SNIA DEVELOPER CONFERENCE
SDC 25
By Developers FOR Developers

Hyatt Regency Santa Clara, CA
September 15-17, 2025

Asynchronous Erasure Coding for Scalable, Resilient, and Efficient Storage

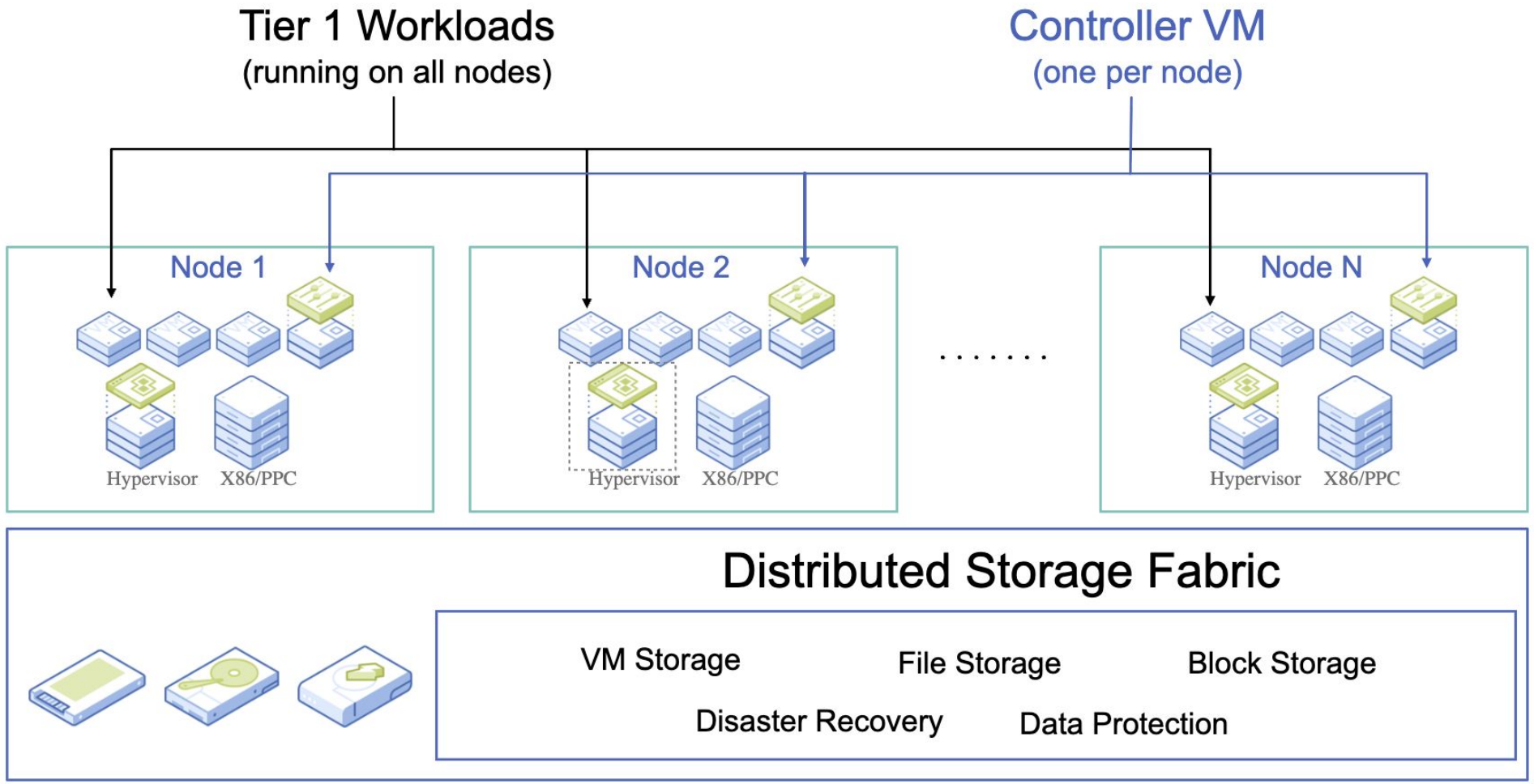
- Sarthak Moorjani, Snehal Kamble
  Nutanix, Inc

www.sniadeveloper.org

Agenda

- Background of Nutanix Architecture
- Data reduction techniques
- Background of Erasure Coding
- Erasure coding at Nutanix - Terminology & Evolution
- Erasure coding Overwrites and garbage
- Flexible strip sizes
- Inline erasure coding

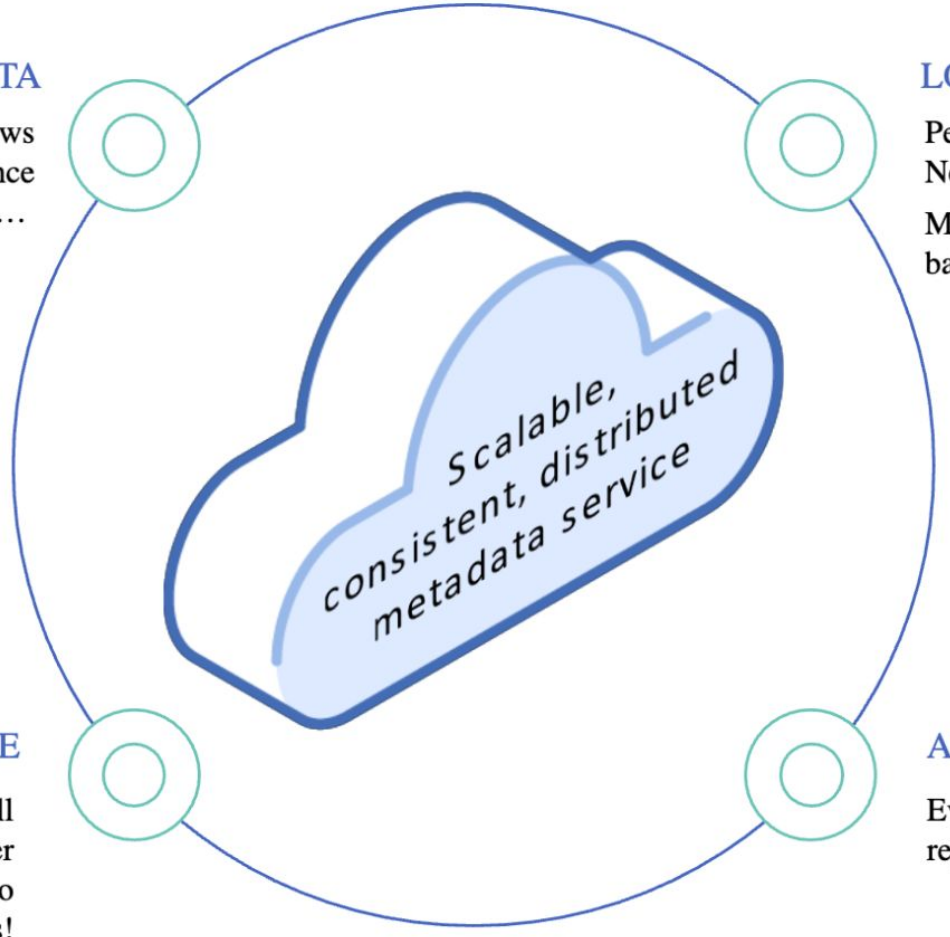
Background of Nutanix Architecture



Background of Nutanix Architecture

FLEXIBLE, FINE-GRAINED METADATA

Dynamic disk / replica selection – allows heterogeneous clusters, flexible fault tolerance
…



LOCALITY OF DATA

Performance, Scalability together with Network efficiency.

MapReduce-based smart algorithms for disk balancing, ILM…

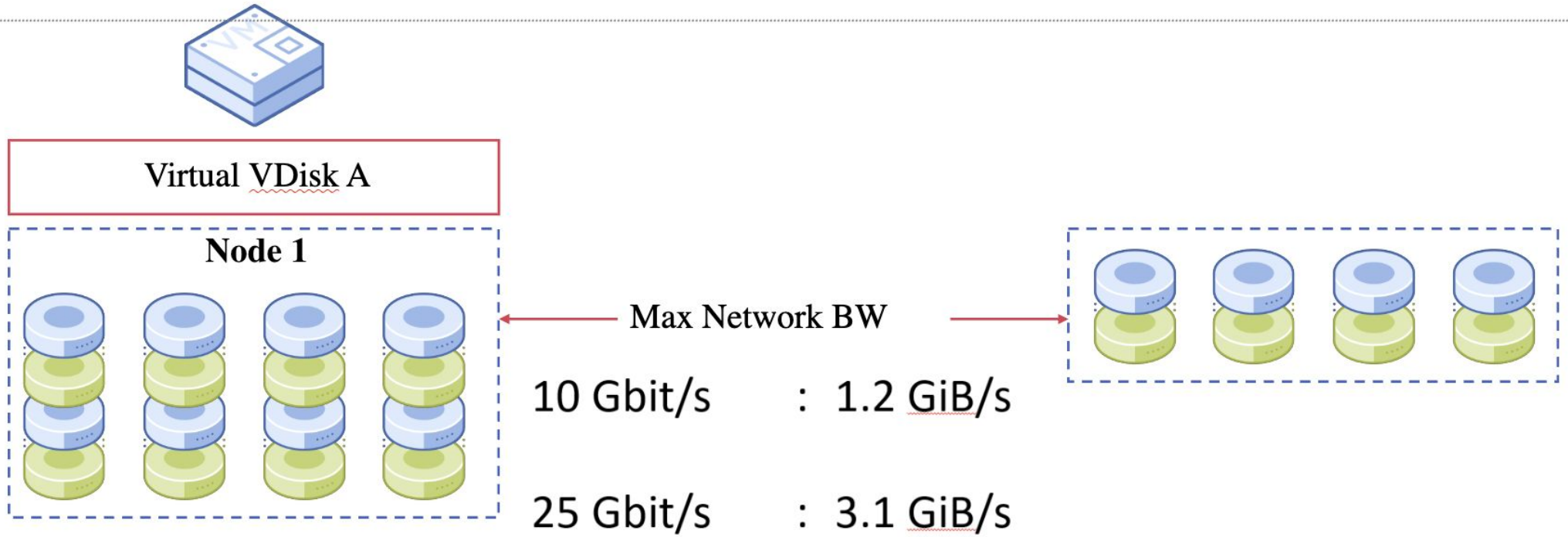
TRUE SCALABLE ARCHITECTURE

Add New Nodes instant perf boost to all VMs. With Node failure, entire cluster participates in healing. Faster recovery, no bottlenecks!

ALL WRITES ALWAYS PROTECTED

Even when nodes fail, new writes are replicated, we never write with RF1

Background of Nutanix Architecture - Data locality



- VM data access not limited by network capacity.
- NVME throughput without investing in costly network infrastructure.

Data reduction techniques

Compression

- Data Dependent

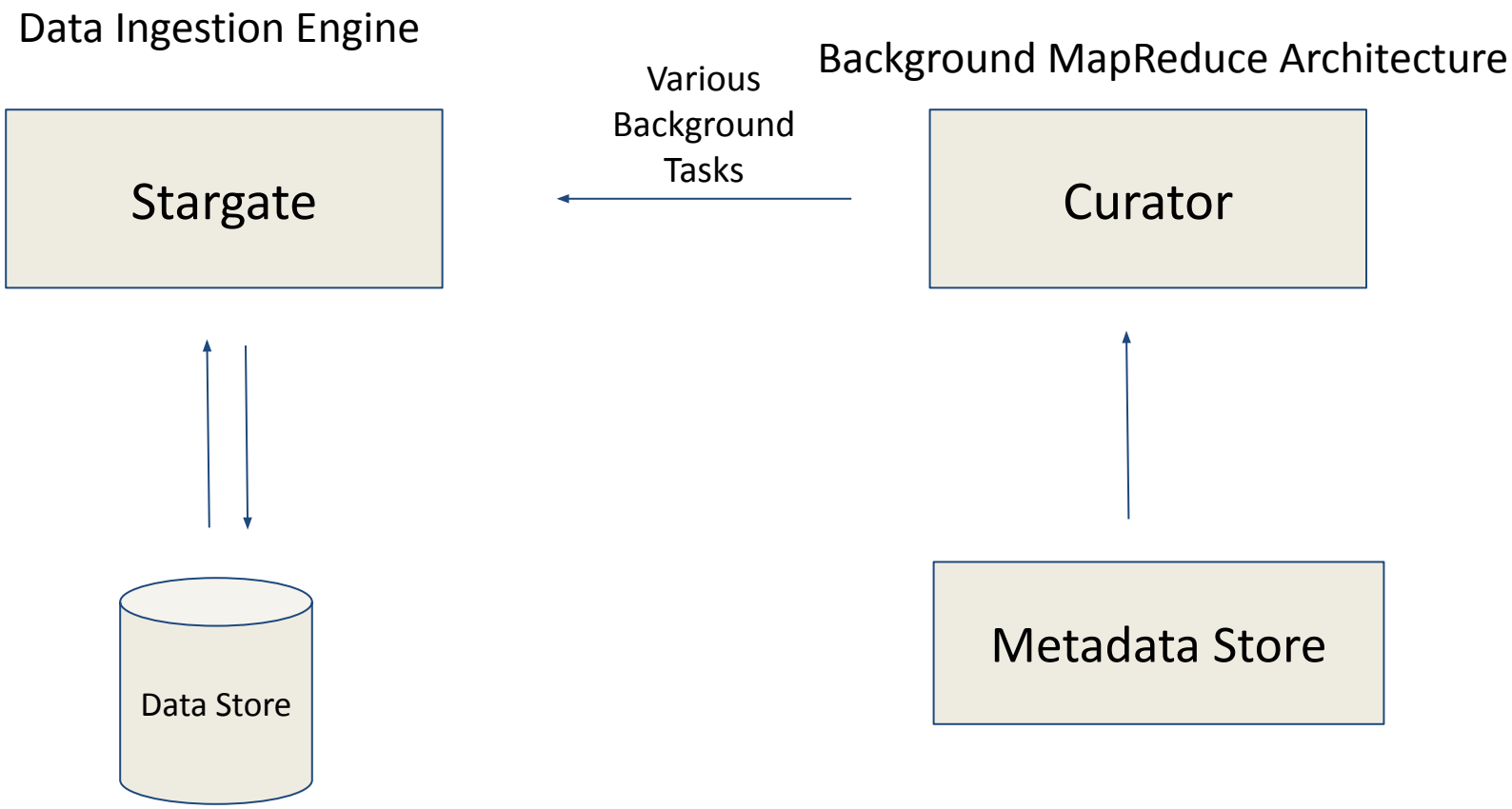
Deduplication

- Peer Data Dependent

Erasure Coding

- Data Agnostic

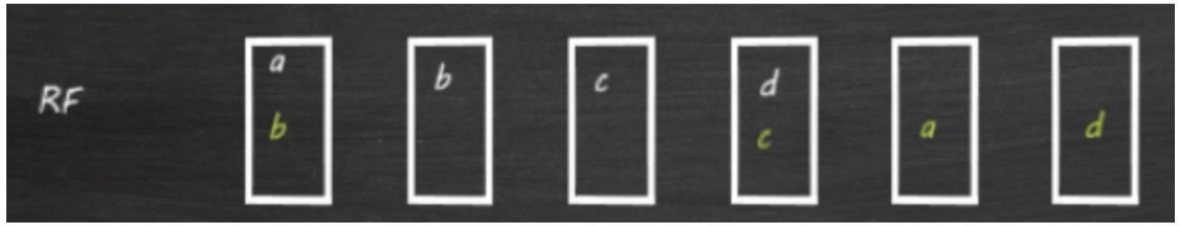
Insert slide about Stargate / Curator



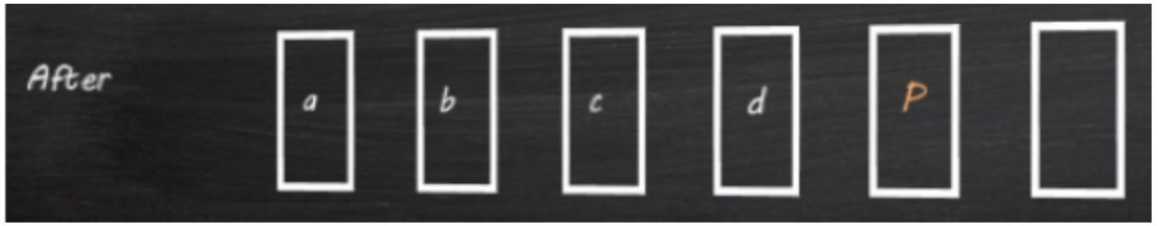
Background of Erasure Coding

- A data protection technique that splits data into chunks and generates parity chunks using mathematical operations.
- Enables reconstruction of lost data even if multiple chunks/nodes fail.
- More space-efficient than replication (less overhead than storing full copies).

Before Erasure Coding:



After Erasure Coding



Erasure Coding - Key Characteristics

1. **Storage Efficiency** – less overhead than replication.

2. **High Durability** – tolerates multiple node/disk failures.

3. **Computational Overhead** – encoding/decoding adds CPU cost.

4. **Applications** – widely used in cloud storage, HDFS, large-scale distributed systems.

Erasure Coding - Terminology

- What is an erasure code strip?
- Erasure code strip members
  - Info member - data blocks to be encoded.
  - Parity members - parity blocks generated after encoding.
- Erasure code strip size
  - N/K (N >= 2, N >= K)

- We internally decide the value of N and also support flexible strip sizes if we cannot fit an optimal strip.

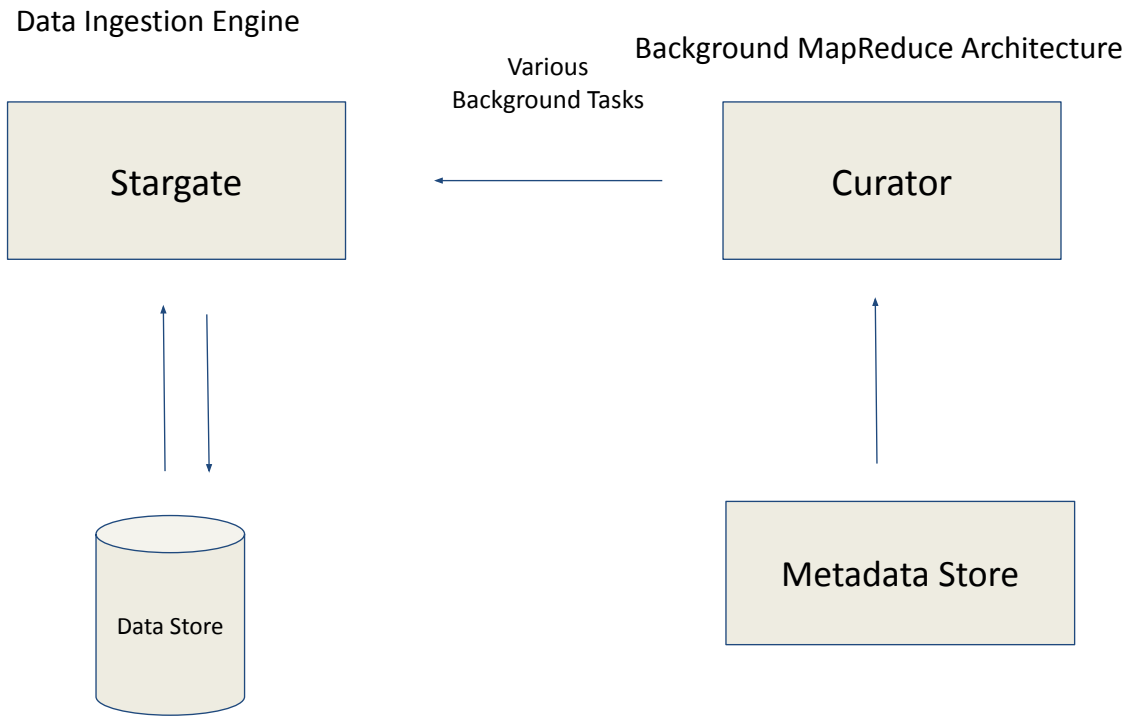
Erasure Coding at Nutanix - Evolution

Erasure Code - A post-process data reduction technique.

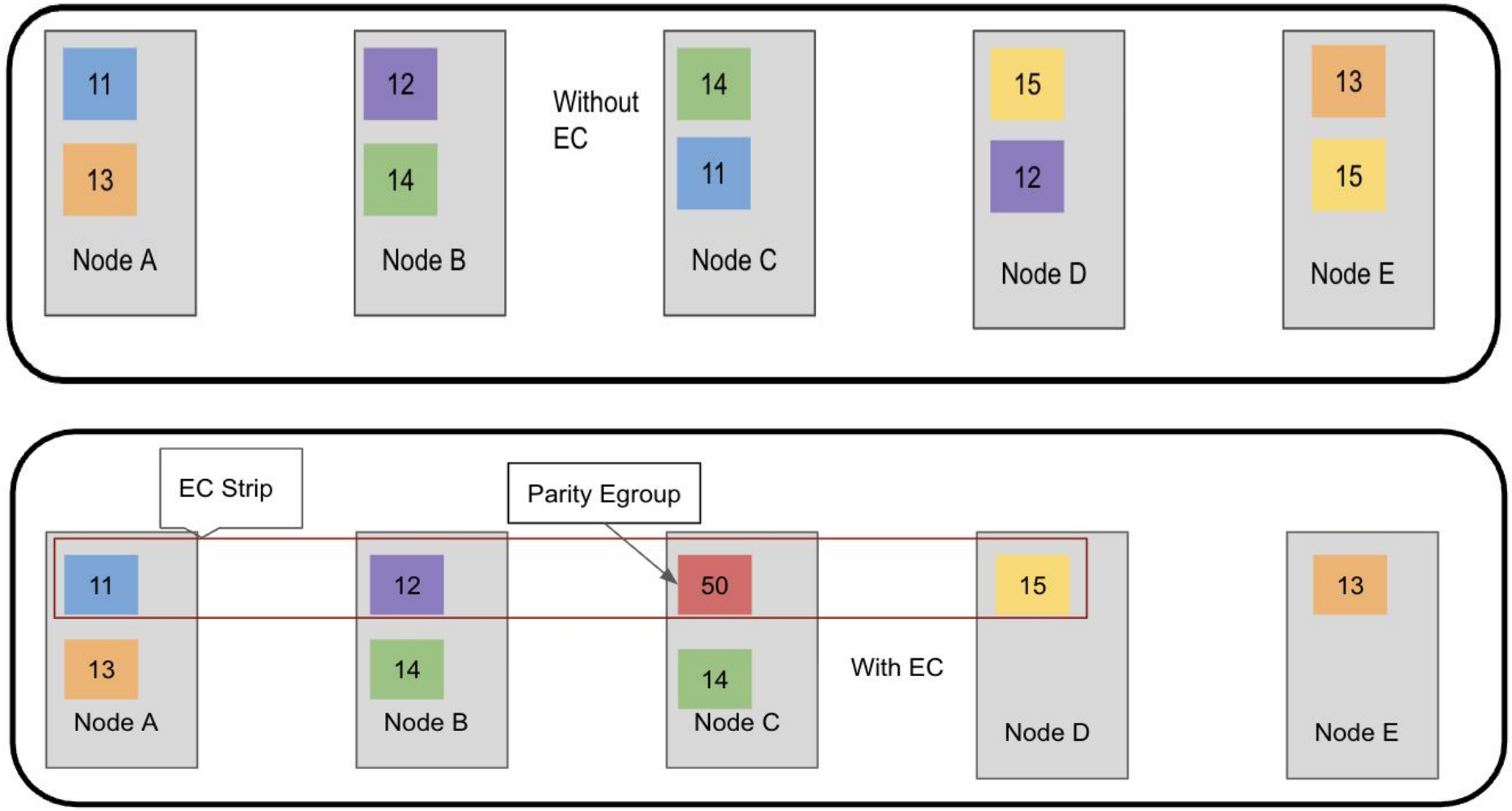
- Works on write cold data blocks.
- Curator full scans identify eligible data blocks from different nodes.
- Erasure code is generally the last transformation performed on the data.

Erasure Coding at Nutanix - Encoding

- Data is ingested and written in a replicated fashion via Stargate.
- Erasure coding is primarily driven through our background operations triggered by Curator.

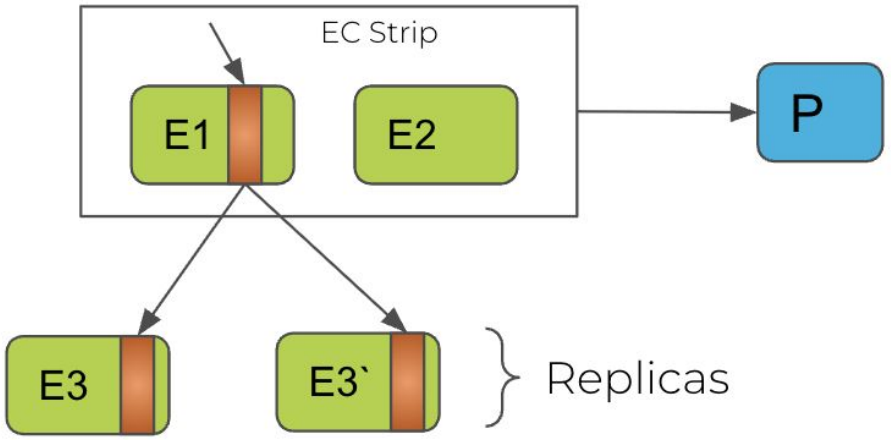


Erasure Coding at Nutanix - Encoding



Erasure Coding at Nutanix - Overwrites and Garbage

- In place overwrites are not supported since the CPU cost of those writes is high.
- If an overwrite comes on an erasure coded data block, the live data is migrated out in replicated form, leaving garbage behind.
- This garbage, however cannot be reclaimed immediately since parity data depends on it.

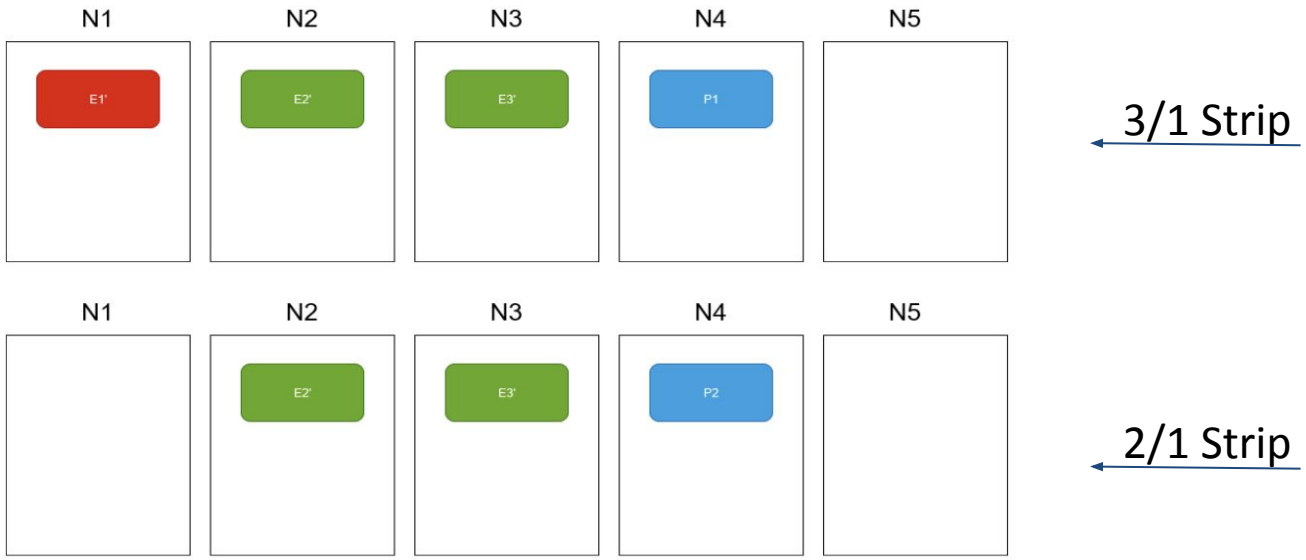


Handling EC Garbage - EC Replace

- The data blocks which are overwritten eventually become garbage.
- When that happens, Curator scans identify this garbage and try to replace it with non erasure coded data blocks.
- During this process, some non erasure coded data becomes erasure coded, and we also clean up garbage.

Addressing garbage: Recode Shrink

- As we saw previously that overwrites create garbage.
- However it is possible that we could not find a replacement candidate.
- We, then extend the EC recode operation to also shrink stripes to smaller sized EC stripes to clean up garbage more aggressively.



Flexible Strip Sizes

- Due to memory limitations, it is sometimes not possible to have an entire picture of the metadata in memory at the same time.

- This means, it is difficult to always find candidates in such a way that we could form similar sized stripes according to the configuration.

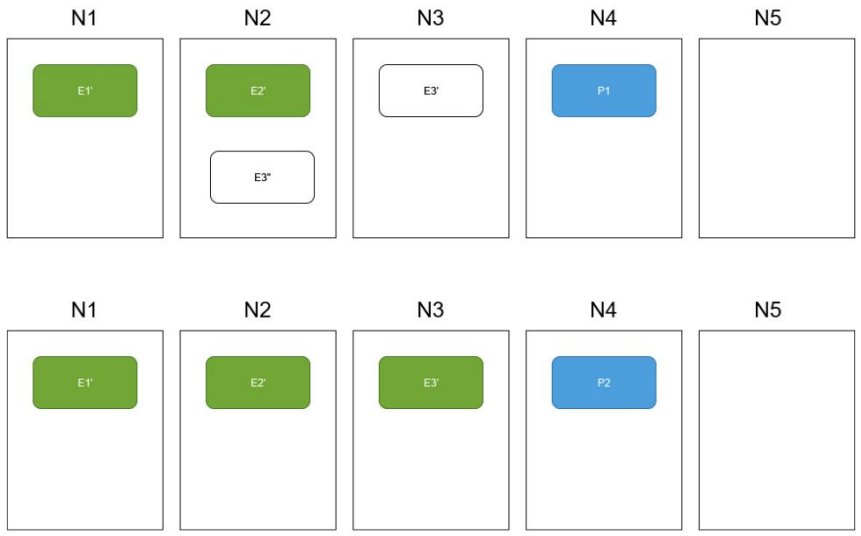
- We introduced the concept of flexible strip sizes to overcome this limitation and still provide space savings.

- For instance, if the configuration allows 4/1 strips and we are unable to find candidates for 4/1 strips, Curator still forms 3/1 and 2/1 strips since they are still better than data blocks in replicated form.

Recode - Augmentation

- As we introduced flexible strip sizes, naturally we had shorter strips in the system.

- To improve this, we introduced a new operation named Recode:Augment. As the name suggests, this would augment the shorter strips to larger ones by using the replicated data blocks.



Total: 5 data block replicas
EC savings ratio: 1.33

Total: 4 data block replicas
EC savings ratio: 1.5

Nearline erasure coding

- Based on data patterns, Stargate detects and organize data in such a way that the data can be immediately EC’ed even though it is write hot instead of waiting for a background scan.

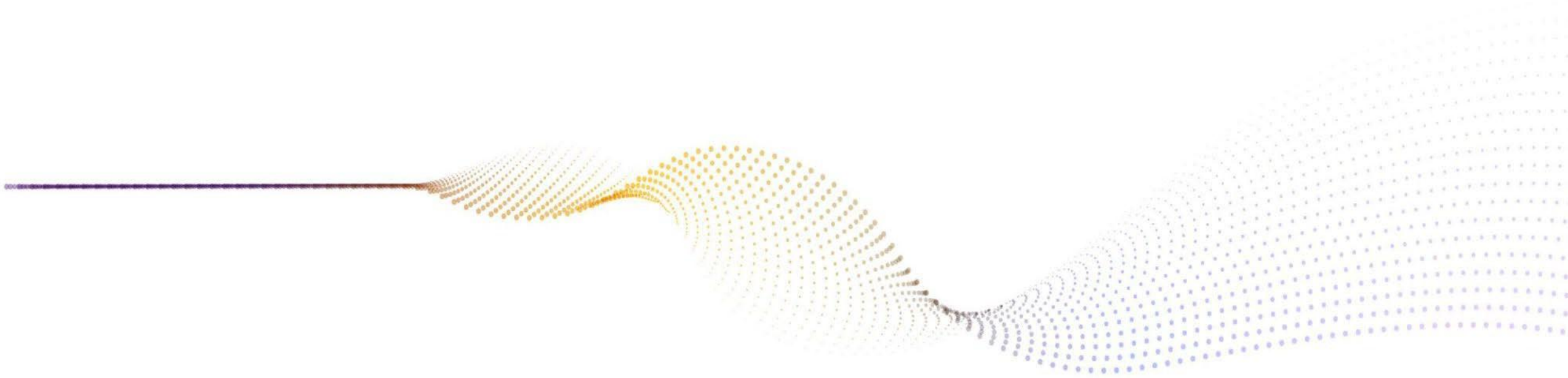
- Since these are in memory computations, nearline erasure coding is limited and we primarily still rely on background scans to erasure code the data.

- This methodology when enabled WORM workloads will achieve Erasure coding savings faster.

- If there are overwrites on such data, then they are treated in the exact same way as the EC stripes created via Curator are treated.

Enabling EC on Nutanix Clusters

- Seamless and easy.
- User are only required to select data stores to enable erasure coding.
- The system is fully capable of figuring out the best size of the EC stripes for maximum EC savings.



Thank you for attending!

