IBM Cloud Object Storage System™
Version 3.14.3

Correlated Failure Mitigation Guide

This edition applies to IBM Cloud Object Storage System and is valid until replaced by new editions.

Contents

Chapter 1. Independent versus correlated failures

Usually, failures are considered to be randomly occurring events. It is true even outside the domain of storage systems.

Independent system failures

Consider that a car has some probability of breaking down during its commute. Often, it is also assumed that failures are statistically independent events. It means the engine that breaks down in one car does not increase the likelihood that another car on the road breaks down.

A correlated failure is an event where this assumption of statistical independence does not hold.

Correlated System Failures

If the first car failed due to defective motor, you can no longer assume that the failure is independent, as some chance exists that other cars on the road were also given the same bad oil. It might cause many vehicles to fail together, within a short period, and due to the same underlying reason. These failures are no longer independent, but correlated.

Correlated failures can cause a storage system, which can have a high degree of redundancy or replication, to nonetheless fail.

In the storage industry, the failure of one hard disk drive, if statistically independent, does not increase the failure chance for other drives.

Chapter 2. How correlated failures occur in components

Correlated failures occur whenever an event or situation causes failures across a population of drives, systems, or data centers.

Drives

Correlated drive failures (CDFs) occur either when an event causes many drives in physical proximity to fail or when a manufacturing defect in the same class, manufacturer, and type of drives occurs.

CDFs can occur due to any of the following conditions:

- Drive controller failure
- System downtime or restart
- Manufacturing defect in batch of drives
- Strong vibration in a rack or site
- Power surge or lightning strike
- Faulty memory or software on a server
- Logical error in operating system or file system

These failures can cause a group of drives to either become temporarily unavailable or experience unrecoverable failures. Storage systems that house all drives within the same server, rack, or site are susceptible to correlated drive failures. It might take only one event to trigger enough drive failures to exceed the system's failure tolerance. Within a system, related slices for the same data are always kept on both different drives and different servers. Therefore, a correlated failure that impacts all the drives within a server cannot cause data unavailability or loss.

Systems

Correlated system failures occur when an event causes many systems in physical proximity to fail.

Servers that are part of the same storage system are often kept in close physical proximity: within the same rack in the same site. It can lead to correlated unavailability, unreachability, or destruction. Events causing correlated system failures include the following items:

- Environmental (heating or cooling) failures
- Network connectivity loss in data center
- Switch or router failure in a rack
- Power supply failure in a rack
- Site destruction (fire, earthquake, flood)

While system deployments can exist entirely within one site or at one physical site, it increases the chance of unavailability or data loss, since a network failure at a site could make the data unavailable. Multiple power outages within a site can also cause loss of recently written data. Methods for mitigating these risks through an appropriate system design are discussed in Chapter 3, “Avoiding and mitigating failures,” on page 5.

In a system using Concentrated Dispersal, the failure of a single device may have greater impact because a device may contain more than one slice.

Data Centers

More rarely, a correlated failure might spread its effects across sites.

It means that data centers in different geographic locations can be impacted by the same event, causing data loss or unavailability. Examples of such correlated failures include the following items:

- Power grid failures
- Internet or backbone failures (cut cables, peering disputes)
- Global internet worm
- Organizational failures (insolvency, dissolution)
- Social unrest (wars, riots, terrorism)

These events occur infrequently, but as they occur they might affect the availability of multiple sites across a geographic area. While it is not possible to defend against every contingency, the correlation between site failures can be greatly reduced by appropriate selection of your sites.

Tip: Use sites on different power grids, and use different internet backbones and providers at each site.

Chapter 3. Avoiding and mitigating failures

Considerations

Using an IBM Cloud Object Storage System™, no data loss or availability occurs during correlated drive failures that are within and limited to a single node.

System considerations

For any Object stored in the system, at most one slice of that Object exists within one node.

When it comes to correlated failures that can affect a System, correlated system failures should be considered.

Site considerations

If all System nodes are deployed at the same physical site, then events that affect any nodes at that site affect the whole system.

Single-site deployments are used often because all data is accessed at that site.

In this case, outages that affect the system affect the reader and writer of the data as well. It decreases the impact of correlated failures that affect only availability. Correlated failures that impact the survivability of data are no less critical.

Important: Choose a site that is not subject to frequent natural disasters. Deploy an adequate fire suppression system.

Data loss is most likely due to a power outage that affects an entire rack or site. It can cause a loss of data that is written within the past minute, which is not yet saved to the drive. To avoid data loss from outages; conditioned, surge protected, and battery-backed up data is a must.

Alternatively, to gain the most in terms of resiliency and reliability out of a System, geographically disperse the Slicestor nodes across different data centers.

Fault tolerance

A system can be affected by outages of all the appliances at the same site.

With the right configuration, the System can tolerate such outages without losing availability at all.

To guarantee write availability when a site outage occurs, the fraction of nodes at any one site should not exceed the following equation:

```
( IDAWidth - WriteThreshold ) / IDAWidth
```

To guarantee read availability, the proportion of nodes at a site should be less than this equation:

```
( WriteThreshold - IDAThreshold ) / IDAWidth
```

By following these guidelines, even entire site failures do not impact access to data. It leaves site correlated failures as the only concern. To mitigate the risk or impact of correlated site failures, several things can be attempted. You can select sites that get power from different locations or live on different power grids. Sites should have multi-homed connections, to different internet backbones. The geographic

distance that separates sites should also be sufficient to avoid issues that could relate to the same natural disaster. By following these steps, the chance of your data being affected by a correlated failure can be substantially reduced.

The use of Vault Profiles represents a departure from the pre-Concentrated Dispersal method for configuring vaults. Due to the various subtleties of Concentrated Dispersal systems, users no longer directly configure IDA Width, Threshold, and Write Threshold. Instead, the IDA configuration is selected based on the number of Slicestors, the type, the number of sites, mirror settings, and the user selected Vault Optimization.

Limitations

Under Concentrated Dispersal, the unit of failure that causes loss or unavailability of a slice is moved from the Server to the Drive. That is, in a non-Concentrated Dispersal system, the fault tolerance implied the system could tolerate the loss and total destruction of (Width - Threshold) Slicestor devices. Under a Concentrated Dispersal system, the fault tolerance implies that the system can tolerate the loss and total destruction of (Width - Threshold) drives. This ensures reliability remains high, as drive failure is the most important determinate of data loss, and because data loss requires just as many drive failures as a non-Concentrated Dispersal system with the same IDA Configuration, a similar level of reliability is achieved.

However, under Concentrated Dispersal, the tolerance for outages of Storage nodes is not as high as the tolerance for disk failure. Therefore, the availability of a Concentrated Dispersal system with the same IDA Configuration is lower than the availability of a non-Concentrated Dispersal system

These issues are mitigated by the selection of the IDA configurations used in Concentrated Dispersal systems. The preconfigured IDA configurations are all selected to provide high levels of availability and reliability. Each IDA configuration is selected such that the system can tolerate the complete loss or destruction of any one Slicestor® Device, plus any other drive in the system.

Notices

This information was developed for products and services offered in the US. This material might be available from IBM® in other languages. However, you may be required to own a copy of the product or product version in that language in order to access it.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 10504-1785
U.S.A.

For license inquiries regarding double-byte character set (DBCS) information, contact the IBM Intellectual Property Department in your country or send inquiries, in writing, to:

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan, Ltd.
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokyo 103-8510, Japan

INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

IBM may use or distribute any of the information you provide in any way it believes appropriate without incurring any obligation to you.

Licensees of this program who wish to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact:

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
US

Such information may be available, subject to appropriate terms and conditions, including in some cases, payment of a fee.

The licensed program described in this document and all licensed material available for it are provided by IBM under terms of the IBM Customer Agreement, IBM International Program License Agreement or any equivalent agreement between us.

The performance data discussed herein is presented as derived under specific operating conditions. Actual results may vary.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

All IBM prices shown are IBM's suggested retail prices, are current and are subject to change without notice. Dealer prices may vary.

This information is for planning purposes only. The information herein is subject to change before the products described become available.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided "AS IS", without warranty of any kind. IBM shall not be liable for any damages arising out of your use of the sample programs.

If you are viewing this information softcopy, the photographs and color illustrations may not appear.

Trademarks

IBM, the IBM logo, and ibm.com® are trademarks or registered trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the web at Copyright and trademark information at www.ibm.com/legal/copytrade.shtml.

Accesser®, Cleversafe®, ClevOS™, Dispersed Storage®, dsNet®, IBM Cloud Object Storage Accesser®, IBM Cloud Object Storage Dedicated™, IBM Cloud Object Storage Insight™, IBM Cloud Object Storage Manager™, IBM Cloud Object Storage Slicestor®, IBM Cloud Object Storage Standard™, IBM Cloud Object Storage System™, IBM Cloud Object Storage Vault™, SecureSlice™, and Slicestor® are trademarks or registered trademarks of Cleversafe, an IBM Company and/or International Business Machines Corp.

Other product and service names might be trademarks of IBM or other companies.

Homologation statement

This product may not be certified in your country for connection by any means whatsoever to interfaces of public telecommunications networks. Further certification may be required by law prior to making any such connection. Contact an IBM representative or reseller for any questions.

Printed in USA