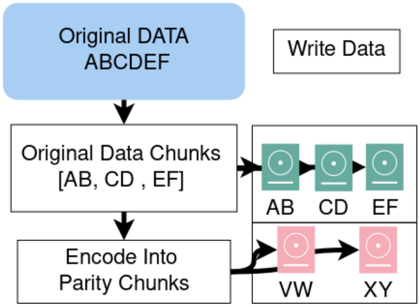
COST. 3-way replication generates 200% storage overhead, rising to 300% with 4-way replication

- Amplifies the storage and energy problem.

[1] Maheswaran Sathiamoorthy, et al "XORing Elephants: Novel Erasure Codes for Big Data"

Erasure Coding

Write data with Erasure Coding



Advantages

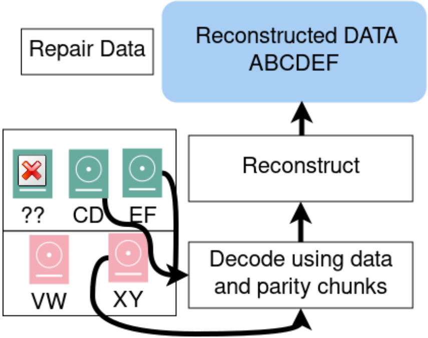
- Reduces storage overhead: RS(3,2) has 60% overhead while 3-way replication has 200% overhead for the same fault tolerance.
- Less data exchange during writes.

Tradeoffs

- CPU overhead (and power) for encoding and decoding.
- Increase in data exchanged during repair.
- No data locality.

Erasure Coding

Recover from disk failure



Advantages

- Reduces storage overhead: RS(3,2) has 60% overhead while 3-way replication has 200% overhead for the same fault tolerance.
- Less data exchange during writes.

Tradeoffs

- CPU overhead (and power) for encoding and decoding.
- Increase in data exchanged during repair.
- No data locality.

Open Questions?

In Real world cluster what can affect the reliability of the system compared to the theory?

Open Questions?

In Real world cluster what can affect the reliability of the system compared to the theory?

When measuring energy what should be taken into account? And what data to collect?

Open Questions?

In Real world cluster what can affect the reliability of the system compared to the theory?

When measuring energy what should be taken into account? And what data to collect?

What type of scenario is interesting to run when comparing energy of those systems?

Open Questions?

In Real world cluster what can affect the reliability of the system compared to the theory?

When measuring energy what should be taken into account? And what data to collect?

What type of scenario is interesting to run when comparing energy of those systems?

Any Idea/Research focus to improve the energy?

Comparative Table

Erasure Coding Family	Storage Overhead	Encoding	Repair Bandwidth	Multiple Failure Repair Bandwidth
Reed Solomon	Best	Fast	High	High
Clay Code (RG)	Best	Slow	Lowest	High
Ceph LRC	Good: Overhead from Local Parity	Fast	Small	High
Azure LRC	Good: Overhead from Local Parity	Fast	Small	High
SHEC	Decent: Overhead from overlapping	Fast	Small	Smaller