DELL EMC

EMC DATA DOMAIN DATA INVULNERABILITY ARCHITECTURE: ENHANCING DATA INTEGRITY AND RECOVERABILITY

A Detailed Review

ABSTRACT

No single mechanism is sufficient to ensure data integrity in a storage system. It is only through the cooperation of a multitude of mechanisms that establish successive lines of defense against all sources of errors that data recoverability can be assured. Unlike traditional general-purpose storage systems, EMC® Data Domain® deduplication storage systems have been designed explicitly as the storage of last resort. Data Domain systems put recovery above all else with data integrity protection built-in through the Data Domain Data Invulnerability Architecture. This white paper focuses on the four key elements of the Data Domain Data Invulnerability Architecture, which, in combination, provide the industry's highest levels of data integrity and recoverability:

- End-to-end verification
- Fault avoidance and containment
- Continuous fault detection and healing
- File system recoverability

August 2016

TABLE OF CONTENTS

EXECUTIVE SUMMARY 4
Storage system data integrity 4

INTRODUCTION 4
Audience 4

DATA DOMAIN DATA INVULNERABILITY ARCHITECTURE 5
End-to-end verification 5
Fault avoidance and containment 6
New data never overwrites good data 6
Fewer complex data structures 6
NVRAM for fast, safe restart 6
Persistent RAM protection 6
No partial stripe writes 6
Continuous fault detection and healing 7
RAID 6: Double disk failure protection, read error correction 7
On-the-fly error detection and correction 7
Scrub to insure data doesn't go bad 7
File system recoverability 8
Self-describing data format to ensure metadata recoverability 8
FS check, if needed, is fast 8

CONCLUSION 8

EXECUTIVE SUMMARY

STORAGE SYSTEM DATA INTEGRITY

Behind all their added value, specialized storage systems are built on software and general-purpose computing components that can all fail. Some failures have an immediate visible impact, such as the total failure of a disk drive. Other failures are subtle and hidden, such as a software bug that causes latent file system corruption that is only discovered at read time. To ensure data integrity in the face of such failures, the best storage systems include various data integrity checks and are generally optimized for performance and system availability, not data invulnerability. In the final analysis, they assume that backups get done, and make design tradeoffs that favor speed over guaranteed data recoverability. For example, no widely used primary storage file system reads data back from disk to ensure it was stored correctly; to do so would compromise performance. But data can't be considered invulnerable if it isn't stored correctly in the first place. With purpose-built backup appliances, the priority must be data invulnerability over performance and even availability. Unless the focus is on data integrity, backup and archive data is at risk. If data is at risk, then when the primary copy of the data is lost, recovery is at risk. Most purpose-built backup appliances are just primary storage systems built out of cheaper disks. As such, they inherit the design philosophy of their primary storage predecessors. Though labeled as purpose-built backup appliances, their designs emphasize performance at the expense of data invulnerability.

INTRODUCTION

This white paper focuses on the four key elements of the EMC® Data Domain® Data Invulnerability Architecture, which, in combination, provide the industry's highest levels of data integrity and recoverability.

AUDIENCE

This white paper is intended for EMC customers, technical consultants, partners, and members of the EMC and partner professional services community who are interested in learning more about the Data Domain Data Invulnerability Architecture.

DATA DOMAIN DATA INVULNERABILITY ARCHITECTURE

Data Domain deduplication storage systems represent a clean break from conventional storage system design thinking and introduce a radical premise: What if data integrity and recoverability was the most important goal? If one imagines a tapeless IT department, one would have to imagine extremely resilient and protective disk storage. Data Domain systems have been designed from the ground up to be the storage of last resort. The Data Domain operating system (DD OS) is purpose-built for data invulnerability. There are four critical areas of focus:

- End-to-end verification
- Fault avoidance and containment
- Continuous fault detection and healing
- File system recoverability

Even with this model, it is important to remember that DD OS is only as good as the data it receives. It can do an end-to-end test of the data it receives within its system boundaries, but DD OS cannot know whether that data has been protected along the network on the way to the system. If there is an error in the network that causes data corruption, or if the data is corrupted in place in primary storage, DD OS cannot repair it. It remains prudent to test recovery to the application level on a periodic basis.

END-TO-END VERIFICATION

Since every component of a storage system can introduce errors, an end-to-end test is the simplest path to ensure data integrity. End-to-end verification means reading data after it is written and comparing it to what was sent to disk, proving that it is reachable through the file system to disk, and proving the data has not been corrupted. When DD OS receives a write request from backup or archive software, it computes a checksum for the data. The system then stores unique data to disk and reads it back to validate the data, immediately correcting I/O errors. Since data is validated after writing to disk and before being released from memory/NVRAM, correcting I/O errors doesn't require restarting the backup job.

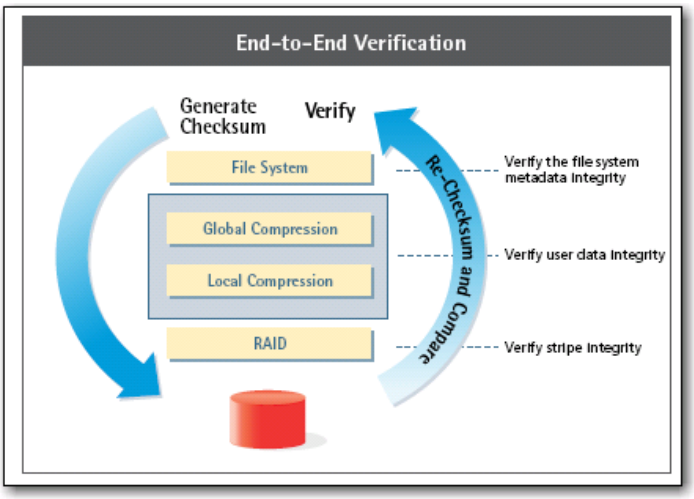

Figure 1. The end-to-end check verifies all file system data and metadata.

End-to-end verification confirms the data is correct and recoverable from every level of the system. If there are problems anywhere along the way, for example if a bit has flipped on a disk drive, it will be caught. Errors can also be corrected through self-healing as described below in the next section. Conventional, primary storage systems cannot afford such rigorous verifications. However, purpose-built backup appliances require them. The tremendous data reduction achieved by Data Domain Global Compression™ reduces the amount of data that needs to be verified and makes such verifications possible.

FAULT AVOIDANCE AND CONTAINMENT

The next step in protecting the data is to make sure the data, which was verified to be correct, stays correct. Ironically, the biggest risk to file system integrity is file system software errors when writing new data. It is only new writes that can accidentally write on existing data, and new updates to file system metadata that can mangle existing structures. Because the Data Domain file system was built to protect data as its primary goal, its design protects even against its own software errors that could put existing data at risk. It accomplishes this through a combination of design simplicity—which reduces the chance of bugs in the first place—and several fault containment features, which make it difficult for potential software errors to corrupt existing data. Data Domain systems are equipped with a specialized log-structured file system that has four important benefits.

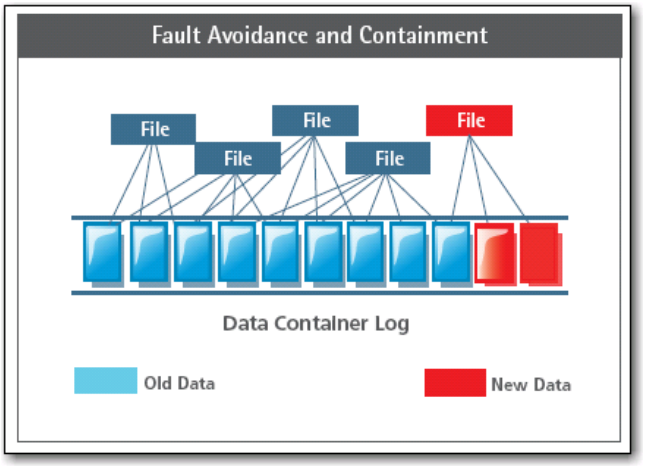

Figure 2. New data never puts old data at risk. The data container log never overwrites or updates existing data. New data is always written in new containers (in red). The old containers and references remain in place and are safe even in the face of software bugs or hardware faults that may occur when storing new backups.

NEW DATA NEVER OVERWRITES GOOD DATA

Unlike a traditional file system, which will often overwrite blocks when data changes, Data Domain systems only write to new blocks. This isolates any incorrect overwrite (a software bug type of problem) to only the newest backup and archive data. Older versions remain safe.

FEWER COMPLEX DATA STRUCTURES

In a traditional file system, there are many data structures (e.g., free block bit maps and reference counts) that support very fast block updates. In a backup application, the workload is primarily simpler sequential writes of new data, meaning fewer data structures are required to support it. As long as the system can keep track of the head of the log, new writes will not touch old data. This design simplicity greatly reduces the chances of software errors that could lead to data corruption.

NVRAM FOR FAST, SAFE RESTART

The system includes a non-volatile RAM write buffer into which it puts all data not yet safely on disk. The file system leverages the security of this write buffer to implement a fast, safe restart capability. The file system utilizes many internal logic and data structure integrity checks. If any problem is found by one of these checks, the file system restarts itself. The checks and restarts provide early detection and recovery from data corruption errors. As it restarts, the Data Domain file system verifies the integrity of the data in the NVRAM buffer before applying it to the file system, ensuring that no data is lost due to the restart. Because the NVRAM is a separate hardware device, it protects the data from errors that can corrupt data in RAM. Because the RAM is non-volatile, it also protects against power failures. Though the NVRAM is important for ensuring the success of new backups, the file system guarantees the integrity of old backups even if the NVRAM itself fails.

PERSISTENT RAM PROTECTION

In some smaller Data Domain systems, the entire system is protected from a power loss by a battery backup. For these systems, data in RAM is protected by a persistent RAM implementation (PRAM). Data Domain systems move data from RAM to disk in a process called vaulting. In the event of a power failure, a battery powers the entire Data Domain system while data is vaulted from RAM to disk. If the event of a system crash, the system reboots into a special vaulting mode where it vaults data from RAM to disk first and then reboots to normal DDOS. This persistent RAM feature maintains power and control to the system RAM through a crash so that the contents of RAM are preserved across the crash.

Persistent RAM implementations protect data during an AC power loss, during a system reboot and shutdown, and during a DDOS or kernel crash. The data vaulted to disk is RAID-protected just like the Data Domain file system.

NO PARTIAL STRIPE WRITES

Traditional primary storage disk arrays, whether RAID 1, RAID 3, RAID 4, RAID 5, or RAID 6, can lose old data if, during a write, there is a power failure that causes a disk to fail. This is because disk reconstruction depends on all the blocks in a RAID stripe being consistent, but during a block write there is a transition window where the stripe is inconsistent, so reconstruction of the stripe would

fail, and the old data on the failed disk would be lost. Enterprise storage systems protect against this with NVRAM or uninterruptible power supplies. But if these fail because of an extended power outage, the old data could be lost and a recovery attempt could fail. For this reason, Data Domain systems never update just one block in a stripe. Following the no-overwrite policy, all new writes go to new RAID stripes and those new RAID stripes are written in their entirety[1]. The verification after write ensures that the new stripe is consistent. New writes don't put existing data at risk. Data Domain systems are designed to minimize the number of standard storage system errors. If more challenging faults happen, it takes less time to find them, correct them, and notify the operator.

CONTINUOUS FAULT DETECTION AND HEALING

No matter the software safeguards in place, it is the nature of computing hardware to have occasional faults. Most visibly in a storage system, disk drives can fail. But other more localized or transient faults also occur. An individual disk block may be unreadable or there could be a bit flip on the storage interconnect or internal system bus. For this reason, DD OS builds in extra levels of data protection to detect faults and recover from them on-the fly and to ensure successful data restore operations.

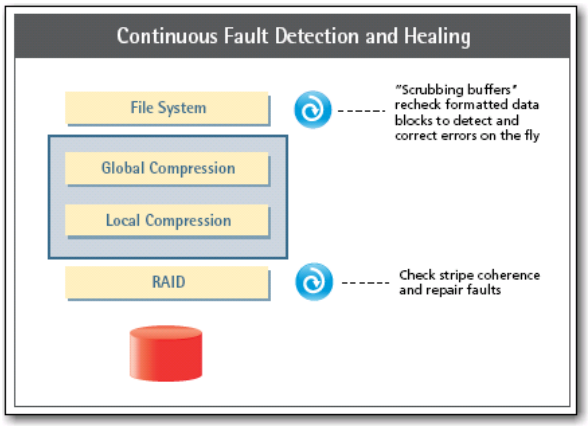

Figure 3. Continuous fault detection and healing protect against storage system faults. The system periodically re-checks the integrity of the RAID stripes and the container log and uses the redundancy of the RAID system to heal any faults. During every read, data integrity is re-verified and any errors are healed on the fly.

RAID 6: DOUBLE DISK FAILURE PROTECTION, READ ERROR CORRECTION

RAID 6 is the foundation for Data Domain's continuous fault detection and healing. Its powerful dual-parity architecture offers significant advantages over conventional architectures including RAID 1 (mirroring), RAID 3, RAID 4, or RAID 5 single-parity approaches. Raid 6:

- protects against two disk failures
- protects against disk read errors during reconstruction
- protects against the operator pulling the wrong disk
- guarantees RAID stripe consistency even during power failure without reliance on NVRAM or UPS
- verifies data integrity and stripe coherency after writes

Each shelf includes a global spare drive, which automatically replaces a failed drive anywhere in the Data Domain system. When the hot swappable failed drive is replaced by EMC, it becomes a new global spare. By comparison, once a single disk is down in the other RAID approaches, any further simultaneous disk error will cause data loss. Any storage system of last resort must include the extra level of protection that RAID 6 provides.

ON-THE-FLY ERROR DETECTION AND CORRECTION

To ensure that all data returned to the user during a restore is correct, the Data Domain file system stores all of its on-disk data structures in formatted data blocks. These are self-identifying and covered by a strong checksum. On every read from disk, the system first verifies that the block read from disk is the block expected. It then uses the checksum to verify the integrity of the data. If any issue is found, it asks RAID 6 to use its extra level of redundancy to correct the data error. Because the RAID stripes are never partially updated, their consistency is ensured, as is the ability to heal an error when it is discovered.

SCRUB TO INSURE DATA DOESN'T GO BAD

On-the-fly error detection works well for data that is being read, but it does not address issues with data that may be unread for weeks or months before it is needed for a recovery. For this reason, Data Domain systems actively re-verify the integrity of all data in an ongoing background process. This scrubbing process intelligently finds and repairs defects on the disk before they can become a problem. Through RAID 6, on-the-fly error detection and correction, and ongoing data scrubbing, most computing-system and disk drive-generated faults can be isolated and overcome with no impact on system operation or data risk.

[1] The gateway product, which relies on external RAID, is unable to guarantee that there are no partial stripe writes.

FILE SYSTEM RECOVERABILITY

Though every effort is made to ensure there are no file system issues, the Data Invulnerability Architecture anticipates that, being man-made, some system some time may have a problem. It therefore includes features to reconstruct lost or corrupted file system metadata and also file system check tools that can bring an ailing system safely back online quickly.

SELF-DESCRIBING DATA FORMAT TO ENSURE METADATA RECOVERABILITY

Metadata structures, such as indices that accelerate access, are rebuildable from the data on disk. All data is stored along with metadata that describes it. If a metadata structure is somehow corrupted, there are two levels of recoverability. First, a snapshot is kept of the file system metadata every several hours; recoverability can rely on this point-in-time copy. Second, the data can be scanned on disk and the metadata structure can be rebuilt. These capabilities enable recoverability even if there is a worst case corruption of the file system or its metadata.

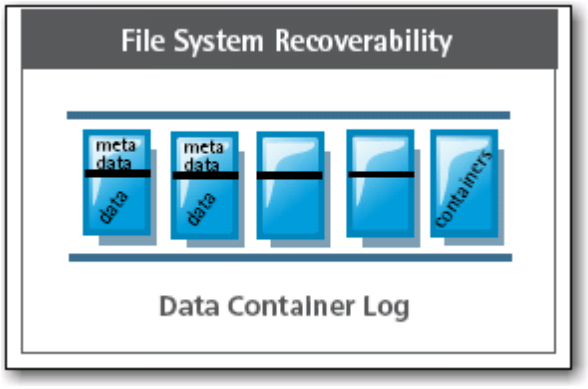

Figure 4. Data is written in a self-describing format. If necessary the file system can be re-created by scanning the log and rebuilding it from the metadata stored with the data.

FS CHECK, IF NEEDED, IS FAST

In a traditional file system, consistency is not checked online at all. Data Domain systems check through an initial inline verification to ensure consistency for all new writes. The usable size of a traditional file system is often limited by the time it would take to recover the file system in the event of some sort of corruption. Imagine running fsck on a traditional file system with more than 80 TB of data. The reason the checking process can take so long is that the file system needs to sort out where the free blocks are so that new writes don't end up overwriting existing data accidentally. Typically this entails checking all references to rebuild free block maps and reference counts. The more data in the system, the longer this takes. In contrast, since the Data Domain file system never overwrites old data and doesn't have block maps and reference counts to rebuild, it only has to verify where the head of the log is to safely bring the system back online to restore critical data.

CONCLUSION

No single mechanism is sufficient to ensure data integrity in a storage system. It is only through the cooperation of a multitude of mechanisms that establish successive lines of defense against all sources of errors that data recoverability can be assured.

Unlike a traditional storage system that has been repurposed from primary storage to data protection, Data Domain systems have been designed from the ground up explicitly as the data store of last resort. The innovative Data Invulnerability Architecture lays out the industry's best defense against data integrity issues. Advanced verification ensures that new backup and archive data is stored correctly. The no-overwrite, log-structured architecture of the Data Domain file system together with the insistence on full-stripe writes ensures that old data is always safe even in the face of software errors from new data. Meanwhile, a simple and robust implementation reduces the chance of software errors in the first place.

The above mechanisms protect against problems during the storage of backup and archive data, but faults in the storage itself also threaten data recoverability. For this reason, the Data Invulnerability Architecture includes a proprietary implementation of RAID 6 that protects against up to two disks failures, can rebuild a failed disk even if there is a data read error, and corrects errors on-the-fly during read. It also includes a background scrub process that actively seeks out and repairs latent faults before they become a problem.

The final line of defense is the recoverability features of the Data Domain file system. The self-describing data format enables the reconstruction of file data even if various metadata structures are corrupted or lost. And, the fast file system check and repair means that even a system holding dozens of terabytes of data won't be offline for long if there is some kind of problem.

Data Domain systems are the only solution built with this relentless attention to data integrity, giving you ultimate confidence in your recoverability.