Quantum

WHITE PAPER

ACTIVESCALE™ ERASURE CODING AND SELF-PROTECTING TECHNOLOGIES

Dynamic Data Placement (DDP) and Dynamic Data Repair (DDR) Technologies Within the ActiveScale Object Storage System Software

Contents

Introduction

Relentless data growth continues to pressure IT budgets and administrators to find more efficient and effective ways to store and protect petabyte-scale data. At the same time, organizations are looking to unlock the value in their data, which makes the task even more challenging. The right storage architecture can allow organizations to leverage more of their data without requiring budgets to scale at the same pace and make facilitating data-forever realistic. An effective storage solution to meet these needs would provide:

- Real-time access from anywhere in the world
- Protection from data loss
- Scaling without limits
- Easy to manage and maintain

ActiveScale is built on patented object storage technologies to address these needs.

Traditional Storage Limits

RAID has long been the data protection method of choice for traditional storage systems, yet it has reached its limits for today's petabyte-scale data stores. As hard drive capacities increase in response to data growth, data rebuild times have become unbearably long, increasing the exposure to data loss. Additionally, scaling system capacity requires reconfiguring RAID volumes and groups, which can quickly become unwieldy and in some cases require complex manual data migrations.

Enterprises are keeping data for longer periods of time for analytics and other purposes. As a result, we find that magnetic disks can suffer from bit errors and associated unrecoverable read errors. In traditional RAID-based systems, bit errors will only be discovered when data is read. Parity RAID schemes may not be able to correct errors through drive recovery before a second drive fails, which will lead to permanent data loss.

A fundamentally different approach that abstracts data management from the underlying hardware is needed. Separating the hardware with its frailties from the software provides a better way to address the limitations of today's traditional storage architectures. A software defined storage architecture that addresses data protection, scalability and ease of management limitations found in traditional RAID storage at scale is needed to meet the data demands of today's world.

ActiveScale Object Storage Software

ActiveScale object storage software was developed from the ground up for the highest levels of scalability, with high durability (19-nines) and simplicity on a cloud scale. The software is at the heart of all ActiveScale object storage systems.

Two key components of the software architecture are Dynamic Data Placement (DDP) Technology and Dynamic Data Repair (DDR) Technology. DDP executes the erasure coding algorithm and performs the hierarchical data spreading function for dynamic data placement. ActiveScale's data placement minimizes the impact of hardware or data integrity failures and capacity expansion. This means that traditional forklift upgrades and performance degradation due to cumulative hardware failures are now a thing of the past. DDR performs data integrity audits and automated repair functions to address "bit rot" or data degradation in the storage media.

ActiveScale object storage systems are designed with an extremely high level of durability, specified up to 19 nines (99.9999999999999999%), in support of the most demanding enterprise and service provider environments. This level of durability corresponds to an average annual expected loss of merely .00000000000000001% of stored objects.

The durability of an ActiveScale storage system is largely determined by the erasure coding. This is the algorithm that divides an object into chunks, a subset of which may be lost without causing object loss. Other key durability factors include the annual failure rate of the disk drives and object repair times.

To better understand how ActiveScale works we'll consider:

- Object storage encoding
- Dynamic data placement
- Data integrity
- Data consistency

Object Storage Encoding

DDP is the erasure encoding algorithm largely responsible for the high durability of ActiveScale object storage systems. An object to be written (PUT) is first broken into chunks. ActiveScale uses DDP to encode and place the data chunks.

Storage Policies

The software's storage policy controls DDP by specifying the configuration of two erasure encoding parameters: spread width and disk safety. The spread width parameter determines the number of disk drives the encoded data is spread across for a given stored object. Disk safety determines how many simultaneous drive losses can be tolerated without impacting the data's readability. In this way, data is protected in the event of a disk, storage node, system rack or even data center loss. The storage policy will consider all available sites for data placement and spread data in a way that optimizes durability, capacity efficiency and repair time.

System Resilience

For an example of data protection and system resilience, consider an object with an 18/5 policy (spread width 18 and disk safety 5) stored in a single data center rack containing 6 storage nodes and 98 drives per node. A selection of 18 drives is made by the system that equally balances the data across this hierarchy. With 6 storage nodes in the rack , it will randomly select 3 drives per node to store object chunks. See Figure 1.

In the 18/5 storage policy, the object is encoded into 18 chunks and can lose up to 5 chunks and still maintain object integrity. Any 13 chunks can be used to re-create the object. If a single drive fails in one of the storage nodes in the cluster, with each object spread across 18 drives, the safety level is reduced by 1 out of the 5 total. If a storage node fails and all 98 drives become unavailable, with each object spread across 3 drives in the storage node, the safety level for an object is reduced by 3 out of 5. All objects that had chunks on drives in the failed storage node still have chunks available on 15 other drives in the remaining 5 storage nodes. As a result, the disk safety for those objects on that storage node is reduced to 2. Therefore, in a full storage node failure, there are still 2 more drives that could fail and the data would still be readable by all users. This ability to absorb numerous, simultaneous drive failures without data loss is what gives the ActiveScale object storage system its extremely high level of durability.

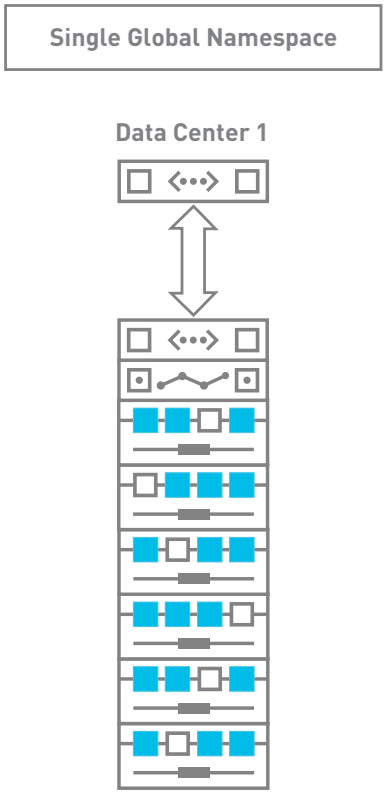

Figure 1: Data spread example for a single rack data center cluster

After the error is repaired, the drive safety level for all objects will be restored to the original 5 drives. This repair process is automated and happens without IT intervention. From an operations standpoint, drives that fail can remain in place. There is no need for IT to immediately replace a failed drive like they must do with RAID-based systems. Failed drives can be replaced during regular maintenance intervals or not at all (fail-in-place).

Large Object and Small Object Policies

Large and small object policies determine how writes (PUT) will be handled. For the large object policy, an incoming object is broken into chunks using Reed-Solomon encoding to take advantage of hardware acceleration in today's processors. This process provides high resiliency and data integrity with a choice of encoding based on customer requirements, such as 3-geo or single-geo configurations. For example, with 18/8 encoding the object is encoded into 18 chunks. The object can be retrieved with any 10 chunks, and up to 8 chunks can be lost or unavailable. This allows an entire site to be lost in a 3-geo configuration, plus two additional chunks in the remaining sites, and the objects would still be available.

The large object policy is optimized for durability and economics. The spread of the chunks will be determined with a three-tier hierarchy designed to minimize disk hot spots and rebalancing while optimizing efficiency to keep costs low. To read (GET) the object only the minimum number of chunks required will be used to reconstitute the object.

Alternatively, the ActiveScale system can employ a small object policy for object sizes 0-8MB. If the application makes use of a majority of, for example, 4MB objects, the system can be tuned for objects of that size. The small object policy is optimized for read performance by maintaining a full copy of the object on a single drive. When the small object policy is applied, DDP will automatically reduce the spread width and store sufficient data in one of the check block files to fully decode the object from a single drive. This is done to minimize latency when storing and retrieving small objects. To provide the reduced latency, a certain amount of overhead is added since a full copy and an encoded copy are written. Both large and small object policies maintain the systems specified durability regardless of object size.

A smaller spread width for small objects will result in writing data to fewer drives, which reduces the disk I/O operations consumed. For systems under high parallel load, this can allow the system to store more objects simultaneously and lower the aggregate latency of the write process.

For optimal system performance, a decision needs to be made regarding the appropriate balance of small and large objects. An ActiveScale system might combine a low percentage of small objects and a high percentage of large objects to deliver both the performance where needed and cost-effective capacity as required.

Metadata Protection for Single Data Center

DDP stores 3 copies of metadata across 3 controllers. A portion of the object's metadata is the spread, which is the list of drives that contain the encoded data for that object. The 3 system nodes form an active cluster for the metadata that supports full read/write operation even in case of a full system node failure. In addition to the 3 copies on 3 system nodes, the metadata is also stored with the encoded data on the storage nodes per the storage policy. This protects the metadata against a major disaster where all 3 system nodes become unavailable.

3-Geo Configurations

ActiveScale object storage systems can be deployed across three geographically dispersed data centers (3-Geo) to protect against a full data center outage. The storage policy for this configuration is 18/8, which results in objects spread across 18 disk drives with a disk safety of 8. Objects can be decoded from any subset of 10 encoded chunks.

Since a 3-Geo system be must able to recover objects if one site becomes unavailable, no more than a third of chunks can be in a single datacenter. With dynamic data placement's hierarchical spreading enabled for each level, the system will equally balance the chunks over 18 drives across three data centers. If each data center has a single rack, 6 drives will be selected per rack, and if there are 6 storage nodes in a rack, a single drive is selected in each storage node. This can be seen in Figure 2.

With only one drive used per node to store encoded chunks, the system is protected against 8 simultaneous storage node failures. Simultaneous means before DDR has had the chance to repair the data from the first failed storage node (covered in more detail later). This configuration protects against a full data center outage because an object's chunks have only been stored on 6 drives per data center. A full data center outage leaves the disk safety for all objects at 2.

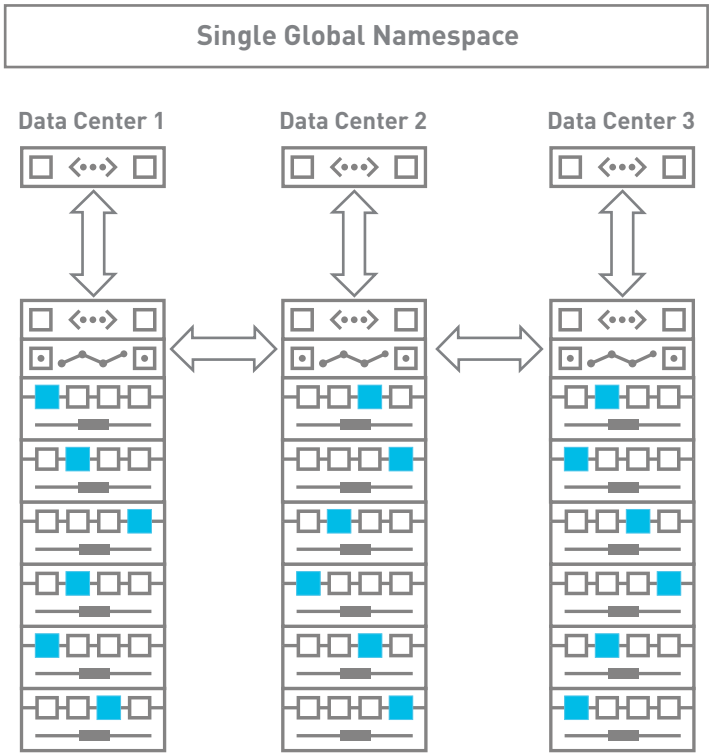

Figure 2: Data spread example for a three rack 3-Geo data center configuration

Metadata Protection for 3-Geo

To maintain object integrity from a metadata perspective, DDP stores 3 copies of metadata across 3 system nodes. In a 3-Geo configuration, copies of the metadata are spread across 3 different data center locations to tolerate a full site outage. The system remains fully read/write operational with 2 of the 3 copies of metadata.

Dynamic Data Placement

Spreads are generated dynamically at the time an object is being stored. DDP will make a selection of disk drives from all available drives in the system that meets the hierarchy rules in the storage policy. As this spread selection is performed for each object (and individually for each superblock of a large object), the encoded data chunks of objects are stored across all available drives within the storage system, always equally balanced in accordance to the current hierarchy. In the case of the small object policy, the 6 encoded chunks are spread out and the single full un-encoded object on a single drive makes the spread unbalanced.

Dynamic spreads are important to avoid disk hot spots and data rebalancing overhead when expanding capacity, which is typical of static or deterministic data placement commonly found in other systems. Static data placement can significantly impact performance by forcing a rebalancing when dealing with component failure or even the addition of new capacity. These problems are avoided with ActiveScale's dynamic data placement.

DDP Hierarchy

DDP technology keeps track of all available disk drives in the system and the location of the data on those drives. It establishes and maintains a hierarchy of drives by configuring the drive location information at the time a storage node is installed. The software automatically maps drives into a three-level hierarchy consisting of storage nodes, racks, and data centers as shown in Figure 3. In this way, a drive is part of a storage node in a rack that is in a data center. ActiveScale object storage systems can be deployed in a single data center or across three geographically dispersed data centers. This allows DDP to place data such that the system can lose a storage node or an entire data center without data unavailability or loss.

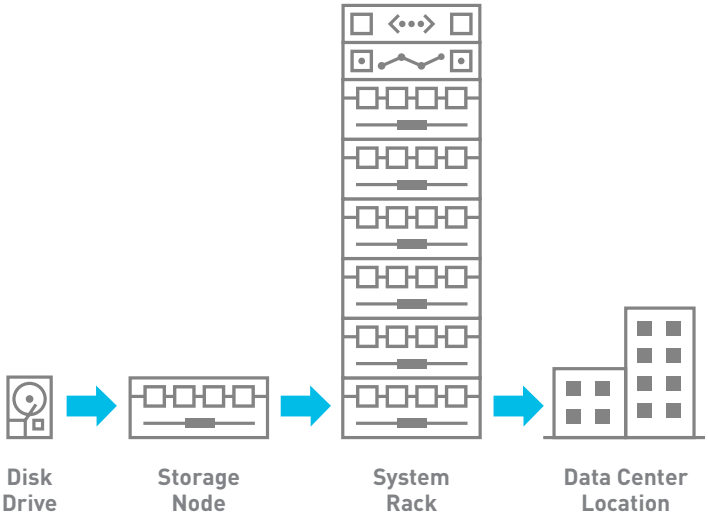

Figure 3: Data is spread across this hierarchy per the storage policy

Data Integrity

DDR – Self Protecting and Healing

DDR is an out-of-band maintenance system that performs a range of system monitoring tasks, data integrity verification and assurance, and self-healing of object data. System monitoring tasks include environmental parameters (temperature, fan) as well as disk health statistics. DDR sends alerts to the systems management layer including SNMP traps, email alerts and the management GUI.

Continuous Integrity Checking and Assurance

Writing data to disk drives is not a failsafe process. This silent data corruption is not proactively detected in a traditional storage system. Data corruption will only surface when it is read, causing the user to get an error. In this case, data will need to be restored from another copy or a backup.

However, DDP generates and stores each erasure encoding equation with an individual CRC checksum. Its fine-grained approach provides superior protection against sector-level bit errors, in contrast with a single, top-level checksum.

Repairing Bit Errors

DDR performs data integrity verification in the background by scanning the storage pool for checksum mismatches. If a check block (equation) becomes corrupted due to an unnoticed write error, bit rot or tampering, DDR will detect the error. It will delete the corrupted check block and generate a new, correct check block in another location on the disk. If a corrupted check block is detected while a user is reading an object from the system, then that check block will be ignored. As long as sufficient check blocks can be read from the system's storage nodes, the object will be accessible without impact to the data availability.

Repairing Spreads

Another aspect of the DDR rebuild process is the ability to automatically self-heal in a parallelized fashion. The effect of this is to shorten rebuild times after a drive failure. If a drive fails in an ActiveScale object storage system, the data chunks that were stored on that drive must be rebuilt and placed on other drives in the system.

With repair spreads generated on the fly per repaired object, the repaired chunks of data can be written to any drive in the system that meets the policy hierarchy spread rules. The dynamic nature of generating repair spreads enables the system to leverage all available disk spindles to target repair data. Also, all network connectivity in multiple storage nodes can be leveraged, which results in the highest possible repair performance.

Data Consistency

Another important aspect of DDP is that data is written using strong consistency. This is an essential requirement for many enterprise workloads. Strong consistency means that a client will never see out-of-date data for normal operations. After a successful write/PUT, the next successful read/GET of that object is guaranteed to show that write. In contrast, several alternative solutions use eventual consistency for normal operations. In this case, a GET after a successful PUT can fail or return an old version of the object. A LIST after a PUT can return a list where the object is not present. Eventual consistency can be very expensive for applications to workaround.

Conclusion

With no slowdown in sight for data growth, the imperative for IT remains the same – find more efficient and effective ways to store and protect the organization's vast store of valuable data. Traditional RAID technology has reached its limits to adequately protect petabyte-scale data stores, and better approaches are now available. The right storage architecture must simplify complexity and help organizations take advantage of their data without requiring budgets to scale at the same pace as data growth. It should deliver disk-based access performance from anywhere in the world, protect the data from loss with high durability, scale without limits and be easy to manage.

ActiveScale is a new class of storage built on patented object storage technology that addresses these needs. ActiveScale's architecture supports tens of petabytes and beyond with high data durability from DDP and high data integrity from DDR. The fundamental design based on a dynamic erasure coding provides better resiliency and seamless adoption of new capacity as customers grow their way into the future.

To learn more visit www.quantum.com/objectstorage

Quantum

Quantum technology and services help customers capture, create, and share digital content—and preserve and protect it for decades at the lowest cost. Quantum's platforms provide the fastest performance for high-resolution video, images, and industrial IoT, with solutions built for every stage of the data lifecycle, from high-performance ingest to real-time collaboration and analysis and low-cost archiving. Every day the world's leading entertainment companies, sports franchises, research scientists, government agencies, enterprises, and cloud providers are making the world happier, safer, and smarter on Quantum. See how at **www.quantum.com**.

www.quantum.com
800-677-6268

WP00254A-v02 Nov 2020