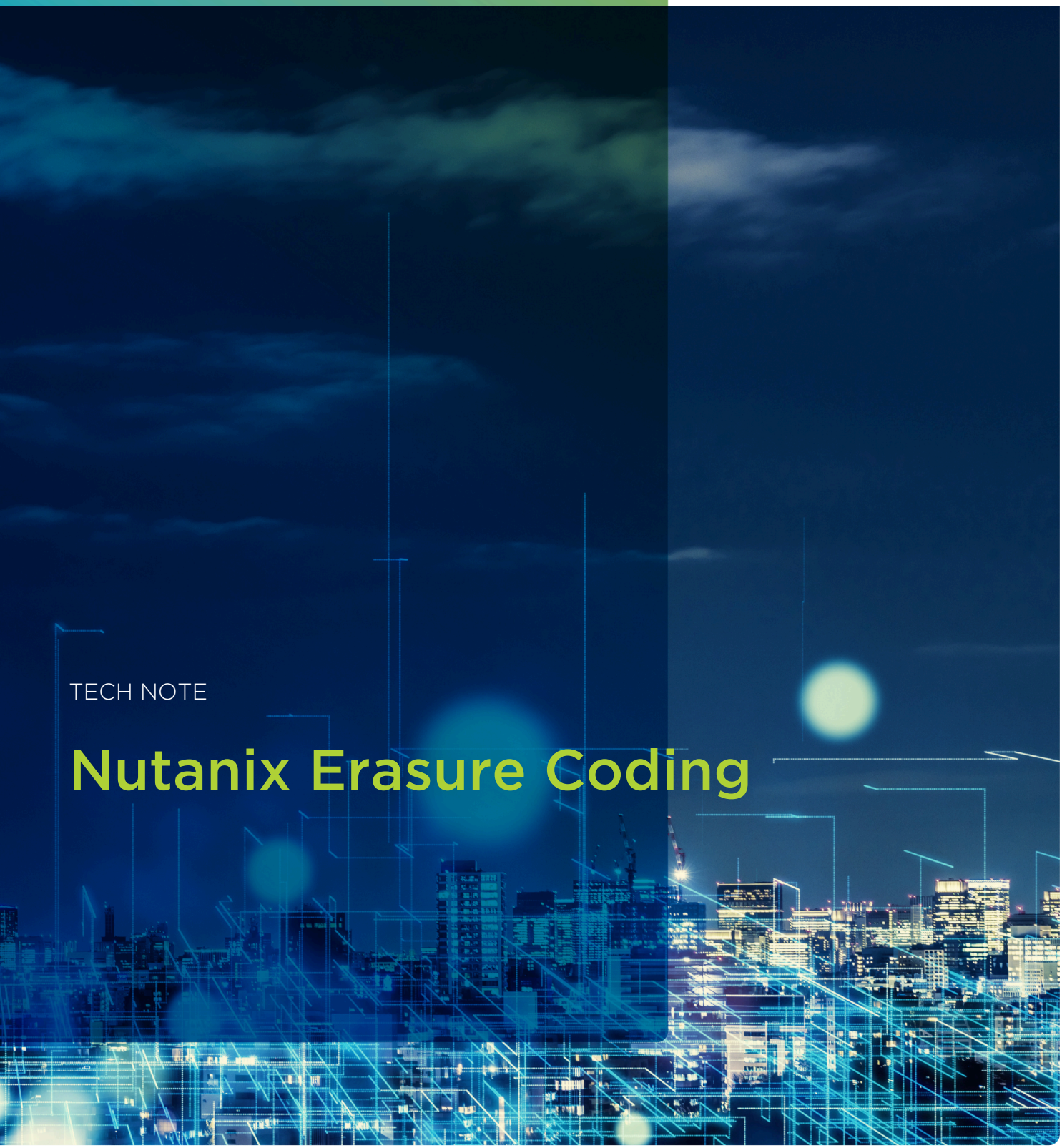
v1.3 | March 2022 | TN-2002



TECH NOTE

Nutanix Erasure Coding

NUTANIX
YOUR ENTERPRISE CLOUD

Copyright

Contents

1. Executive Summary

This document provides information about the Nutanix storage-saving feature erasure coding (EC-X), including how it works and when to use it. The amount of space you can save when using EC-X varies. At a minimum, the overhead drops from 2 to 1.5 when you are using Nutanix replication factor 2 with EC-X.

2. Introduction

Audience

This tech note is part of the Nutanix Solutions Library. We wrote it for everyone who wants to know how erasure coding (EC-X) works, what to expect in terms of storage gains, and the requirements for using EC-X.

Purpose

In this document, we cover the following topics:

- Overview of the Nutanix EC-X feature.
- EC-X in action.
- What you can expect when you use EC-X.

Document Version History

Version Number	Published	Notes
1.0	January 2020	Original publication.
1.1	June 2020	Updated Nutanix overview, AOS versions, and the Recommendations section.
1.2	February 2021	Updated to include inline EC-X with Nutanix Objects.
1.3	March 2022	Updated the Erasure Coding Basics section.

3. Erasure Coding Basics

This section covers some of the basic information about erasure coding (EC-X) and Nutanix storage.

When you configure Nutanix storage with a replication factor of 2 or 3, the Nutanix cluster maintains two or three exact copies of the same data on different nodes to ensure data availability. The actual logical capacity available depends on the replication factor you choose. When you use replication factor 2 (also called fault tolerance 1), you have approximately 50 percent capacity available. When you use replication factor 3 (also called fault tolerance 2), you have approximately 33 percent capacity available.

The Nutanix OS provides storage-saving techniques like compression, deduplication, and EC-X, with EC-X applied or compiled as the last step. Before EC-X starts, the data must be write-cold (in other words, there has been no write access to the data for seven days). The amount of space you can save when using EC-X varies based on:

- The amount of cold data.
- The size of your Nutanix cluster.
- The replication factor you configure for Nutanix storage.

The Nutanix storage fabric works with different logical constructs:

- Slice: 32 KB (8 KB is a subregion of a slice that you can address).
- Extent: 1 MB (for deduplicated data, an extent is 16 KB).
- Extent group: Between 1 and 4 MB.
- Container: Logical grouping construct where you place VMs and enable EC-X.

EC-X works at the extent group layer, meaning it uses 1–4 MB data sets when performing its calculations. By default, the cluster automatically uses extent groups that belong to the same virtual disk (vDisk), as this method makes

it easier to perform garbage cleanup, but the cluster can use extent groups from different vDisks if necessary. vDisks on Nutanix are made of virtual blocks (vBlocks), which are 1 MB chunks of virtual address space. Each vDisk in the system is owned by a Nutanix Controller Virtual Machine (CVM) that typically runs on the same Nutanix node as the VM the vDisk belongs to.

EC-X makes a strip from existing data to create parity. The strip width depends on the number of nodes in the Nutanix cluster and the data replication factor configured for the Nutanix container.

EC-X tries to delete the copy of the extent group that is not local to the CVM. For example, if VM1 runs on node1 and has egroup1 on node1 and node2, AOS keeps the egroup1 copy on node1 after the EC-X operation. EC-X places the parity extent group on a different node (not node1) and does not compress the parity bit, even if compression is enabled at the container level. In a hybrid system, AOS places the parity bit on HDD if possible.

Table: EC-X Configuration Options

Number of Nutanix Nodes	Replication Factor 2 EC-X Strip Size (Data/Parity)	Replication Factor 3 EC-X Strip Size (Data/Parity)
4	2/1	N/A
5	3/1	N/A
6	4/1	2/2
7	4/1	3/2
8	4/1	4/2
9	4/1	4/2

The number of Nutanix nodes includes at least one node used for failover with replication factor 2 and at least two nodes used for failover with replication factor 3.

4. Erasure Coding in Action

We provide a few scenarios showing EC-X in action in this section.

The following table compares the EC-X overhead in the different configurations available with traditional replication factor 2 or 3 overhead.

Table: EC-X Overhead by Strip Size: Fault Tolerance 1

Number of Nutanix Nodes	EC-X Strip Size	EC-X Overhead	Replication Factor 2 Overhead
4	2/1	1.5	2
5	3/1	1.33	2
6	4/1	1.25	2
7	4/1	1.25	2
8+	4/1	1.25	2

Table: EC-X Overhead by Strip Size: Fault Tolerance 2

Number of Nutanix Nodes	EC-X Strip Size	EC-X Overhead	Replication Factor 3 Overhead
4	N/A	N/A	N/A
5	N/A	N/A	N/A
6	2/2	2	3
7	3/2	1.67	3
8+	4/2	1.5	3

Four-Node Nutanix Cluster, Replication Factor 2

This scenario includes the following items:

- Four Nutanix nodes

- Two vDisks
  - The CVM on node1 owns vDisk1.
  - The CVM on node3 owns vDisk2.
- Three extent groups per vDisk
- Two identical extent group copies

Four-Node Cluster Before EC-X

Before the EC-X parity calculation, AOS writes each extent group twice to the Nutanix cluster on different nodes:

- vDisk1 egroup1 goes on nodes 1 and 4.
- vDisk1 egroup2 goes on nodes 1 and 2.
- vDisk1 egroup3 goes on nodes 1 and 3.
- vDisk2 egroup1 goes on nodes 2 and 3.
- vDisk2 egroup2 goes on nodes 3 and 4.
- vDisk2 egroup3 goes on nodes 1 and 3.

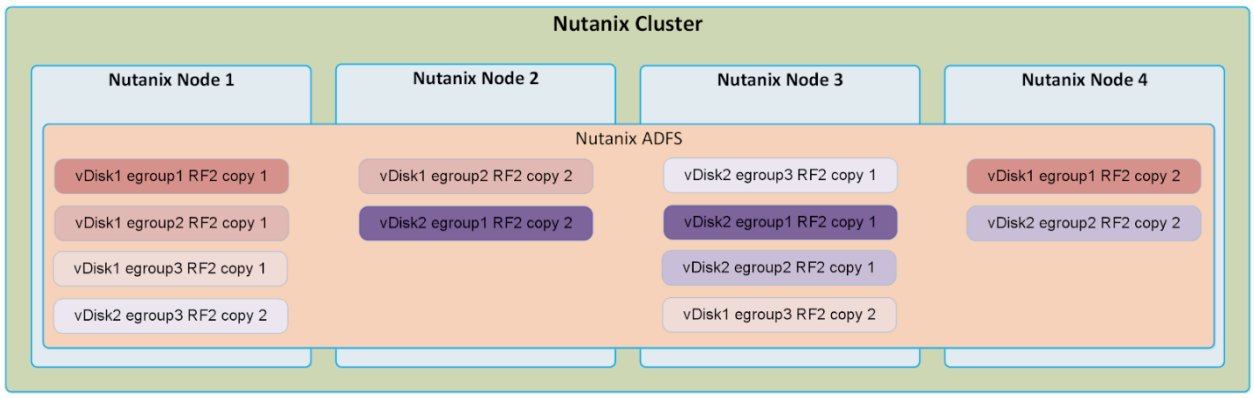


Figure 1: Four-Node Nutanix Cluster Before EC-X

Four-Node Cluster After EC-X

After the EC-X parity calculation, we get the following results:

- vDisk1 egroup1 goes on node1 (data) and node2 (parity).

- vDisk1 egroup2 goes on node1 (data) and node4 (parity).
- vDisk1 egroup3 goes on node1 (data) and node4 (parity).
- vDisk2 egroup1 goes on node3 (data) and node2 (parity).
- vDisk2 egroup2 goes on node3 (data) and node4 (parity).
- vDisk2 egroup3 goes on node3 (data) and node4 (parity).

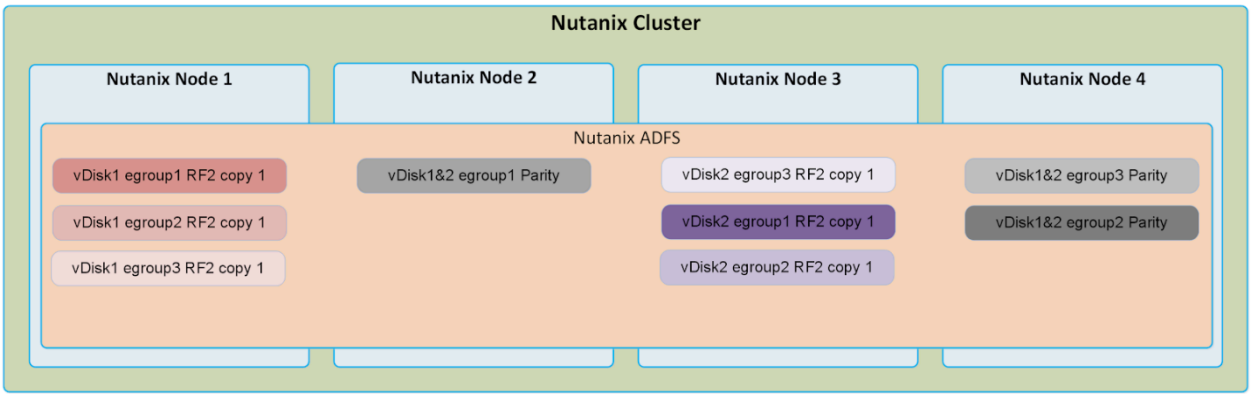


Figure 2: Four-Node Nutanix Cluster After EC-X

Six-Node Nutanix Cluster, Replication Factor 3

This scenario includes the following items:

- Six Nutanix nodes
- Two vDisks
  - The CVM on node1 owns vDisk1.
  - The CVM on node3 owns vDisk2.
- Three extent groups per vDisk
- Three identical extent group copies

Six-Node Cluster Before EC-X

Before the EC-X parity calculation, AOS writes each extent group three times to the Nutanix cluster on different nodes:

- vDisk1 egroup1 goes on nodes 1, 2, and 5.
- vDisk1 egroup2 goes on nodes 1, 3, and 6.
- vDisk1 egroup3 goes on nodes 1, 2, and 4.
- vDisk2 egroup1 goes on nodes 1, 3, and 4.
- vDisk2 egroup2 goes on nodes 2, 3, and 5.
- vDisk2 egroup3 goes on nodes 3, 4, and 6.

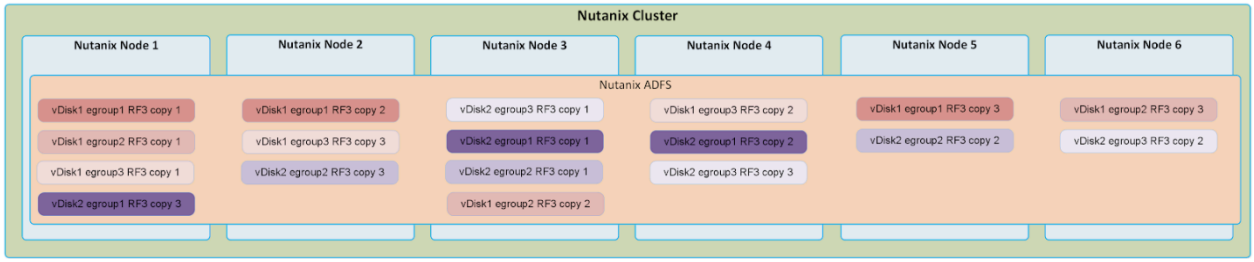


Figure 3: Six-Node Nutanix Cluster Before EC-X

Six-Node Cluster After EC-X

After the EC-X parity calculation, we get the following results:

- vDisk1 egroup1 goes on node1 (data) and nodes 2 and 4 (parity).
- vDisk1 egroup2 goes on node1 (data) and nodes 5 and 6 (parity).
- vDisk1 egroup3 goes on node1 (data) and nodes 2 and 4 (parity).
- vDisk2 egroup1 goes on node3 (data) and nodes 2 and 4 (parity).
- vDisk2 egroup2 goes on node3 (data) and nodes 5 and 6 (parity).
- vDisk2 egroup3 goes on node3 (data) and nodes 2 and 4 (parity).

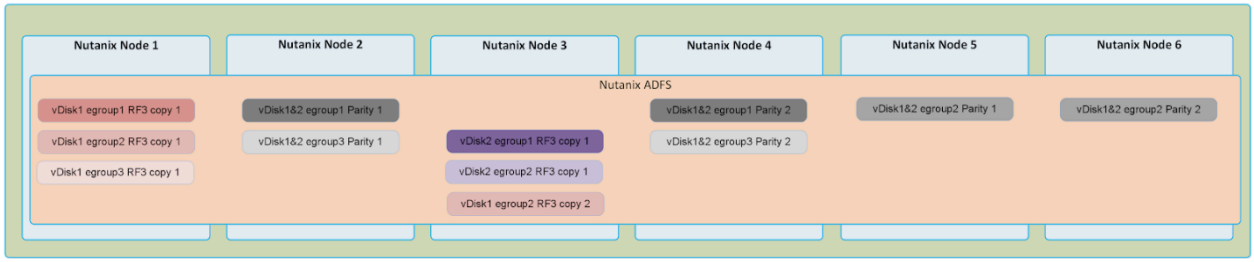


Figure 4: Six-Node Nutanix Cluster After EC-X

5. Recommendations

- Write one, read many (WORM) workloads and other workloads with similarly limited overwrites are the ideal candidates for EC-X. Examples include:
  - Backups
  - Archives
  - Nutanix Files and file servers
  - Log servers
  - Email (depending on usage)
  - Nutanix Objects
- A cluster must have at least four nodes with replication factor 2 or six nodes with replication factor 3 to enable EC-X.
- Erasure coding mandates that you put data and parity strips on separate failure domains (nodes) and have an additional node available for recovery. For example, a strip size of 4/1 requires you to have at least six nodes.
- With post-process EC-X, space savings might not appear for some time.
- Erasure coding effectiveness (data reduction savings) might be reduced on workloads that have many overwrites outside the erasure coding window, which by default is seven days.
- Read performance can degrade during failure scenarios.
- If you enable EC-X on any storage container, you need at least four blocks or racks for replication factor 2 or six blocks or racks for replication factor 3 to maintain block or rack awareness.
- Turning off EC-X recreates the data in a replication factor 2 or replication factor 3 configuration, so your environment returns to using more storage.

- Small EC-X overwrites (equal to or less than 512 KB) happen in-place in existing extent groups. Large overwrites happen in a replication factor 2 or replication factor 3 format, depending on the container configuration, then the write-cold timer starts over.

About Nutanix

Nutanix is a global leader in cloud software and a pioneer in hyperconverged infrastructure solutions, making clouds invisible and freeing customers to focus on their business outcomes. Organizations around the world use Nutanix software to leverage a single platform to manage any app at any location for their hybrid multicloud environments. Learn more at www.nutanix.com or follow us on Twitter @nutanix.

List of Figures