How objects are stored (replication or erasure coding)

StorageGRID 11.7

NetApp
April 12, 2024

Table of Contents

How objects are stored (replication or erasure coding) . . . 1
- What is replication? . . . 1
- Why you should not use single-copy replication . . . 2
- What is erasure coding? . . . 4
- What are erasure coding schemes? . . . 6
- Advantages, disadvantages, and requirements for erasure coding . . . 8

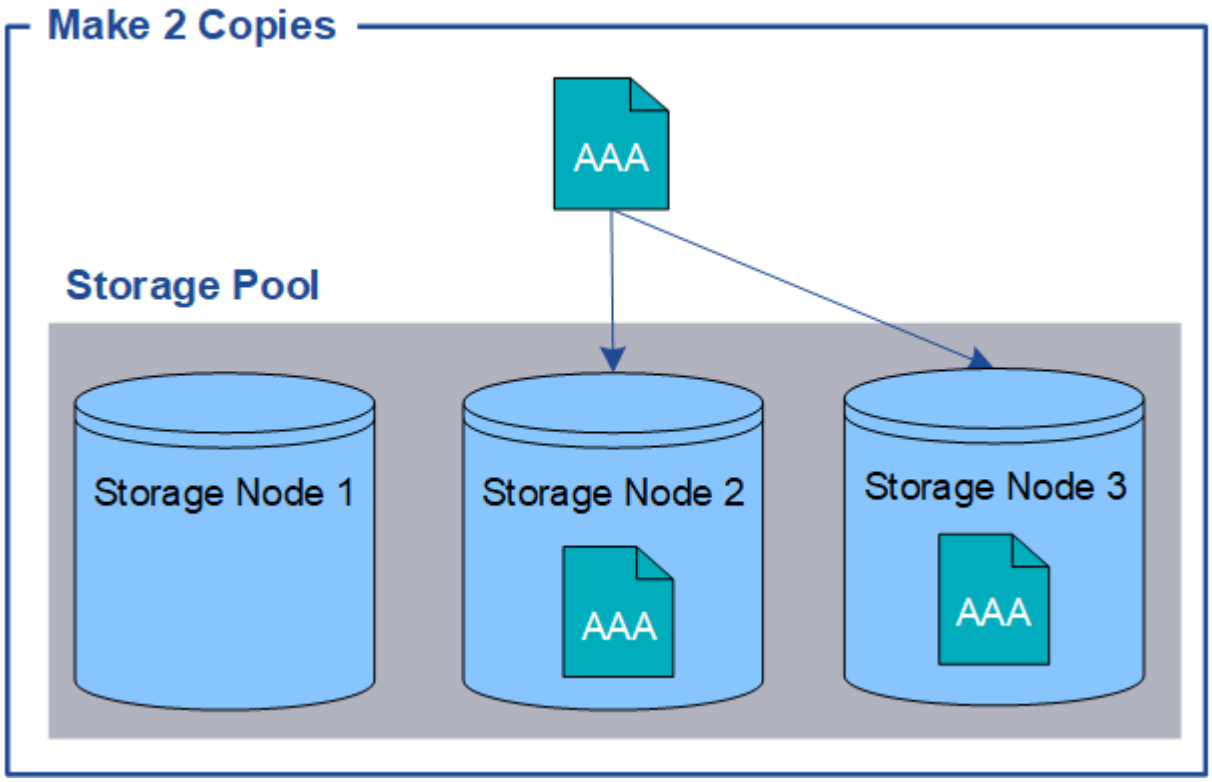
How objects are stored (replication or erasure coding)

What is replication?

Replication is one of two methods used by StorageGRID to store object data. When objects match an ILM rule that uses replication, the system creates exact copies of object data and stores the copies on Storage Nodes or Archive Nodes.

When you configure an ILM rule to create replicated copies, you specify how many copies should be created, where those copies should be placed, and how long the copies should be stored at each location.

In the following example, the ILM rule specifies that two replicated copies of each object be placed in a storage pool that contains three Storage Nodes.



When StorageGRID matches objects to this rule, it creates two copies of the object, placing each copy on a different Storage Node in the storage pool. The two copies might be placed on any two of the three available Storage Nodes. In this case, the rule placed object copies on Storage Nodes 2 and 3. Because there are two copies, the object can be retrieved if any of the nodes in the storage pool fails.

StorageGRID can store only one replicated copy of an object on any given Storage Node. If your grid includes three Storage Nodes and you create a 4-copy ILM rule, only three copies will be made—one copy for each Storage Node. The **ILM placement unachievable** alert is triggered to indicate that the ILM rule could not be completely applied.

Related information

- What is erasure coding?

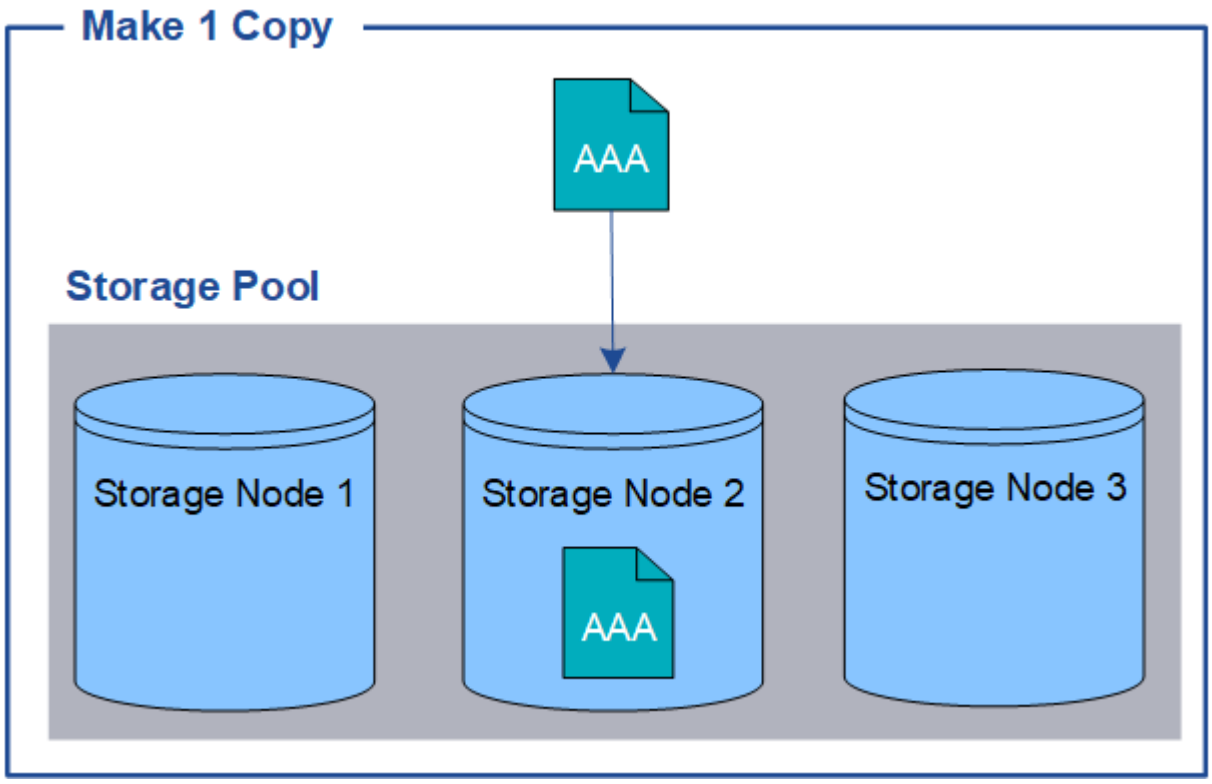
- What is a storage pool?
- Enable site-loss protection using replication and erasure coding

Why you should not use single-copy replication

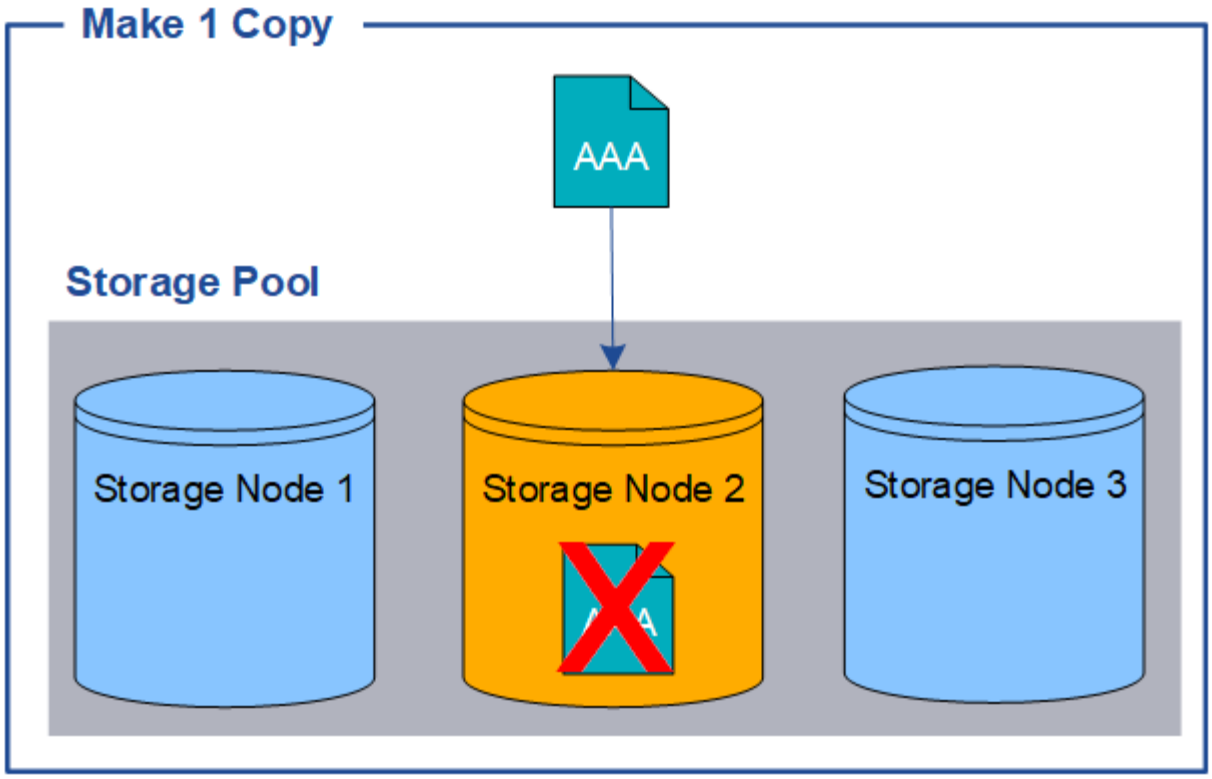
When creating an ILM rule to create replicated copies, you should always specify at least two copies for any time period in the placement instructions.

> Don’t use an ILM rule that creates only one replicated copy for any time period. If only one replicated copy of an object exists, that object is lost if a Storage Node fails or has a significant error. You also temporarily lose access to the object during maintenance procedures such as upgrades.

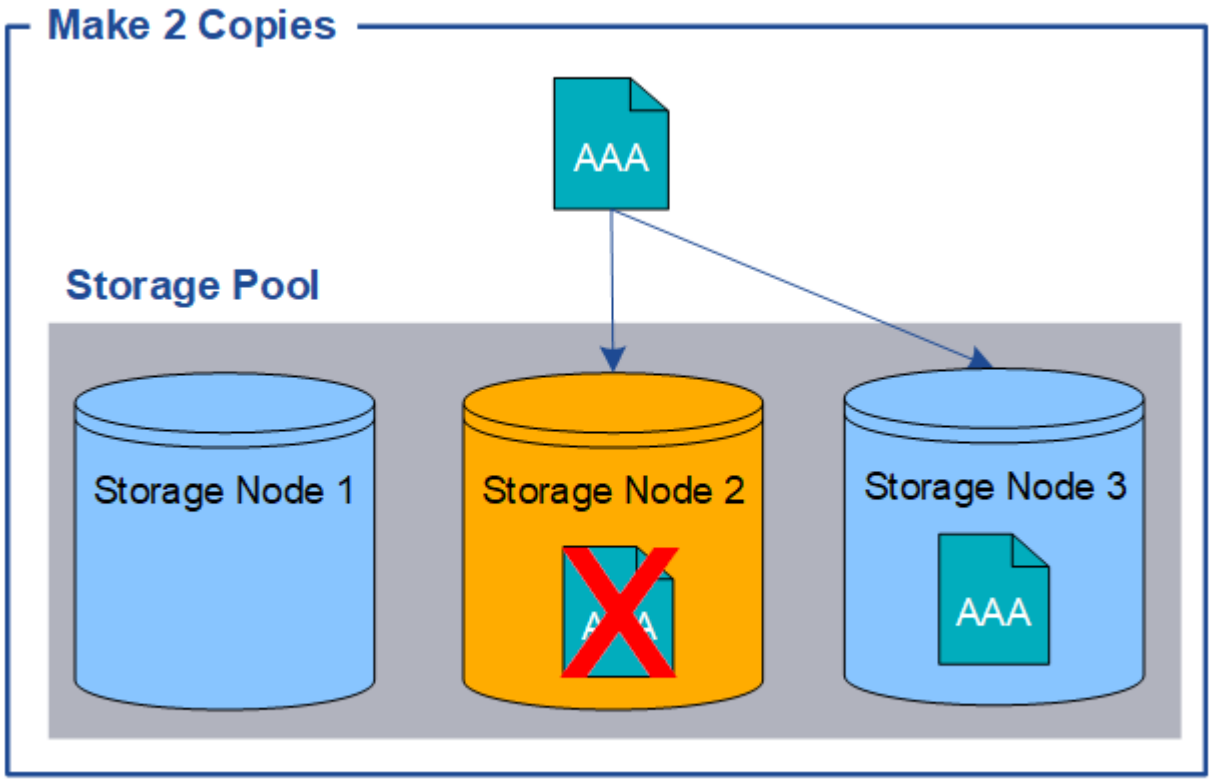
In the following example, the Make 1 Copy ILM rule specifies that one replicated copy of an object be placed in a storage pool that contains three Storage Nodes. When an object is ingested that matches this rule, StorageGRID places a single copy on only one Storage Node.



When an ILM rule creates only one replicated copy of an object, the object becomes inaccessible when the Storage Node is unavailable. In this example, you will temporarily lose access to object AAA whenever Storage Node 2 is offline, such as during an upgrade or other maintenance procedure. You will lose object AAA entirely if Storage Node 2 fails.



To avoid losing object data, you should always make at least two copies of all objects you want to protect with replication. If two or more copies exist, you can still access the object if one Storage Node fails or goes offline.



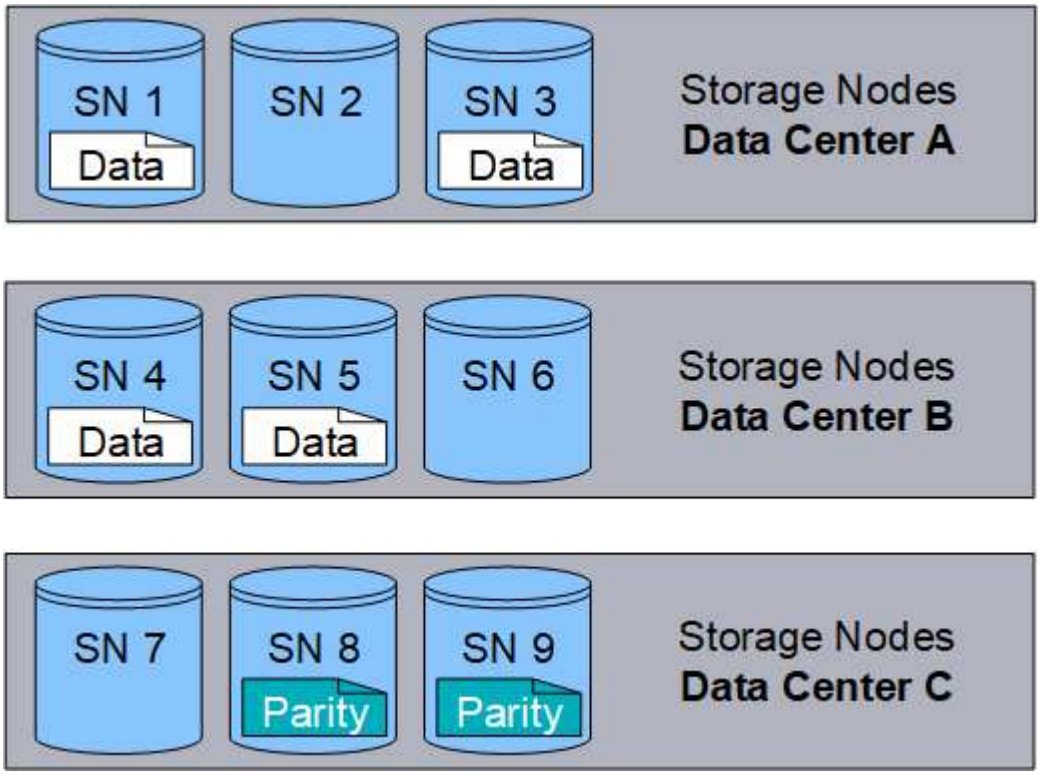
What is erasure coding?

Erasure coding is one of two methods StorageGRID uses to store object data. When objects match an ILM rule that uses erasure coding, those objects are sliced into data fragments, additional parity fragments are computed, and each fragment is stored on a different Storage Node.

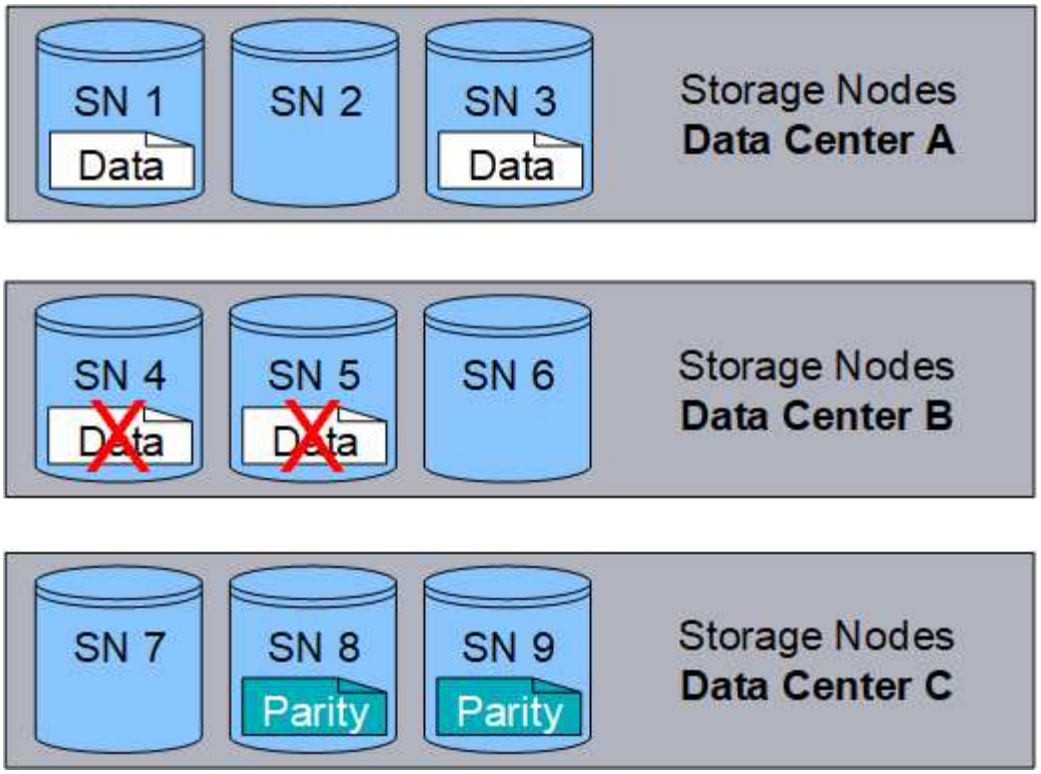
When an object is accessed, it is reassembled using the stored fragments. If a data or a parity fragment becomes corrupt or lost, the erasure-coding algorithm can recreate that fragment using a subset of the remaining data and parity fragments.

As you create ILM rules, StorageGRID creates erasure coding profiles that support those rules. You can view a list of erasure coding profiles, rename an erasure coding profile, or deactivate an erasure coding profile if it is not currently used in any ILM rules.

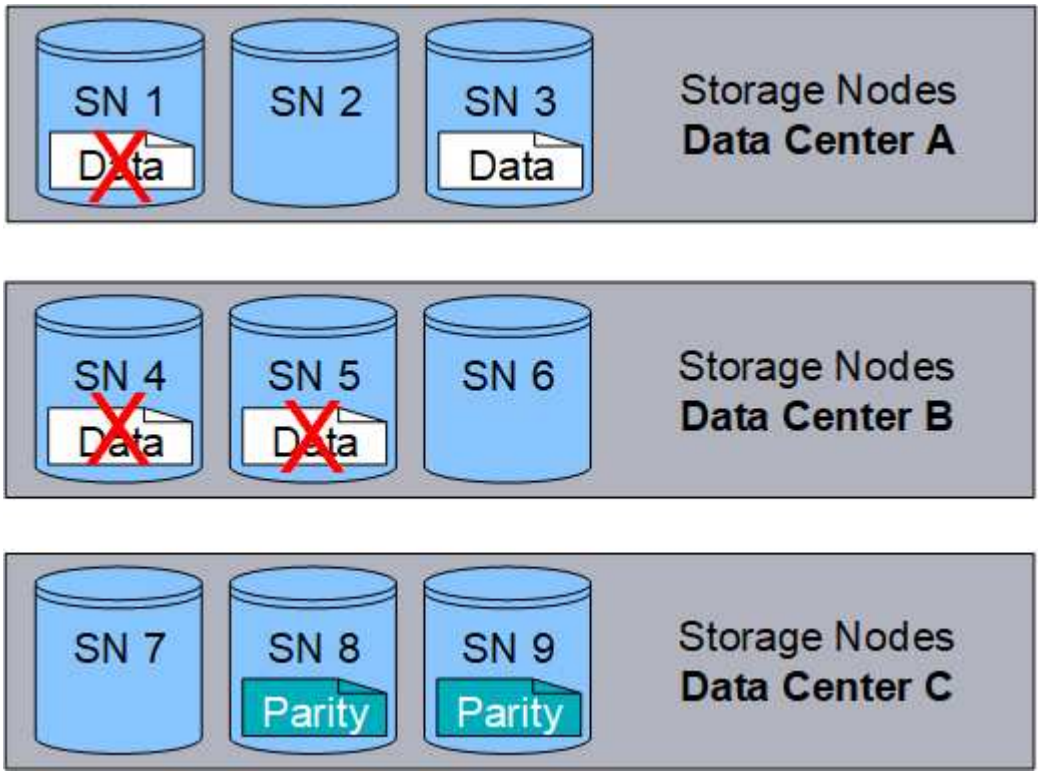
The following example illustrates the use of an erasure-coding algorithm on an object's data. In this example, the ILM rule uses a 4+2 erasure-coding scheme. Each object is sliced into four equal data fragments, and two parity fragments are computed from the object data. Each of the six fragments is stored on a different node across three data center sites to provide data protection for node failures or site loss.



The 4+2 erasure-coding scheme can be configured in various ways. For example, you can configure a single-site storage pool that contains six Storage Nodes. For site-loss protection, you can use a storage pool containing three sites with three Storage Nodes at each site. An object can be retrieved as long as any four of the six fragments (data or parity) remain available. Up to two fragments can be lost without loss of the object data. If an entire site is lost, the object can still be retrieved or repaired, as long as all of the other fragments remain accessible.



If more than two Storage Nodes are lost, the object is not retrievable.



Related information

- What is replication?
- What is a storage pool?
- What are erasure coding schemes?

- Rename an erasure coding profile
- Deactivate an erasure coding profile

What are erasure coding schemes?

Erasure-coding schemes control how many data fragments and how many parity fragments are created for each object.

When you configure the erasure coding profile for an ILM rule, you select an available erasure-coding scheme based on how many Storage Nodes and sites make up the storage pool you plan to use.

The StorageGRID system uses the Reed-Solomon erasure-coding algorithm. The algorithm slices an object into `k` data fragments and computes `m` parity fragments. The `k + m = n` fragments are spread across `n` Storage Nodes to provide data protection. An object can sustain up to `m` lost or corrupt fragments. To retrieve or repair an object, `k` fragments are needed.

When selecting the storage pool to use for a rule that will create an erasure-coded copy, use the following guidelines for storage pools:

- The storage pool must include three or more sites, or exactly one site.

  > You can't use erasure coding if the storage pool includes two sites.

  - Erasure-coding schemes for storage pools containing three or more sites
  - Erasure-coding schemes for one-site storage pools
- Don't use a storage pool that includes the default site, All Sites.
- The storage pool should include at least *k+m* `+1` Storage Nodes.

  The minimum number of Storage Nodes required is `k+m`. However, having at least one additional Storage Node can help prevent ingest failures or ILM backlogs if a required Storage Node is temporarily unavailable.

The storage overhead of an erasure-coding scheme is calculated by dividing the number of parity fragments (`m`) by the number of data fragments (`k`). You can use the storage overhead to calculate how much disk space each erasure-coded object requires:

*disk space = object size + (object size * storage overhead)*

For example, if you store a 10 MB object using the 4+2 scheme (which has 50% storage overhead), the object consumes 15 MB of grid storage. If you store the same 10 MB object using the 6+2 scheme (which has 33% storage overhead), the object consumes approximately 13.3 MB.

Select the erasure-coding scheme with the lowest total value of `k+m` that meets your needs. Erasure-coding schemes with a lower number of fragments are overall more computationally efficient, as fewer fragments are created and distributed (or retrieved) per object, can show better performance due to the larger fragment size, and can require fewer nodes be added in an expansion when more storage is required. (For information about planning a storage expansion, see the instructions for expanding StorageGRID.)

Erasure-coding schemes for storage pools containing three or more sites

The following table describes the erasure-coding schemes currently supported by StorageGRID for storage pools that include three or more sites. All of these schemes provide site-loss protection. One site can be lost, and the object will still be accessible.

For erasure-coding schemes that provide site-loss protection, the recommended number of Storage Nodes in the storage pool exceeds `k+m +1` because each site requires a minimum of three Storage Nodes.

Erasure-coding scheme (*k+m*)	Minimum number of deployed sites	Recommended number of Storage Nodes at each site	Total recommended number of Storage Nodes	Site loss protection?	Storage overhead
4+2	3	3	9	Yes	50%
6+2	4	3	12	Yes	33%
8+2	5	3	15	Yes	25%
6+3	3	4	12	Yes	50%
9+3	4	4	16	Yes	33%
2+1	3	3	9	Yes	50%
4+1	5	3	15	Yes	25%
6+1	7	3	21	Yes	17%
7+5	3	5	15	Yes	71%

> StorageGRID requires a minimum of three Storage Nodes per site. To use the 7+5 scheme, each site requires a minimum of four Storage Nodes. Using five Storage Nodes per site is recommended.

When selecting an erasure-coding scheme that provides site protection, balance the relative importance of the following factors:

- **Number of fragments**: Performance and expansion flexibility are generally better when the total number of fragments is lower.
- **Fault tolerance**: Fault tolerance is increased by having more parity segments (that is, when `m` has a higher value.)
- **Network traffic**: When recovering from failures, using a scheme with more fragments (that is, a higher total for `k+m`) creates more network traffic.
- **Storage overhead**: Schemes with higher overhead require more storage space per object.

For example, when deciding between a 4+2 scheme and 6+3 scheme (which both have 50% storage

overhead), select the 6+3 scheme if additional fault tolerance is required. Select the 4+2 scheme if network resources are constrained. If all other factors are equal, select 4+2 because it has a lower total number of fragments.

If you are unsure of which scheme to use, select 4+2 or 6+3, or contact technical support.

Erasure-coding schemes for one-site storage pools

A one-site storage pool supports all of the erasure-coding schemes defined for three or more sites, provided that the site has enough Storage Nodes.

The minimum number of Storage Nodes required is `k+m`, but a storage pool with `k+m +1` Storage Nodes is recommended. For example, the 2+1 erasure-coding scheme requires a storage pool with a minimum of three Storage Nodes, but four Storage Nodes is recommended.

Erasure-coding scheme (*k+m*)	Minimum number of Storage Nodes	Recommended number of Storage Nodes	Storage overhead
4+2	6	7	50%
6+2	8	9	33%
8+2	10	11	25%
6+3	9	10	50%
9+3	12	13	33%
2+1	3	4	50%
4+1	5	6	25%
6+1	7	8	17%
7+5	12	13	71%

Advantages, disadvantages, and requirements for erasure coding

Before deciding whether to use replication or erasure coding to protect object data from loss, you should understand the advantages, disadvantages, and the requirements for erasure coding.

Advantages of erasure coding

When compared to replication, erasure coding offers improved reliability, availability, and storage efficiency.

- **Reliability**: Reliability is gauged in terms of fault tolerance—that is, the number of simultaneous failures that can be sustained without loss of data. With replication, multiple identical copies are stored on different nodes and across sites. With erasure coding, an object is encoded into data and parity fragments and distributed across many nodes and sites. This dispersal provides both site and node failure protection. When compared to replication, erasure coding provides improved reliability at comparable storage costs.
- **Availability**: Availability can be defined as the ability to retrieve objects if Storage Nodes fail or become inaccessible. When compared to replication, erasure coding provides increased availability at comparable storage costs.
- **Storage efficiency**: For similar levels of availability and reliability, objects protected through erasure coding consume less disk space than the same objects would if protected through replication. For example, a 10 MB object that is replicated to two sites consumes 20 MB of disk space (two copies), while an object that is erasure coded across three sites with a 6+3 erasure-coding scheme only consumes 15 MB of disk space.

> ⓘ Disk space for erasure-coded objects is calculated as the object size plus the storage overhead. The storage overhead percentage is the number of parity fragments divided by the number of data fragments.

Disadvantages of erasure coding

When compared to replication, erasure coding has the following disadvantages:

- An increased number of Storage Nodes and sites is recommended, depending on the erasure coding scheme. In contrast, if you replicate object data, you need only one Storage Node for each copy. See Erasure coding schemes for storage pools containing three or more sites and Erasure coding schemes for one-site storage pools.
- Increased cost and complexity of storage expansions. To expand a deployment that uses replication, you add storage capacity in every location where object copies are made. To expand a deployment that uses erasure coding, you must consider both the erasure-coding scheme in use and how full existing Storage Nodes are. For example, if you wait until existing nodes are 100% full, you must add at least `k+m` Storage Nodes, but if you expand when existing nodes are 70% full, you can add two nodes per site and still maximize usable storage capacity. For more information, see Add storage capacity for erasure-coded objects.
- There are increased retrieval latencies when you use erasure coding across geographically distributed sites. The object fragments for an object that is erasure coded and distributed across remote sites take longer to retrieve over WAN connections than an object that is replicated and available locally (the same site to which the client connects).
- When you use erasure coding across geographically distributed sites, there is higher WAN network traffic usage for retrievals and repairs, especially for frequently retrieved objects or for object repairs over WAN network connections.
- When you use erasure coding across sites, the maximum object throughput declines sharply as network latency between sites increases. This decrease is due to the corresponding decrease in TCP network throughput, which affects how quickly the StorageGRID system can store and retrieve object fragments.
- Higher usage of compute resources.

When to use erasure coding

Erasure coding is best suited for the following requirements:

- Objects greater than 1 MB in size.

> Erasure coding is best suited for objects greater than 1 MB. Don’t use erasure coding for objects smaller than 200 KB to avoid the overhead of managing very small erasure-coded fragments.

- Long-term or cold storage for infrequently retrieved content.
- High data availability and reliability.
- Protection against complete site and node failures.
- Storage efficiency.
- Single-site deployments that require efficient data protection with only a single erasure-coded copy rather than multiple replicated copies.
- Multiple-site deployments where the inter-site latency is less than 100 ms.

Copyright information

Trademark information