WHITE PAPER

INTRODUCTION TO DELL EMC XTREMIO X2 STORAGE ARRAY

A Detailed Review

Abstract

This white paper introduces the Dell EMC XtremIO X2 Storage Array (with XIOS Version 6.3.0). It provides detailed descriptions of the system architecture, theory of operation, and features. The paper also explains how the unique features of XtremIO—including Inline Data Reduction techniques, scalable performance, and data protection—provide unmatched solutions to today's data storage problems.

November, 2019

Contents

Executive Summary

Dell EMC XtremIO's 100% flash-based Scale-Out enterprise storage array delivers not only high levels of performance and scalability, but also brings new levels of ease-of-use to SAN storage, while offering advanced features that have never been possible before.

XtremIO's ground-up all-flash array design was created from the start for maximum performance and consistent low latency response times, and with enterprise grade high availability features, real-time Inline Data Reduction that dramatically lowers costs, and advanced functions such as thin provisioning, tight integration to VMware, copy, volume copies, and very efficient data protection.

The product architecture addresses all requirements for flash-based storage, including achieving longevity of the flash media, lowering the effective cost of flash capacity, delivering performance and scalability, providing operational efficiency, and delivering advanced storage array functionality.

This white paper provides a broad introduction to the Dell EMC XtremIO X2 Storage Array, with detailed descriptions of the system architecture, theory of operation, and its various features.

Introduction

XtremIO is an all-flash storage array that has been designed from the ground-up to unlock flash's full performance potential by uniquely leveraging characteristics of SSDs, based on flash media.

XtremIO uses industry-standard components and proprietary intelligent software to deliver unparalleled levels of performance. Achievable performance ranges from hundreds of thousands to millions of IOPS, and consistent low latency of under one millisecond.[1]

The system is also designed with simplicity, providing a user-friendly interface that makes provisioning and managing the array very easy. The system can be provisioned with minimal planning.

XtremIO leverages flash to deliver value across the following main dimensions:

- **Performance**

  Regardless of the system load, written capacity and workload characteristics, latency, and throughput remain consistently predictable and constant. Latency within the array for an I/O request is typically far less than one millisecond.[1]

- **Scalability**

  The XtremIO storage can be expanded for only capacity (Scale-Up) or both capacity and performance (Scale-Out). The system begins with a single building block, called an X-Brick with a minimum of 18 SSDs. When additional capacity is required, the system scales up, with up to 72 SSDs for a single X-Brick. When additional performance and capacity is required, the system can be expanded by adding additional X-Bricks.

  Performance scales linearly, ensuring that two X-Bricks can deliver twice the IOPS of a single X-Brick configuration, three X-Bricks deliver three times the IOPS, four X-Bricks deliver four times the IOPS, and so on. Latency remains consistently low as the system scales out.

[1] As measured for small block sizes. Large block I/O by nature incurs higher latency on any storage system.

- **Efficiency**

  The core engine implements content-based Inline Data Reduction. The XtremIO X2 Storage Array automatically reduces (deduplicates and compresses) data on the fly, as it enters the system. This reduces the amount of data written to flash, improving longevity of the media and driving down cost. XtremIO arrays allocate capacity to volumes on-demand in granular data blocks. Volumes are always thin-provisioned without any loss of performance, over-provisioning of capacity, or fragmentation. Once content-based inline deduplication is implemented, the remaining data is compressed even further, reducing the number of writes to the flash media. The data compression is carried out inline on the deduplicated (unique) data blocks.

  - Benefits gained from avoiding a large percentage of writes include:
  - Better performance due to reduced data
  - Increased overall endurance of the flash array's SSDs
  - Less required physical capacity to store the data, increasing the storage array's efficiency and dramatically reducing the $/GB cost of storage

- **Data Protection**

  XtremIO leverages a proprietary flash-optimized data protection algorithm (XtremIO Data Protection or XDP), which provides performance that is superior to any existing RAID algorithm. Optimizations in XDP also result in fewer writes to flash media for data protection purposes.

- **Integrated Copy Data Management**

  XtremIO allows consolidation of many different workloads and copies of those workloads (for example, test and development) on one array. You can create a larger number of high performance and space-efficient copies using XtremIO Virtual Copies (XVC) or using highly efficient XCOPY using VMware VAAI integration.

System Overview

The XtremIO X2 Storage Array is an all-flash system, based on flexible scaling options. The system uses building blocks called X-Bricks, which can be clustered together to grow performance or capacity, or both, as required.

System operation is controlled via a stand-alone dedicated Linux-based server called the XtremIO Management Server (XMS). An XMS host, which can be either a physical or a virtual server, can manage multiple XtremIO clusters. An array continues operating if it is disconnected from the XMS, but cannot be configured or monitored.

XtremIO architecture is based on a metadata-centric, content-aware system. The metadata-centric model helps to streamline data operations efficiently without requiring any movement of data. The system is content-aware where the data is laid out uniformly across all SSDs using the unique fingerprint of the incoming data.

This means that the architecture is specifically designed to deliver the full performance potential of flash, while linearly scaling and utilizing all resources such as CPU, RAM, SSDs, and host ports in a balanced manner. This allows the array to achieve any desired performance level, while maintaining consistency of performance that is critical to predictable application behavior.

The XtremIO Storage System provides a very high level of performance that is consistent over time, system conditions and access patterns. It is designed for true random I/O.

The system's performance level is not affected by its capacity utilization level, number of volumes, or aging effects. Moreover, performance is not based on a "shared cache" architecture and therefore it is not affected by the dataset size or data access pattern.

Due to its content-aware storage architecture, XtremIO provides:

- Even distribution of data blocks between the available SSDs, inherently leading to maximum performance and minimal flash wear
- Even distribution of metadata between the available SSDs
- No data or metadata hotspots
- Easy setup and no tuning
- Advanced storage functionality, including Inline Data Reduction (deduplication and data compression), thin provisioning, advanced data protection (XDP), Snapshots, and more

X-Brick

Figure 1 shows an X-Brick.

Figure 1. X-Brick

An X-Brick is the basic building block of an XtremIO array. Three X-Brick types are available: X2-R, X2-S and X2-T[2]. Systems using either type of X-Bricks can be used for any workload based on the workload's characteristics, capacity, and performance requirement. For example, if a deployment requires hundreds of terabytes of usable capacity, then X2-R is a better choice because each X2-R X-Brick can scale up to over 200 TB[3] of usable capacity. If a workload is I/O intensive, contains more duplicated data, and hence a smaller capacity footprint, X2-S might be a better fit.

Table 1. Single X-Brick High-Level Specifications

System	Supported Drive Types	SSDs per X-Brick
X2S	400GB	72
X2T	1.92TB	36
X2R	1.92TB / 3.84TB[4]	72 / 60

Customers must decide in advance which option is better for their deployment. The two X-Brick types cannot be mixed in a single cluster/system, and there is no way to transform an X-Brick from one type to the other. For detailed information on different models and their differences, refer to the “XtremIO X2 Specifications Sheet”.

[2] An X2-T model can be upgraded to X2-R.
[3] Fully-populated cluster with 3.84TB SSDs.
[4] In version 6.2, mixed configuration of 1.92TB and 3.84TB drives in the same cluster is not supported.

Each X-Brick is comprised of:

- One 2U Disk Array Enclosure (DAE), containing:
    - Up to 72 SSDs[5]
    - Two redundant power supply units (PSUs)
    - Two redundant SAS interconnect modules
- Two 1U Storage Controllers (redundant storage processors)

  Each Storage Controller includes:
    - Two redundant power supply units (PSUs)
    - Two 1/10GbE iSCSI ports
    - Two user interface interchangeable ports, either 4/8/16Gb FC or 1/10GbE iSCSI
    - Two 56Gb/s InfiniBand ports
    - One 100/1000/10000 Mb/s management port

Scale-Out Architecture: Linear Expansion of Capacity and Performance

An XtremIO storage system can include a single X-Brick or multiple X-Bricks, as shown in Figure 2.

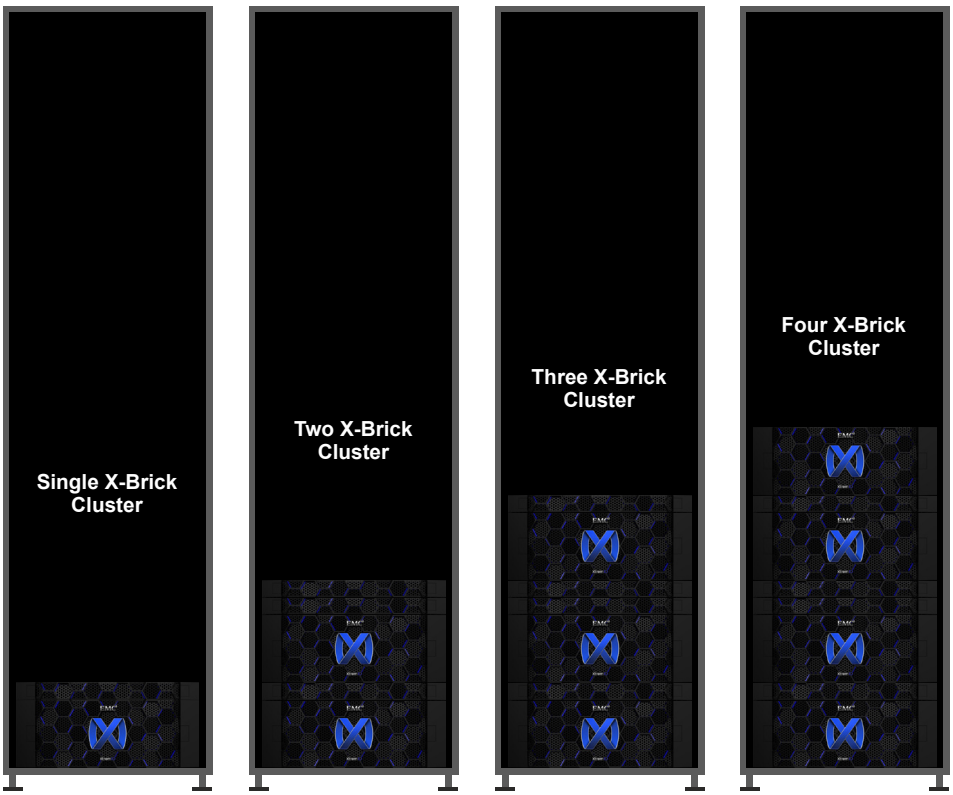


Figure 2. System Configurations as Single and Multiple X-Brick Clusters

With clusters of two or more X-Bricks, XtremIO uses a redundant 56Gb/s (4xFDR) InfiniBand network for back-end connectivity between the Storage Controllers, ensuring a highly available, ultra-low latency network.

[5] For 1.92TB SSDs. 60 SSDs is supported for a DAE populated with 3.84TB drives.

Multiple X-Brick clusters include two InfiniBand Switches. A single X-Brick cluster does not require any InfiniBand switches.

Table 2. InfiniBand Switches Types

	X2-R	X2-S	X2-T
Footprint	1U	1U	N/A
Port Count	36	12	N/A
Supported X-Bricks	4	4	N/A

Scale-Up Architecture: Capacity Expansion

Scale-Up refers to adding more storage capacity (physical SSDs) to an existing configuration without adding computing resources. The new capacity utilizes the existing X-Brick DAEs.

Each DAE holds up to 72 SSDs and divides them into two equal groups of 36 SSDs, known as Data Protection Groups (DPGs). A customer can either expand a partially populated DPG with new SSDs or start a new one if the old DPG is full. When expanding a DPG, there is no need to rebalance data between SSDs. DPG Layout is shown in Figure 3.

Every DPG can hold between 18 and 36 SSDs. The number of SSDs in a DPG can be increased by 2 or 18 SSDs at a time, depending on the existing number of SSDs in the DPG.

When scaling up a DPG:

- All SSDs in the cluster must have the same capacity
- The second DPG can only start to populate when the first DPG is full

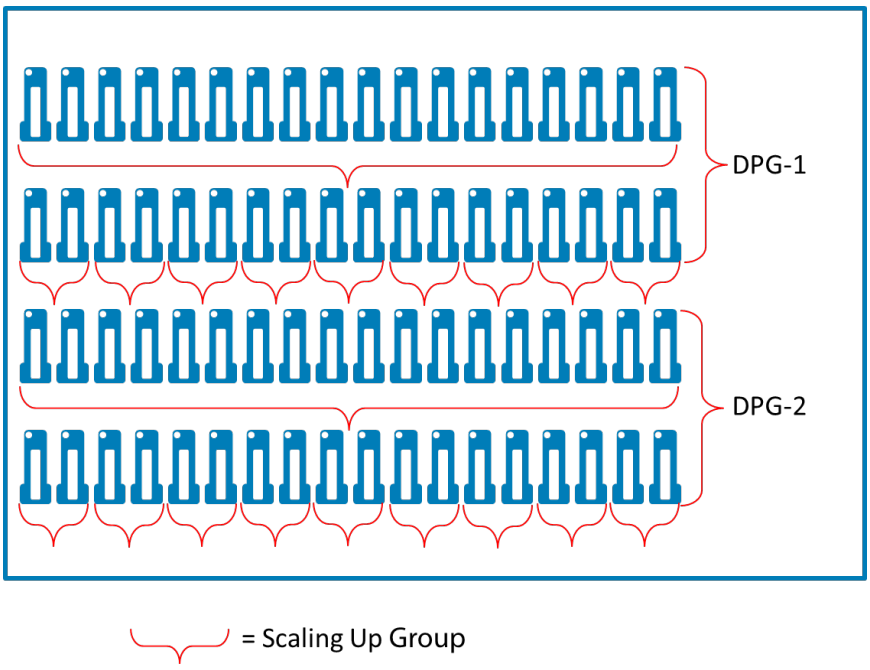


Figure 3. DAE

System Architecture

XtremIO integrates with existing SANs through 16Gb/s Fibre Channel or 10Gb/s Ethernet iSCSI connectivity to the hosts.

Unlike other block arrays, XtremIO is a purpose-built flash storage system, designed to deliver the ultimate in performance, ease-of-use and advanced data management services. Each Storage Controller within the XtremIO array runs a specially-customized lightweight Linux distribution as the base platform. The XtremIO Operating System (XIOS), runs on top of Linux and handles all activities within a Storage Controller, as shown in Figure 4. XIOS is optimized for handling high I/O rates and manages the system's functional modules, the RDMA over InfiniBand operations, monitoring and memory pools.

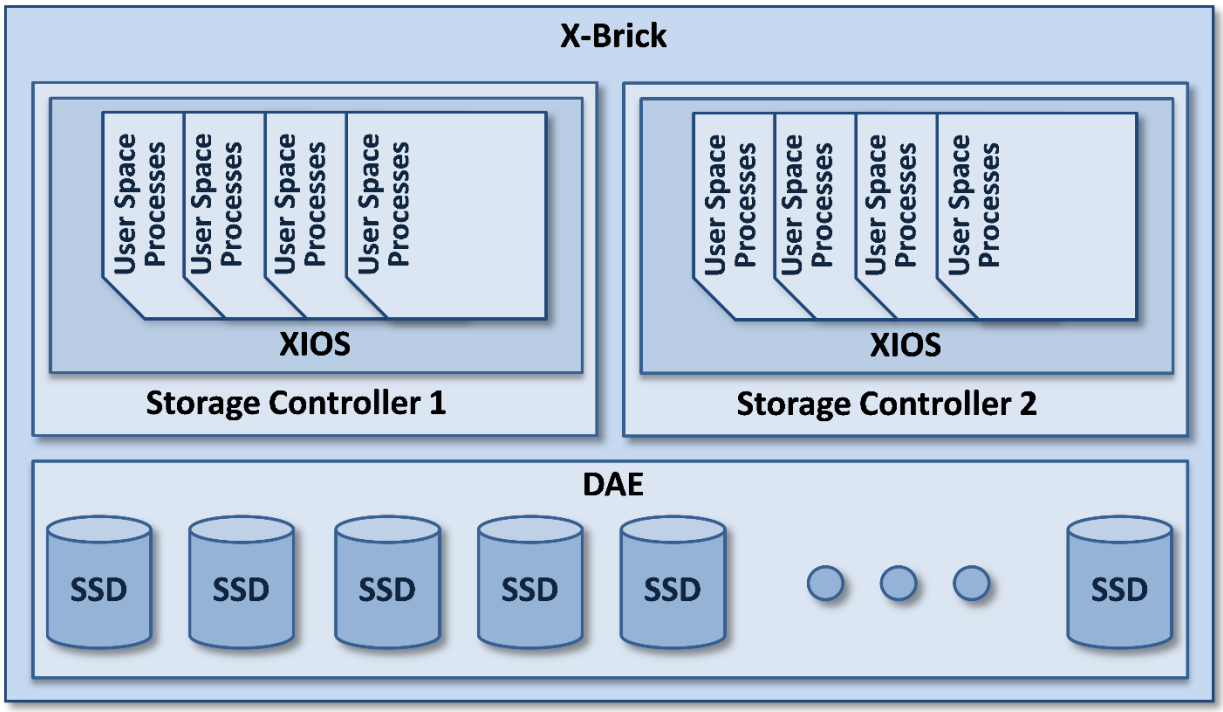


Figure 4. X-Brick Block Diagram

XIOS has a proprietary process-scheduling-and-handling algorithm designed to meet the specific requirements of the content-aware, low-latency, and high-performance storage subsystem.

XIOS provides:

- **Low-latency scheduling**: To enable efficient context switching of sub-processes, optimize scheduling, and minimize wait time
- **Linear CPU scalability**: To enable full exploitation of any CPU resources, including multi-core CPUs
- **Limited CPU inter-core sync**: To optimize the inter-sub-process communication and data transfer
- **No CPU inter-socket sync**: To minimize synchronization tasks and dependencies between the sub-processes that run on different sockets
- **Cache-line awareness**: To optimize latency and data access

The Storage Controllers on each X-Brick connect the enclosed disk array enclosure (DAE) via redundant SAS interconnects. The Storage Controllers from different X-Bricks are inter-connected to a redundant and highly available InfiniBand fabric. Regardless of which Storage Controller receives an I/O request from a host, multiple Storage Controllers on multiple X-Bricks cooperate to process the request. The data layout in the XtremIO system ensures that all components inherently share the load and participate evenly in I/O operations.

Theory of Operation

The XtremIO X2 Storage Array automatically reduces (deduplicates and compresses) data as it enters the system, processing it in data blocks. Deduplication is global (over the entire system), is always on, and is performed in real time (never as a post-processing operation). After the deduplication, the data is compressed inline, before it is written to the SSDs.

XtremIO uses a global memory cache which is aware of the deduplicated data, and content-based distribution that inherently spreads the data evenly across the entire array. All volumes are accessible across all X-Bricks and across all storage array host ports.

The system back-end connectivity is based on highly available InfiniBand network that provides high speeds with ultra-low latency and Remote Direct Memory Access (RDMA) between all Storage Controllers in the cluster. By leveraging RDMA, the XtremIO system is, in essence, a single, shared memory space spanning all Storage Controllers.

The effective logical capacity of a single X-Brick varies depending upon the data set being stored.

- For highly duplicated information, which is typical of many virtualized cloned environments such as Virtual Desktop Integration (VDI), the effective usable capacity is much higher than the available physical flash capacity. Deduplication ratios in the range of 5:1 to 10:1 are routinely achieved in such environments.
- For compressible data, which is typical in many databases and in application data, compression ratios are in 2:1 to 3:1 range.
- Systems benefitting from both data compression and data deduplication, such as Virtual Server Infrastructures (VSI), commonly achieve a 6:1 ratio.

XtremIO Metadata

The XtremIO system gathers a variety of metadata (fingerprint, location, mappings, and reference counts) on incoming data as explained below. This metadata is used as the fundamental insight for performing different system operations. For example, it is used to lay out incoming data uniformly and for implementing data services such as Thin Provisioning and Inline Data Reduction to eliminate data movement. While the metadata is used only within the system, it can also be used to communicate to external systems to reduce data movement (for example VMware XCOPY and Microsoft ODX).

Mapping Table

Each Storage Controller maintains a table that manages the location of each data block on SSDs, as shown in Table 3.

Table 3. Mapping Table Example

		LBA Offset		Fingerprint		SSD Offset / Physical Location		
Data	→	Address 0	→	20147A8	→	40	→	Data
Data	→	Address 1	→	AB45CB7	→	8	→	Data
Data	→	Address 2	→	F3AFBA3	→	88	→	Data
Data	→	Address 3	→	963FE7B	→	24	→	Data
Data	→	Address 4	→	0325F7A	→	64	→	Data
Data	→	Address 5	→	134F871	→	128	→	Data
Data	→	Address 6	→	CA38C90	→	516	→	Data
Data	→	Address 7	→	963FE7B	—	Deduplicated	—	X

Note: In Table 3, the colors of the data blocks correspond to their contents. Unique contents are represented by different colors while duplicate contents are represented by the same color (red).

The table has two parts:

- The first part of the table maps the host LBA offset to its content fingerprint.
- The second part of the table maps the content fingerprint to its physical location (offset) on SSDs. This provides XtremIO with the unique capability to distribute the data evenly across the array and place each block in the most suitable location on SSDs. It also enables the system to skip a non-responding drive or to select a location for writing new blocks when the array is almost full and there are no empty stripes on which to write.

How the Write I/O Flow Works

In a typical write operation, the incoming data stream reaches any one of the Active-Active Storage Controllers and is broken into data blocks. For every data block, the array fingerprints the data with a unique identifier.

The array maintains a table with this fingerprint (as shown in Table 3) to determine if incoming writes already exist within the array. The fingerprint is also used to determine the storage location of the data. The LBA-to-fingerprint mapping is recorded in the metadata within the Storage Controller's memory.

The system checks if the fingerprint and the corresponding data block have already been stored previously.

If the fingerprint is new, the system:

- Chooses a location on the array where the block will be written (based on the fingerprint, not the LBA)
- Creates the "fingerprint-to-physical location" mapping
- Compresses the data
- Writes the data
- Increments the reference count for the fingerprint by one

In case of a "duplicate" write, the system records the new LBA-to-fingerprint mapping, and increments the reference count on this specific fingerprint. Since the data already exists in the array, it is unnecessary to change the fingerprint-to-physical location mapping or to write anything to SSDs. All metadata changes occur within the memory. Therefore, the deduplicated write is carried out faster than the first unique block write. This is one of the unique advantages of XtremIO's Inline Data Reduction, where deduplication actually improves the write performance.

The actual write of the data block to SSDs is carried out asynchronously. At the time of the application write, the system places the data block into the in-memory write buffer. The data block is protected by journaling to local NVRAM. The data is then replicated to different Storage Controllers via RDMA when journaling to remote NVRAM, and the Storage Controller returns an acknowledgement to the host.

When enough blocks are collected in the buffer, the system writes them to the XDP (XtremIO Data Protection) stripe(s) on SSDs. This process is carried out in the most efficient manner, and is explained in detail in the XtremIO Data Protection White Paper.

When a Write I/O is issued to the array:

1. The system analyzes the incoming data and segments it into data blocks, as shown in Figure 5.

Figure 5. Data Broken into Fixed Blocks

2. For every data block, the array allocates a unique fingerprint to the data, as shown in Figure 6.

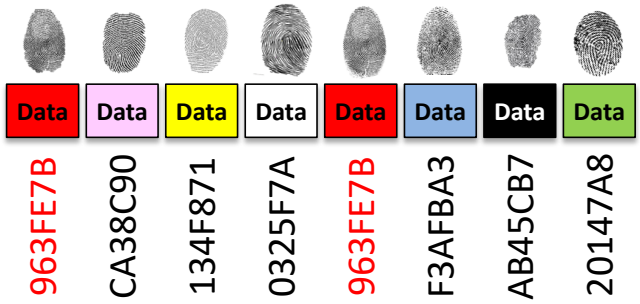


Figure 6. Fingerprint Allocated to Each Block

The array maintains a table with this fingerprint to determine if subsequent writes already exist within the array, as shown in Table 3.

- If a data block fingerprint does not exist in the system, it will be asynchronously written to the Unified Data Cache and then to DAE SSDs (as described in the next steps). The reference record is updated to point to the physical location, registering a reference count 1.
- If a data block fingerprint already exists in the system, it is not written in the Mapping table (as shown in Figure 7). The reference count for fingerprint will be increased by one.

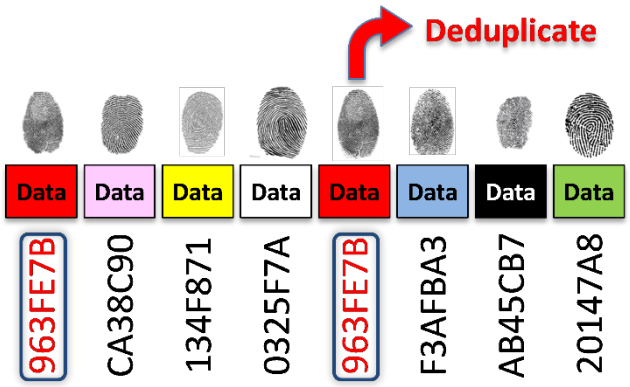


Figure 7. Deduplicating the Existing/Repeated Block

3. The user data is cached in both the local and remote Storage Controller.
4. The Mapping table is updated with fingerprint and the relevant LBA.
5. The system sends back an acknowledgement to the host.
6. Asynchronously, and using a consistent distributed mapping, the fingerprint is routed to the Storage Controller that is responsible for the relevant fingerprint address space.

   The consistent distributed mapping is based on the content fingerprint. The mathematical process that calculates the fingerprints results in a uniform distribution of fingerprint values and the fingerprint mapping is evenly spread among all Storage Controllers in the cluster, as shown in Figure 8.

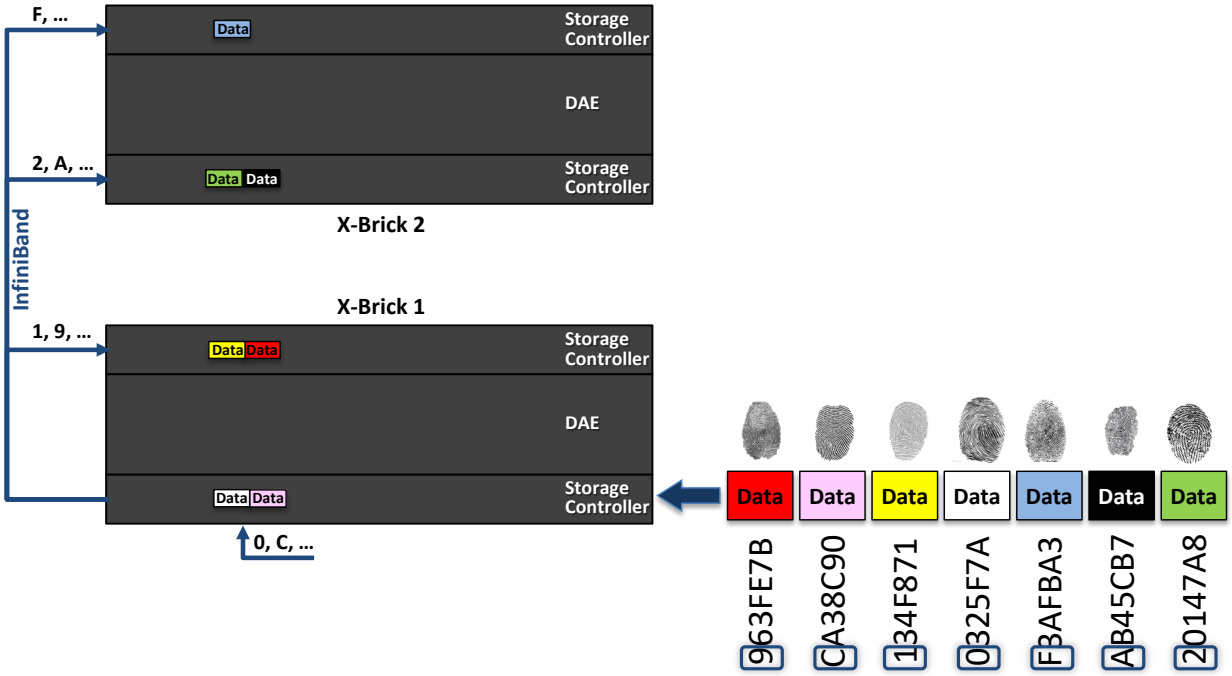


Figure 8. Data Spread Across the Cluster

Note: Data transfer across the cluster is carried out over the low latency, high-speed InfiniBand network using RDMA, as shown in Figure 8.

7. Due to the even distribution of the fingerprint function, each Storage Controller in the cluster receives an even share of the data blocks. When additional blocks arrive, they populate the stripes, as shown in Figure 9.

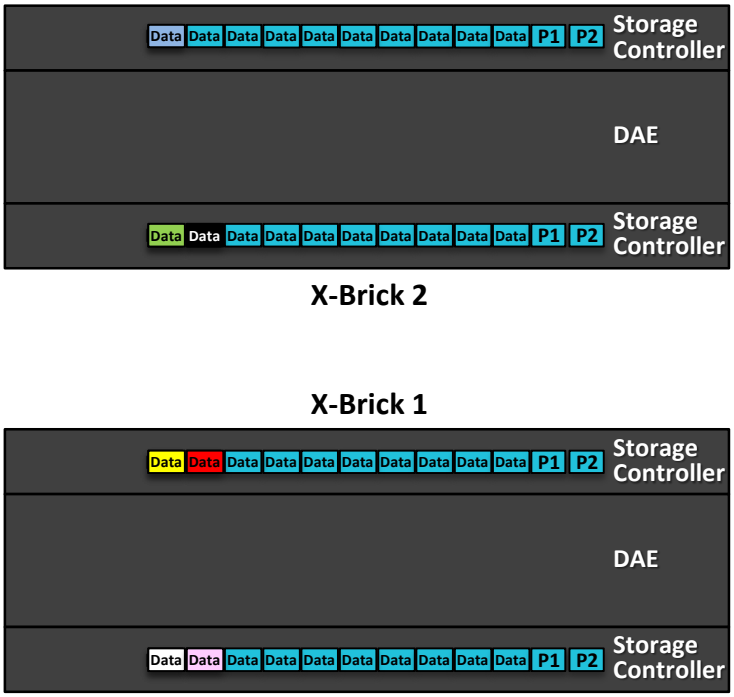


Figure 9. Additional Blocks Populated Full Stripes

8. The system compresses the data blocks to further reduce the size of each data block.
9. Once a Storage Controller has enough data blocks to fill the emptiest stripe in the array (or a full stripe if available), it transfers them from the cache to SSDs, as shown in Figure 10.

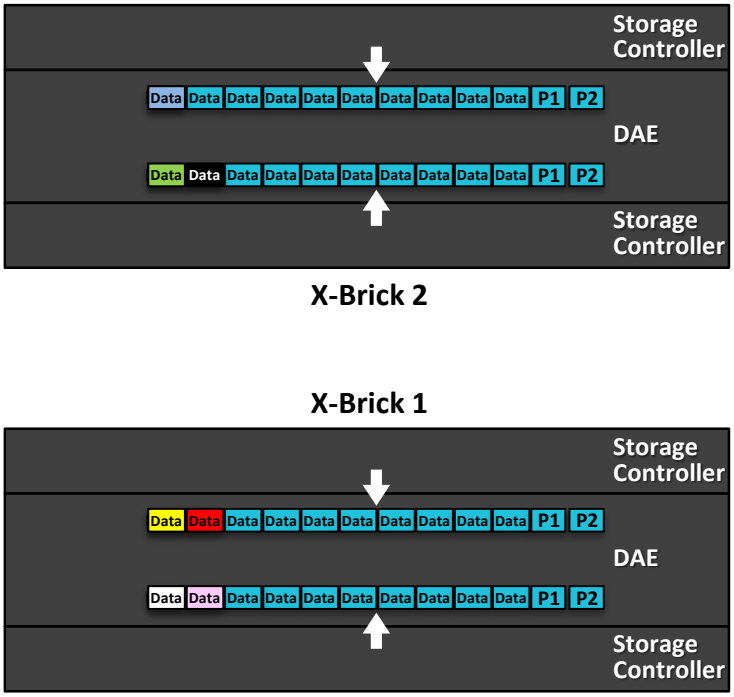


Figure 10. Stripes Committed to SSDs

How the Read I/O Flow Works

In a data block read operation, the system performs a look-up of the logical address in the LBA-to-fingerprint mapping. Once the fingerprint is found, it looks up the fingerprint-to-physical mapping, and retrieves the data block from the specific physical location. Because the data is evenly written across the cluster and SSDs, the read load is also evenly shared.

XtremIO has a memory-based read cache in each Storage Controller.

- In traditional arrays, the read cache is organized by logical addresses. Blocks at addresses that are more likely to be read are placed in the read cache.
- In the XtremIO array, the read cache is organized by content fingerprint. Blocks whose contents (represented by their fingerprint IDs) are more likely to be read are placed in the cache.

If the requested block size is larger than the data block size, XtremIO performs parallel data block reads across the cluster and assembles them into bigger blocks, before returning them to the application.

A compressed data block is decompressed before it is delivered to the host.

When a Read I/O is issued to the array:

1. The system analyzes the incoming request to determine the LBA for each data block and creates a buffer to hold the data.
2. The following processes occur in parallel:

- For each data block, the array finds the stored fingerprint. The fingerprint determines the location of the data block on an X-Brick. For larger I/Os (e.g., 256K), multiple X-Bricks are involved in retrieving each data block.
- The system transmits the requested Read data to the processing Storage Controller, via RDMA, over InfiniBand.

3. The system sends the fully populated data buffer back to the host.

System Features

The XtremIO X2 Storage Array offers a wide range of features that are always available and do not require special licenses. While some of these features may be seen on other vendor storage arrays, the architecture of these features is unique to XtremIO and is designed around the capabilities and the limitations of flash media.

System features include the following.

- Data services features—applied in sequence (as listed below) for all incoming writes
    - Inline Data Reduction:
        - Inline data deduplication
        - Inline data compression
    - XtremIO Virtual Copies ("Snapshots")
    - Thin Provisioning
    - XtremIO Data Protection (XDP)
    - Data at Rest Encryption

- System-wide features:
    - Scalable Performance
    - Even Data Distribution across all cluster resources
    - High Availability
    - Non-Disruptive Upgrade
    - Scale-Up (capacity expansion)
    - Scale-Out (capacity and performance expansion)
    - Integrated Copy Data Management
- Integrated features:
    - VMware VAAI Integration
    - Microsoft ODX Integration
    - RecoverPoint Integration for Local and Remote replication

Inline Data Reduction

Unique XtremIO Inline Data Reduction is achieved by utilizing the following techniques:

- Inline data deduplication
- Inline data compression

Inline Data Deduplication

Inline data deduplication is the removal of duplicate I/O blocks from data **before** it is written to the flash media.

XtremIO deduplication is always on and inline, which means that, unlike many systems in the market, XtremIO always deduplicates data as it enters the system without needing post-processing operation. With XtremIO, there are no resource-consuming background processes and no additional reads/writes (which are associated with post-processing). Therefore, the process does not negatively affect the performance of the storage array, does not waste the available resources that are allocated for the host I/O, and does not consume flash wear cycles. In addition, deduplication is at global level, which means no duplicate blocks are written over the entire array.

With XtremIO, data blocks are stored according to their content, and not according to their user level address within the volumes. This ensures perfect load balancing across all devices in the system in terms of capacity and performance. Each time a data block is modified, it can be placed on any set of SSDs in the system, or not written at all if the block's content is already known to the system.

The system inherently spreads the data across the array, using all SSDs evenly and providing perfect wear leveling. Even if the same LBA is repeatedly overwritten by a host computer, each write is directed to a different location within the XtremIO array. If the host writes the same data repeatedly, it will be deduplicated, resulting in no additional writes to the flash.

XtremIO uses a content-aware, globally deduplicated Unified Data Cache for highly efficient data deduplication. The system's unique content-aware storage architecture enables achieving a substantially larger cache size with a small DRAM allocation. Therefore, XtremIO is the ideal solution for difficult data access patterns, such as the "boot storms" that are common in virtual desktop (VDI) environments.

The system also uses the content fingerprints, not only for Inline Data Deduplication, but also for uniform distribution of data blocks across the array. This provides inherent load balancing for performance and enhances flash wear level efficiency, since the data never needs to be rewritten or rebalanced.

Performing this process inline, and globally across the array, translates into fewer writes to the SSDs. This increases SSD endurance and eliminates the performance degradation that is associated with post-processing deduplication.

XtremIO Inline Data Deduplication and intelligent data storage process ensure:

- Balanced usage of the system resources, maximizing the system performance
- Minimum amount of flash operations, maximizing the flash longevity
- Equal data distribution, resulting in evenly balanced flash wear across the system
- No system level garbage collection
- No deduplication post-processing
- Smart usage of SSD capacity, minimizing storage costs

Inline Data Compression

Inline Data Compression is the compression of the already-deduplicated data **before** it is written to the flash media. XtremIO automatically compresses data after all duplications have been removed. This ensures that the compression is performed only for unique data blocks. Data compression is performed in real time and not as a post-processing operation. The nature of the data set determines the overall compressibility rate. The compressed data block is then stored on the array.

Compression reduces the total amount of physical data that needs to be written on SSDs. This reduction minimizes the Write Amplification (WA) of the SSDs, thereby improving the endurance of the flash array.

XtremIO Inline Data Compression provides the following benefits:

- Data compression is always inline and is never performed as a post-processing activity. Therefore, the data is always written only once.
- Compression is supported for a diverse variety of data sets (e.g. database data, VDI, VSI environments, etc.).
- Data compression complements data deduplication in many cases. For example, in a VDI environment, deduplication dramatically reduces the required capacity for cloned desktops. Consequently, compression reduces the specific user data. As a result, an increased number of VDI desktops can be managed by a single X-Brick.
- Compression saves storage capacity by storing data blocks in the most efficient manner.
- When combined with powerful XtremIO Virtual Copies capabilities, XtremIO can easily support petabytes of functional application data.

Total Data Reduction

XtremIO data deduplication and data compression complement each other. Data deduplication reduces physical data by eliminating redundant data blocks. Data compression further reduces the data footprint by eliminating data redundancy within the binary level of each data block.

Figure 11 demonstrates the benefits of both the data deduplication and data compression processes combined, resulting in total data reduction.

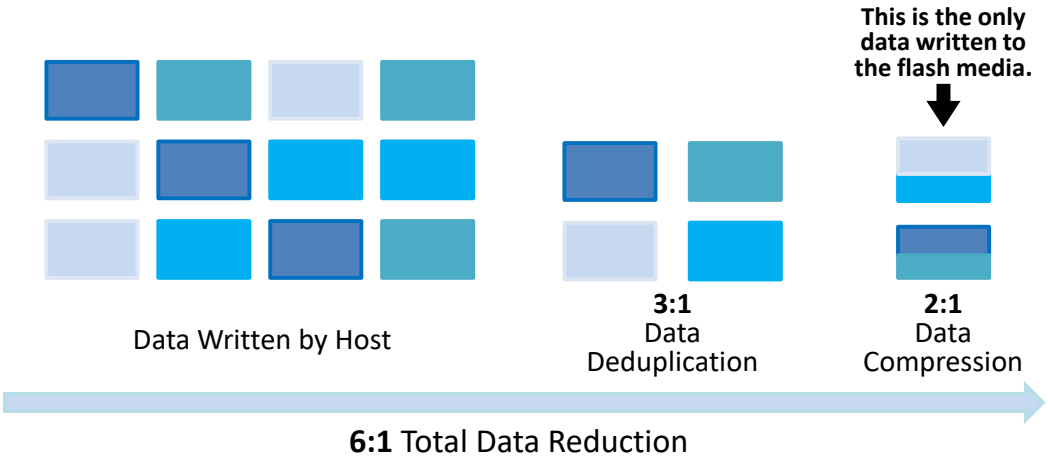


Figure 11. Data Deduplication and Compression Combined

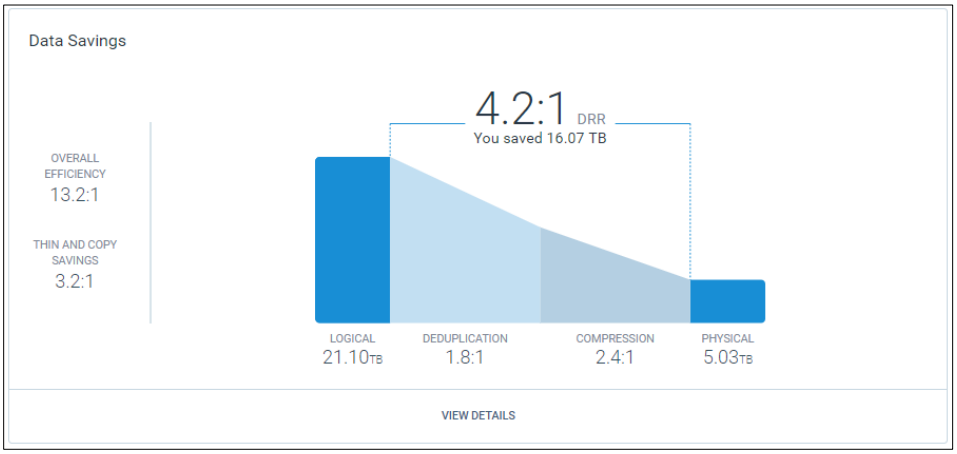
In the above example, the twelve data blocks written by the host are first deduplicated to four data blocks, demonstrating a 3:1 data deduplication ratio. Following the data compression process, the four data blocks are then each compressed by a ratio of 2:1, resulting in a total data reduction ratio of 6:1.

Data Reduction Measurement Mechanisms

Cluster Level Data Savings

Shows the cluster level data reduction counters including deduplication, compression, thin and copy savings and overall efficiency.

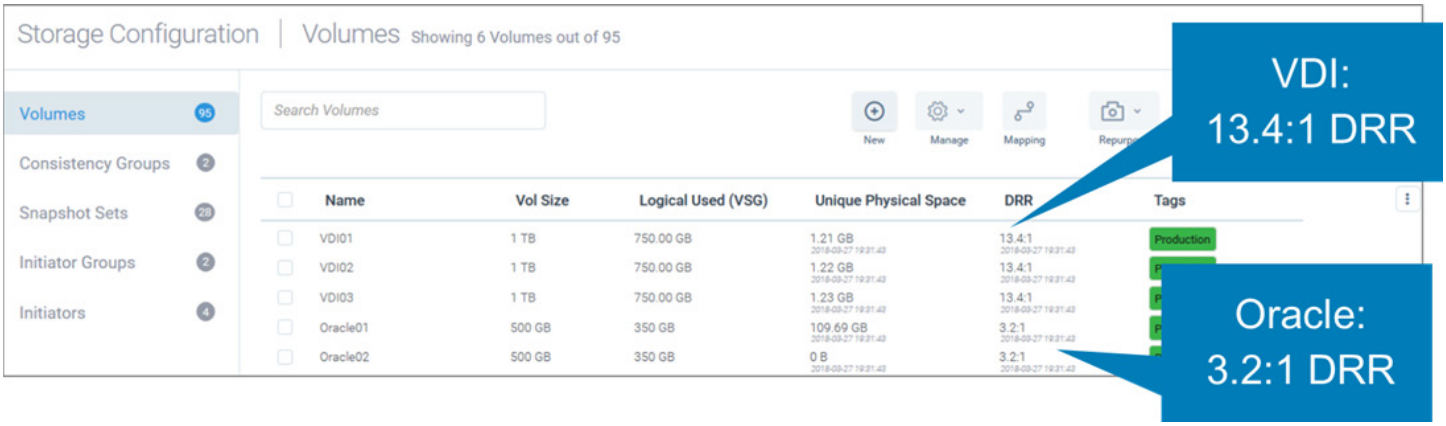
The following graph shows the DRR as 4.2:1, where Thin and Copy savings are 3.2:1, Deduplication is 1.8:1, and compression is 2.4:1



Data Reduction Ratio on Volume Level or VSG (Volume Snapshot Group)

The DRR per Volume and Volume Snapshot Group[6] is calculated by taking the data blocks of the Volume, and calculating their Deduplication and Compression rates. In most cases, Volumes in the same Snapshot tree have similar DRR rates.

The following example shows the DRR rate of a VDI Volume as 13.4:1. The DRR rate of an Oracle volume is 3.2:1.



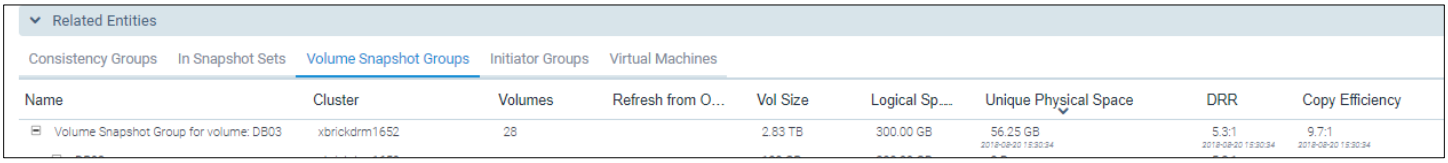
Unique Physical Space per Volume or VSG (Volume Snapshot Group)

The Unique Physical Space Counter counts all the blocks with a single occurrence (reference count) in the array, meaning the blocks that are not deduplicated. XtremIO does not differentiate between blocks that are unique to a Volume and blocks that are shared between different Volumes. This counter is useful for helping to understand what the current Volume or VSG space footprint on the array is, and what the amount of space projected to be released is, once the Volume or all Volumes belonging to the same VSG, are deleted.

Copy Efficiency per VSG (Volume Snapshot Group)

The Copy Efficiency calculates the ratio between all the VSG's host-accessible Volume blocks and the amount of logical blocks that are actually stored in the VSG. The relative amount of shared blocks affects the Copy Efficiency ratio. The more blocks that are shared within the VSG copies, the higher the Copy Efficiency.

The following example shows the DRR rate of a VSG for volume DB03 as 5.3:1. The Unique Physical Space for this VSG is 56.25GB and the Copy Efficiency is 9.7:1.



Integrated Copy Data Management (iCDM)

XtremIO pioneered the concept of integrated copy data management (iCDM)—the ability to consolidate both primary data and its associated copies on the same Scale-Out all-flash array for unprecedented agility and efficiency.

[6] All volumes belong to the same volume tree, as described in XtremIO Virtual Copies (Snapshots) on page 20.

XtremIO is the only storage platform capable of consolidating multiple workloads and entire business process workflows safely and efficiently, providing organizations with a new level of agility and self-service for new on-demand processes. To provide this kind of consolidation, the array must be able to deliver all performance SLAs in a consistent and predictable way, all the time while supporting on-demand copy operations at scale.

This has the following benefits:

- **Development**

  For development or testing purposes, the process of making a development Repurpose copy on-demand from production, pushing the copy to a QA host, pushing the QA copy to a scalability test-bed, and refreshing the output back into production leverages the fresh copies of production data. The result: 30 to 50 percent faster development, dramatically less infrastructure and complexity, higher productivity, reduced risks for your development, and increased quality of your product.

- **Analytics**

  For analytics, production data can be extracted and pushed to all downstream analytics applications on demand, as a simple in-memory operation. Copies are high performance and can get the same SLAs as production, without risking the production SLA. XtremIO offers this on demand, as both self-service and automated workflows for both application and infrastructure teams.

- **Operations**

  For the Application team, operations like patches, upgrades, and tuning copies can be made instantly. This allows the Application team to carry out tasks such as fast diagnosis of app/database performance problems, testing and validating patching processes, validating and applying tuning, or testing new technologies and infrastructure options for databases, all at production scale and performance.

- **Data Protection**

  XVC enables the creation of many copies at low interval points-in-time for recovery. Point-in-time copies can be used for recovery when needed or can be leveraged for other purposes. Application integration and orchestration policies can be set to auto-manage data protection, using different SLAs.

The iCDM offers the following efficient tools and workflows for managing and optimizing usability of Virtual Copies:

- **Consistency Groups**

  Consistency Groups (CG) are used to create a consistent image of a set of volumes, usually used by a single application, such as database. With XtremIO CGs, you can create a Virtual Copy of all volumes in a group, using a single command. This ensures that all volumes are created at the same time. Many operations that are applied on a single volume can also be applied on a CG.

  - **Snapshot Set**

    A Snapshot Set is a group of Virtual Copies Volumes that were taken using a single command and represents a point in time of a group. A Snapshot Set can be the result of a Snapshot taken on a CG, on another Snapshot Set, or on a set of volumes that were selected manually. The Snapshot Set taken from CG maintains a relationship with the ancestor from which it was created and allows subsequent management operations between the objects.

  - **Repurposing Copies**

    XtremIO Repurposing Virtual Copies are created as regular volumes. This volume type allows read-write, read-only, or no-access volume access types. The volume access can be changed without restrictions over time and according to data life cycle mandates.

- **Protection Copies**

  XtremIO Protection Copies are created as immutable read-only copies to satisfy the need for data protection and immediate recovery. The volume from this copy type can be mapped to an external host such as a backup application but cannot be written to. The volume access cannot be changed and always remains read-only, protecting the data from modification.

- **Protection Scheduler**

  The Protection Scheduler can be used for local protection use cases. It can be applied to a volume or CG. Each Scheduler can be defined to run at an interval of seconds, minutes or hours. Alternatively, it can be set to run at a specific time of a day or a week. Each Scheduler has a retention policy, based on the number of copies the customer would like to hold or based on the age of the oldest Snapshot.

- **Restore from Protection**

  Using a single command, it is possible to restore a production volume or a CG from one of its descendant Snapshot Sets.

- **Refresh a Repurposing Copy**

  The refresh command is a powerful tool for test and development environments and for the offline processing use case. With a single command, a Virtual Copy of the production volume or CG could be refreshed from a related copy, such as the parent object, or another dev/test copy. The operation does not require any changes in volume provisioning on XtremIO array or host SCSI rescan, but only host side Logical Volume Management operation to discover the changes.

iCDM is unique to XtremIO and all these benefits are available as a no-cost, array-based application service.

XtremIO Virtual Copies (Snapshots)

XtremIO Virtual Copies (XVC) are created by capturing the state of data in volumes at a particular point in time and allowing users to access that data when needed, even when the source volume has been changed or deleted. XtremIO Virtual Copies are inherently writeable, but can be created or changed to be read-only to maintain immutability. Virtual Copies can be taken from either the source or any Virtual Copy of the source volume.

XtremIO Virtual Copy technology is implemented by leveraging the content-aware capabilities of the system (Inline Data Reduction), optimized for SSDs, with a unique metadata tree structure that directs I/O to the right timestamp of the data. This allows efficient copy creation that can sustain high performance while maximizing the media endurance—both in terms of the ability to create multiple copies and the amount of I/O that a copy can support.

When creating a Virtual Copy, the system generates a pointer to the ancestor metadata (of the actual data in the system). Therefore, creating a Virtual Copy is a very quick operation that does not have any impact on the system and does not consume any capacity. Virtual Copy capacity consumption occurs only if a change requires writing a new unique block.

When a Virtual Copy is created, its metadata is identical to that of the ancestor volume. When a new block is written to the ancestor, the system updates the metadata of the ancestor volume to reflect the new write (and stores the block in the system, using the standard write flow process). As long as this block is shared between the Virtual Copy and the ancestor volume, it will not be deleted from the system following a write. This applies both to a write in a new location on the volume (a write on an unused LBA) and to a rewrite on an already-written location.

The system manages the Virtual Copy's and ancestor's metadata using a tree structure. The copy and the ancestor volumes are represented as leaves in this structure, as shown in Figure 12.

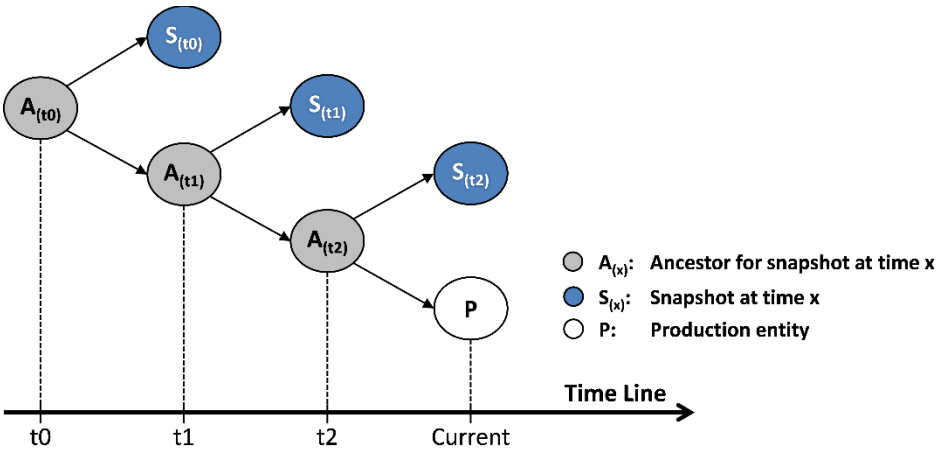


Figure 12. Metadata Tree Structure

The metadata is shared between all copy blocks that have not been changed (from the copy's original ancestor). The copy maintains unique metadata only for an LBA whose data block is different from that of its ancestor. This provides economical metadata management.

When a new Virtual Copy is created, the system always creates two leaves (two descendant entities) from the copied entity. One of the leaves represents the Virtual Copy, and the other one becomes the source entity. The copied entity will no longer be used directly, but will be kept in the system for metadata management purposes only.

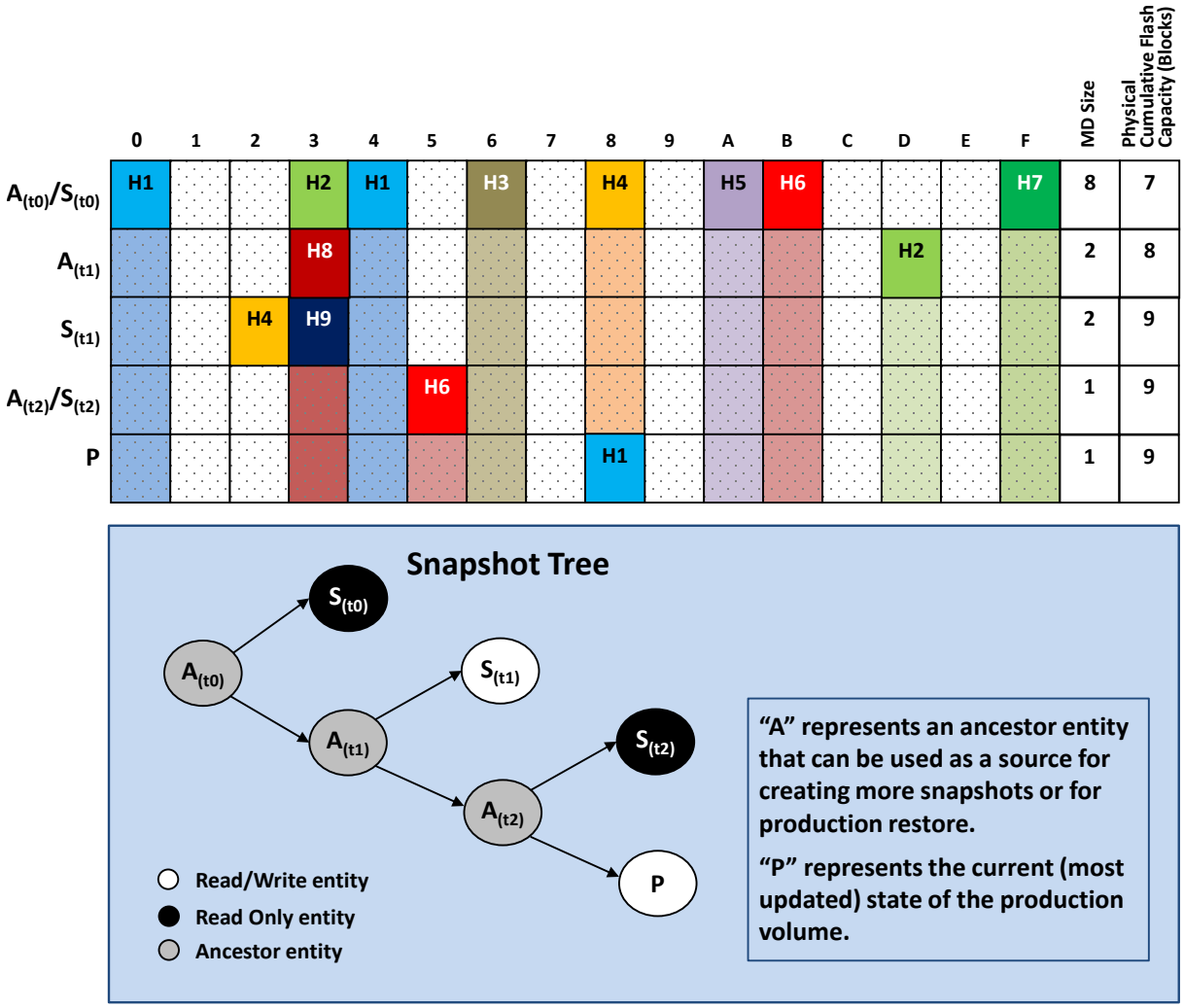


Figure 13. Creating Virtual Copies (Snapshots)

Figure 13 illustrates a 16-block volume in the XtremIO system. The first row (marked as $A_{(t0)}/S_{(t0)}$) shows the volume at the time the first Virtual Copy was taken (t0). At t0, the ancestor ($A_{(t0)}$) and the copy ($S_{(t0)}$) have the same data and metadata, because $S_{(t0)}$ is the read-only copy of $A_{(t0)}$ (containing the same data as its ancestor).

Note: Out of the 16 blocks, only 8 blocks are used. Blocks 0 and 4 consume only one block of physical capacity as a result of deduplication. The blanked dotted blocks represent the blocks that are thinly provisioned and do not consume any physical capacity.

In Figure 13, before creating the Virtual Copy at $S_{(t1)}$, two new blocks are written to P:

- H8 overwrites H2.
- H2 is written to block D. But it does not take up more physical capacity because it is the same as H2, stored in block 3 in $A_{(t0)}$.

$S_{(t1)}$ is a read-write copy. It contains two additional blocks (2 and 3) that are different from its ancestor.

Unlike traditional snapshots (which need reserved spaces for changed blocks and an entire copy of the metadata for each snap), XtremIO does not need any reserved space for Virtual Copies and never has metadata bloat.

At any time, XtremIO Virtual Copies consumes only the unique metadata, which is used only for the blocks that are not shared with the copy's ancestor entities. This allows the system to efficiently maintain large numbers of Virtual Copies, using a very small storage overhead, which is dynamic and proportional to the amount of changes in the entities.

For example, at time t2, blocks 0, 3, 4, 6, 8, A, B, D and F are shared with the ancestor's entities. Only block 5 is unique for this copy. Therefore, XtremIO consumes only one metadata unit. The rest of the blocks are shared with the ancestors and use the ancestor data structure to compile the correct volume data and structure.

The system supports the creation of Virtual Copies on a set of volumes. All Virtual Copies from the volumes in the set are cross-consistent and contain the exact same-point-in-time for all volumes. This can be created manually, by selecting a set of volumes for copying, or by placing volumes in a Consistency Group container and creating a Virtual Copy of the Consistency Group.

During the Virtual Copy creation, there is no impact on the system performance or overall system latency (performance is maintained). This is regardless of the number of copies in the system or the size of the Virtual Copy tree.

Virtual Copy deletions are lightweight and proportional only to the amount of changed blocks between the entities. The system uses its content-aware capabilities to handle copy deletions. Each data block has a counter that indicates the number of instances of that block in the system.

When a block is deleted, the counter value is decreased by one. Any block whose counter value is zero (meaning that there is no logical block address [LBA] across all volumes or Virtual Copies in the system that refers to this block) is overwritten by XDP when new unique data enters the system.

Deleting a child with no descendants requires no additional processing by the system. Deleting a Virtual Copy in the middle of the tree triggers an asynchronous process. This process merges the metadata of the deleted entity's children with that of their grandparents. This ensures that the tree structure is not fragmented.

With XtremIO, every block that needs to be deleted is immediately marked as freed. Therefore, there is no garbage collection and the system does not have to perform a scanning process to locate and delete the orphan blocks. Furthermore, with XtremIO, Virtual Copy deletion has no impact on system performance and SSD endurance.

The Virtual Copy implementation is entirely metadata driven and leverages the array's Inline Data Reduction to ensure that data is never copied within the array. Thus, many Snapshots can be maintained.

XtremIO Virtual Copies:

- Require no capacity.
- Allow for the creation of immutable Protection copies and/or Repurposing writable copies of the source volume.
- Are created instantaneously.
- Have negligible performance impact on the source volume and the copy itself.

Note: For more detailed information on Virtual Copies, refer to the XtremIO Virtual Copies White Paper.

XtremIO Remote Protection

XtremIO Metadata-Aware Replication

XtremIO Metadata-Aware Asynchronous Replication leverages the XtremIO architecture to provide the most efficient replication that reduces the bandwidth consumption. XtremIO Content-Aware Storage (CAS) architecture and in-memory metadata allow the replication to transfer only unique data blocks to the target array. Every data block that is written in XtremIO is identified by a fingerprint which is kept in the data block's metadata information.

- If the fingerprint is unique, the data block is physically written and the metadata points to the physical block.
- If the fingerprint is not unique, it is kept in the metadata and points to an existing physical block.

A non-unique data block, which already exists on the target array, is not sent again (deduplicated). Instead, only the block metadata is replicated and updated at the target array.

The transferred unique data blocks are sent compressed over the wire.

XtremIO Asynchronous Replication is based on Snapshot-shipping method that allows XtremIO to transfer only the changes, by comparing the changes between two subsequent Snapshots, benefiting from write-folding.

This efficient replication is not limited per volume, per replication session or per single source array, but is a global deduplication technology across all volumes and all source arrays.

In a fan-in environment, replicating from four sites to a single target site, as displayed in Figure 14, overall storage capacity requirements (in all primary sites and the target site) are reduced by up to 38 percent [7], providing the customers with considerable cost savings.

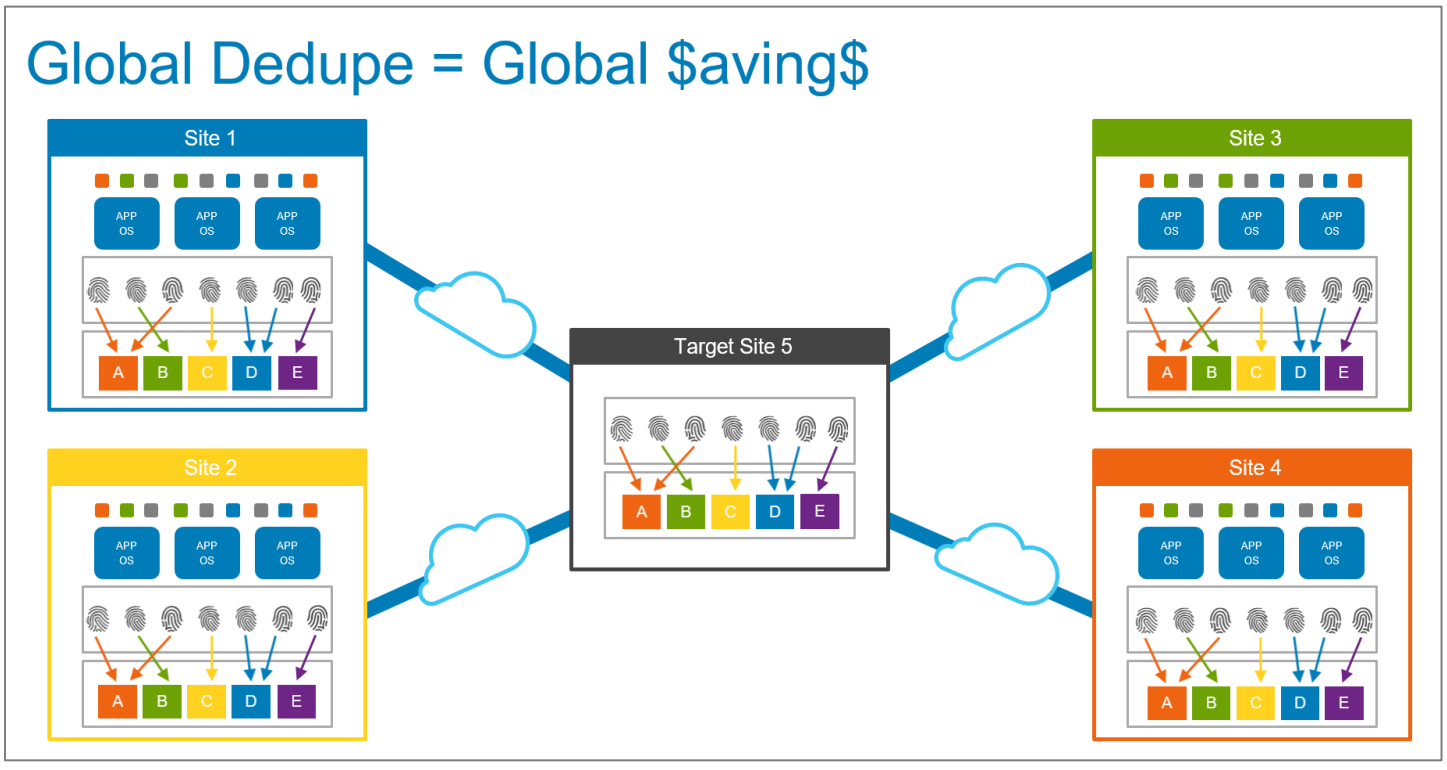


Figure 14. Creating Virtual Copies (Snapshots)

[7] Based on Dell EMC internal analysis, February 2019, with XtremIO Replication for fan-in 4:1 central DR site topology.

XtremIO Synchronous Replication

XtremIO enables you to protect the data, both asynchronously and synchronously, when "zero data loss" data protection policy is required.

XtremIO Synchronous Replication is fully integrated with asynchronous replication, in-memory Snapshots and iCDM capabilities, making it very efficient.

The challenge with Synchronous replication arises when the source and target are out of sync. This is true during the initial sync phase as well as when a disconnection occurs due to link failure or a user-initiated operation (for example, pausing the replication or performing failover).

Synchronous replication is highly efficient due to the following unique capabilities:

- Metadata-Aware replication – For the initial synchronization phase and when the target becomes out of sync, the replication uses the metadata-aware replication to efficiently and quickly replicate the data to the target. The replication uses multiple cycles until the gap is minimal, and then switches to synchronous replication. This reduces the impact on the production to a minimum and accelerates the sync time.
- Recovery Snapshots – To avoid the need for a full copy or even a full metadata copy, XtremIO leverages the in-memory Snapshot capabilities. Recovery-Snapshots are created on both sides every few minutes, which can be used as a baseline in case a disconnection occurs. When the connection is resumed, the system only needs to replicate the changes made since the most recent recovery Snapshot prior to the disconnection.
- Prioritization – In order to ensure the best performance for the applications using Synchronous Replication, XtremIO automatically prioritizes the I/O of Synchronous Replication over Asynchronous Replication. Everything is done automatically, and no tuning or special definition is required.
- Auto-Recovery from link disconnection – The replication resumes automatically when the link is back to normal.

XtremIO Synchronous Replication is managed at the same location in the GUI as the Asynchronous Replication and supports all disaster recovery operations, as supported by Asynchronous Replication.

Switching between asynchronous to synchronous is performed simply, using a single command.

Best Protection

XtremIO replication efficiency allows XtremIO to support the replication for All-Flash Arrays (AFA) high performance workloads. The replication supports both Synchronous Replication and Asynchronous Replication with an RPO as low as 30 seconds and can keep up to 500 PITs[8]. XtremIO offers simple operations and workflows for managing the replication and integration with iCDM for both Synchronous and Asynchronous Replication:

- **Test a Copy (current or specific PIT) at the remote host** –
  Testing a copy does not impact the production and the replication, continues to replicate the changes to the target array, and is not limited by time. The "Test Copy" operation uses the same SCSI identity for the target as the one used in case of Failover.

- **Failover** –
  Using the Failover command makes it possible to select the current or any PIT at the target and promote it to the remote host. Promoting a PIT is instantaneous.

- **Failback** –
  Fails over back from the target array to the source array.

- **Repurposing Copies** –
  XtremIO offers a simple command to create a new environment from any of the replication PITs.

- **Refresh a Repurposing Copy** –
  With a single command, a repurpose copy can be refreshed from any replication PIT. This is very useful when refreshing the data from the production to a test environment that resides on a different cluster or when refreshing the DEV environment from any of the existing build versions.

- **Creating a Bookmark on demand** –
  An ad-hoc PIT can be created when needed. This option is very useful when an application-aware PIT is required or before performing maintenance or upgrades.

[8] The values shown are of the initial release version. Refer to the relevant Release Notes for the most current scalability numbers.

Unified View for Local and Remote Protection

A dedicated tab for data protection is provided in the XtremIO GUI for managing XtremIO local and remote protection (Synchronous and Asynchronous Replication). In the **Data Protection Overview** screen, a high-level view displays the status for all Local and Remote protections, as shown in Figure 15.

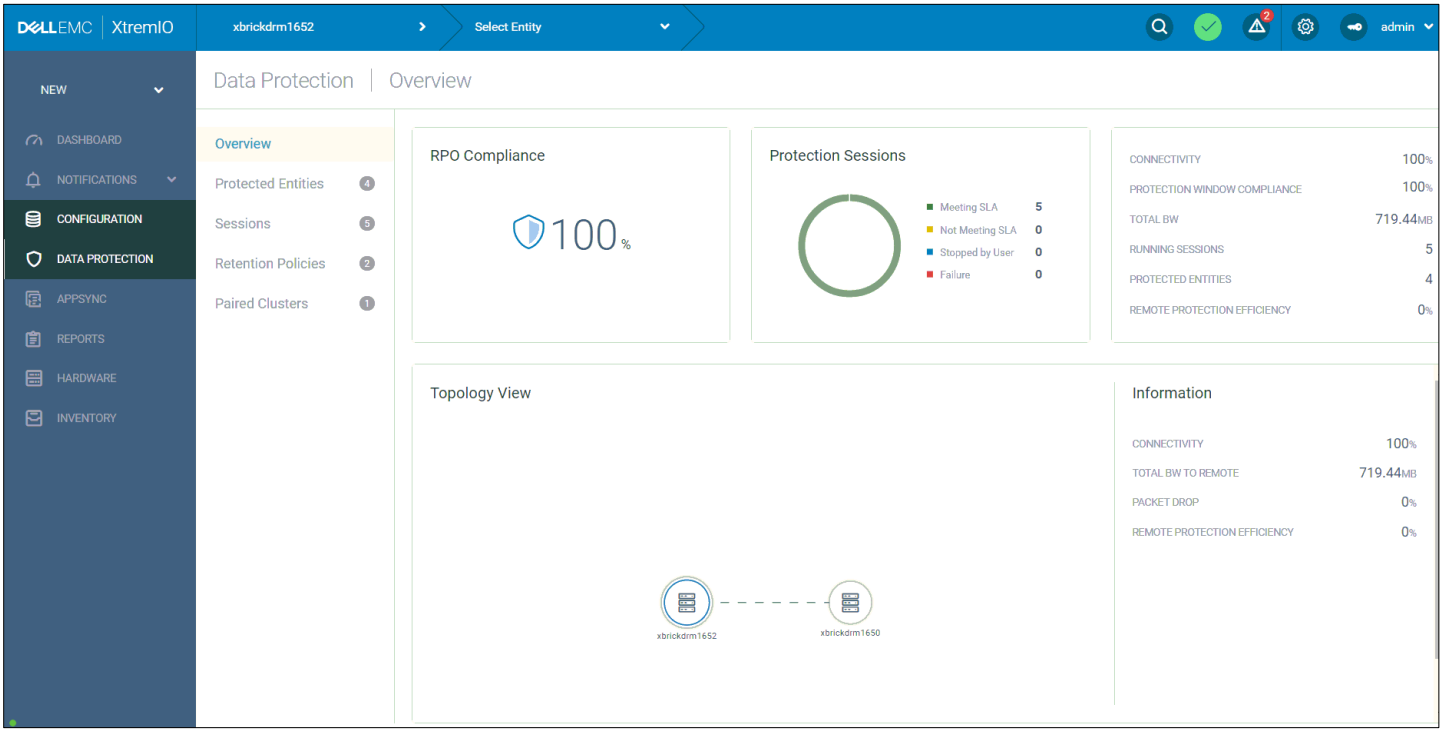


Figure 15. Data Protection Overview Screen

The Overview section includes:

- The minimum RPO compliance for all Local and Protection sessions
- Protection sessions status chart
- Connectivity information between peer clusters

From the overview screen it is easy to drill down to the session.

Consistency Group Protection View

With the new unified protection approach in one view it is easy to understand the protection for the consistency group. The **Topology View** pane displays the local and remote protection topology of the Consistency Group, as shown in Figure 16. Clicking each of the targets displays the detailed information in the **Information** pane.

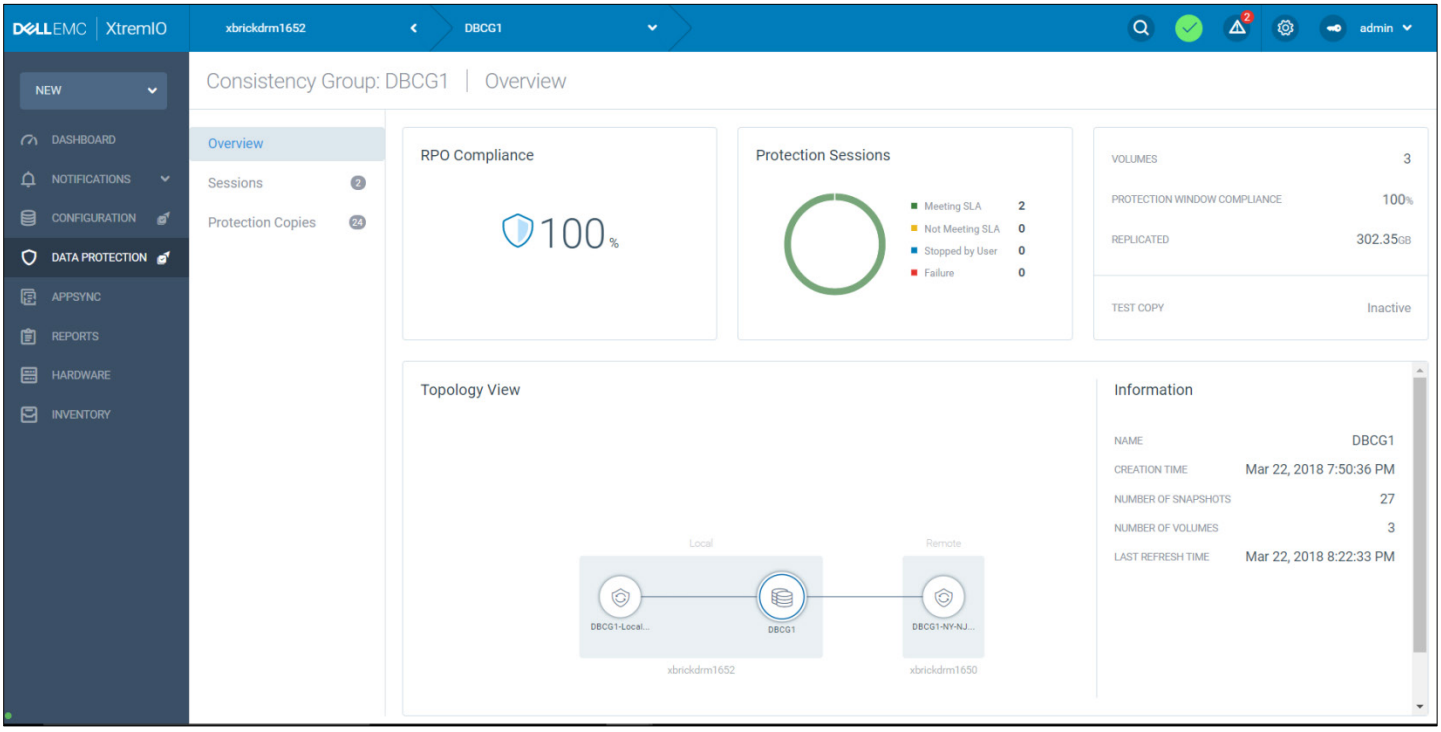


Figure 16. Consistency Group Overview Screen

Thin Provisioning

XtremIO storage is natively thin provisioned, using a small internal block size. This provides fine-grained resolution for the thin provisioned space.

All volumes in the system are thin provisioned, meaning that the system consumes capacity only when it is actually needed. XtremIO determines where to place the unique data blocks physically inside the cluster after it calculates their fingerprint IDs. Therefore, it never pre-allocates or thick-provisions storage space before writing.

As a result of the XtremIO content-aware architecture, blocks can be stored at any location in the system (and only metadata is used to refer to their locations) and the data is written only when unique blocks are received.

Therefore, unlike thin provisioning with many disk-oriented architectures, with XtremIO there is no space creeping and no garbage collection. Furthermore, the issue of volume fragmentation over time is not applicable to XtremIO (as the blocks are scattered all over the random-access array) and no defragmentation utilities are needed.

XtremIO inherent thin provisioning also enables consistent performance and data management across the entire life cycle of the volumes, regardless of the system capacity utilization or the write patterns to the system.

XtremIO Data Protection (XDP)

The XtremIO storage system provides "self-healing" double-parity data protection with a very high efficiency.

The system requires very little capacity overhead for data protection and metadata space. It does not require dedicated spare drives for rebuilds. Instead, it leverages the "hot space" concept, where any free space available in the array can be utilized for failed drive reconstructions. The system always reserves sufficient distributed capacity for performing a single rebuild.

In a rare case of double SSD failure, even with a full capacity of data, the array uses the free space to rebuild the data of one of the drives. It rebuilds the second drive once one of the failed drives is replaced. If there is enough free space to rebuild the data of both drives, it is performed simultaneously.

XtremIO maintains its performance, even at high capacity utilization, with minimal capacity overhead. The system does not require mirroring schemes (and their associated 100 percent capacity overhead).

XtremIO requires far less reserved capacity for data protection, metadata storage, Virtual Copies, spare drives, and performance, leaving much more space for user data. This lowers the cost per usable GB.

The XtremIO data protection algorithm provides:

- N+2 data protection
- Incredibly low data protection capacity overhead of 8 percent
- Performance superior to any RAID algorithm (RAID 1, the RAID algorithm that is most efficient for writes, requires over 60 percent more writes than XtremIO Data Protection)
- Flash endurance superior to any RAID algorithm, due to the smaller number of writes and even distribution of data
- Automatic rebuild in case of drive failure and faster rebuild times than traditional RAID algorithms
- Superior robustness with adaptive algorithms that fully protect incoming data, even when failed drives exist in the system
- Administrative ease through fail-in-place support

Table 4. Comparison of XtremIO Data Protection against RAID Schemes

Algorithm	Performance	Data Protection	Capacity Overhead	Reads per Stripe Update	Traditional Algorithm Read Disadvantage	Writes per Stripe Update	Traditional Algorithm Write Disadvantage
RAID 1	High	1 failure	50%	0	–	2 (64%)	1.6x
RAID 5	Medium	1 failure	25% (3+1)	2 (64%)	1.6x	2 (64%)	1.6x
RAID 6	Low	2 failures	20% (8+2)	3 (146%)	2.4x	3 (146%)	2.4x
XtremIO XDP	**60% better than RAID 1**	**2 failures per X-Brick**	**Ultra-Low 5.5% (34+2)**	**1.22**	**–**	**1.22**	**–**

How XDP Works

XtremIO Data Protection (XDP) is designed to take advantage of flash media specific properties and the XtremIO content addressable storage architecture.

Benefiting from the fact that it can control where data is stored without any penalty, XDP achieves high protection levels and low storage overhead, but with better performance than RAID 1. As an additional benefit, XtremIO Data Protection also significantly enhances the endurance of the underlying flash media, compared to any previous RAID algorithm, which is an important consideration for an enterprise flash array.

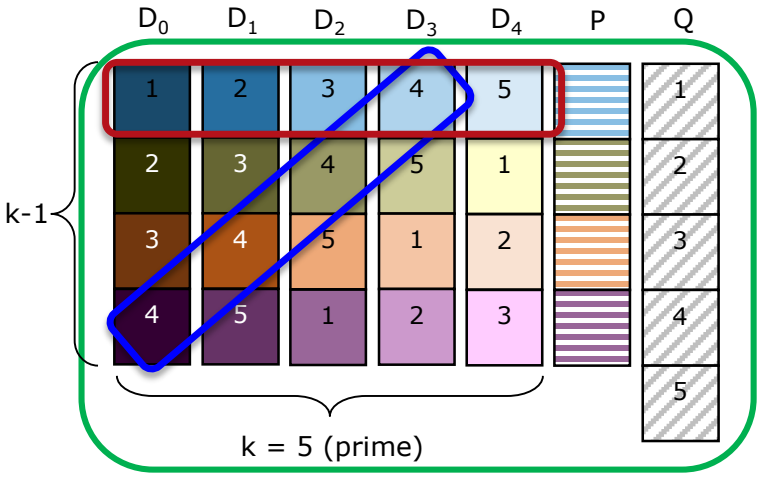


Figure 17. Row and Diagonal Parity

XDP uses a variation of N+2 row and diagonal parity, as shown in Figure 17, which provides protection from two simultaneous SSD errors. The X-Brick DAE may contain up to 72 SSDs organized in two Data Protection Groups (DPGs). The XDP is managed independently on the DPG level. With DPG of up to 36 SSDs, this results in 5.5 percent of capacity overhead.

Traditional arrays update logical block addresses (LBAs) in the same physical location on the disk (causing the high I/O overhead of a stripe update). XtremIO always places the data in the emptiest stripe. Writing data to the emptiest stripe effectively amortizes the overhead of read and write I/O operations for every stripe update and is only feasible in the XtremIO all-flash, content-based architecture. This process ensures that XtremIO performs consistently as the array fills and is in service for extended periods of time when overwrites and partial stripe updates become the norm.

XtremIO also provides a superior rebuild process. When a traditional RAID 6 array faces a single disk failure, it uses RAID 5 methods to rebuild it by reading each stripe and computing the missing cell from the other cells in the stripe. In contrast, XtremIO uses both the P and Q parity to rebuild the missing information and uses an elaborated algorithm that reads only the needed information for the next cell rebuild. As result, up to two simultaneous SSD failures are allowed per DPG.

Table 5. Comparison of XDP Reads for Rebuilding a Failed Disk with those of Different RAID Schemes

Algorithm	Reads to Rebuild a Failed Disk Stripe of Width K	Traditional Algorithm Disadvantage
XtremIO XDP	3K/4	–
RAID 1	1	None
RAID 5	K	33%
RAID 6	K	33%

Note: For more detailed information on XDP, refer to the XtremIO Data Protection White Paper.

Data at Rest Encryption

Data at Rest Encryption (DARE) provides a solution to securing critical data even when the media is removed from the array. XtremIO arrays utilize a high performance inline encryption technique to ensure that all data stored on the array is unusable if the SSD media is removed. This prevents unauthorized access in the event of theft or loss during transport, and makes it possible to return/replace failed components containing sensitive data.

DARE is a mandatory requirement that has been established in several industries, such as health care (where patient records must be kept closely guarded), banking (where financial data safety is extremely important), and in many government institutions.

At the heart of the XtremIO DARE solution is the use of Self-Encrypting Drive (SED) technology. An SED has dedicated hardware which is used to encrypt and decrypt data as it is written to or read from SSDs. Offloading the encryption task to the SSDs enables XtremIO to maintain the same software architecture whenever encryption is enabled or disabled on the array. All of the XtremIO features and services, including Inline Data Reduction, XtremIO Data Protection (XDP), thin provisioning, and XtremIO Virtual Copies are available on an encrypted cluster (as well as on non-encrypted clusters).

A unique Data Encryption Key (DEK) is created during the drive manufacturing process. The key does not leave the drive at any time. It is possible to erase the DEK or change it, but this causes the data on the drive to become unreadable and no option is provided to retrieve the DEK. To ensure that only authorized hosts can access the data on the SED, the DEK is protected by an Authentication Key (AK). Without this key, the DEK is encrypted and cannot be used to encrypt or decrypt data.

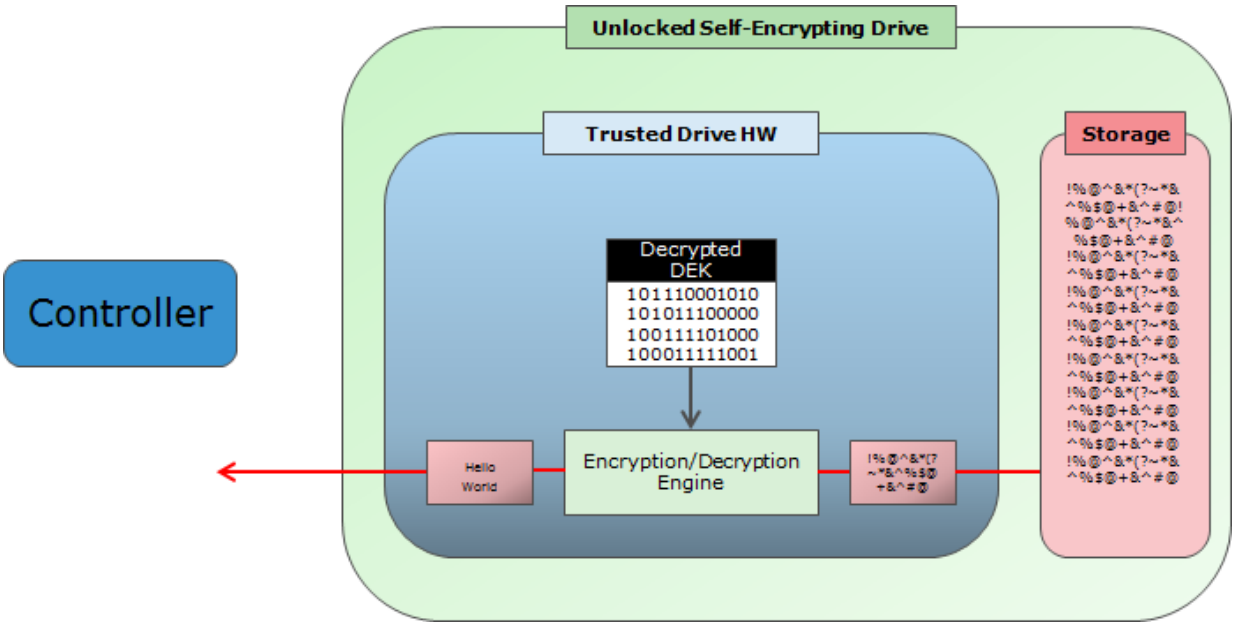


Figure 18. Unlocked SED

SEDs are shipped out of the factory in an unlocked state, meaning that any host can access the drive data. In unlocked drives, the data is always encrypted but the DEK is always decrypted and no authentication is required.

Locking the drive is made possible by changing the drive's default AK to a new, private AK and changing the SED settings so that it remains locked after a boot or power fail (such as when an SSD is taken out of the array). When an SSD is taken out of the array, it is turned off and will require the AK upon booting up. Without the correct AK, the data on the SSD is unreadable and safe.

To access the data, the hosts must provide the correct AK, a term that is sometimes referred to as "acquiring" or "taking ownership of" the drive, which unlocks the DEK and enables data access.

Drive acquisition is achieved only upon boot, and the SED remains unlocked for as long as the array is up. Since data passes through the encryption or decryption hardware in all cases, there is no performance impact when locking an SED.

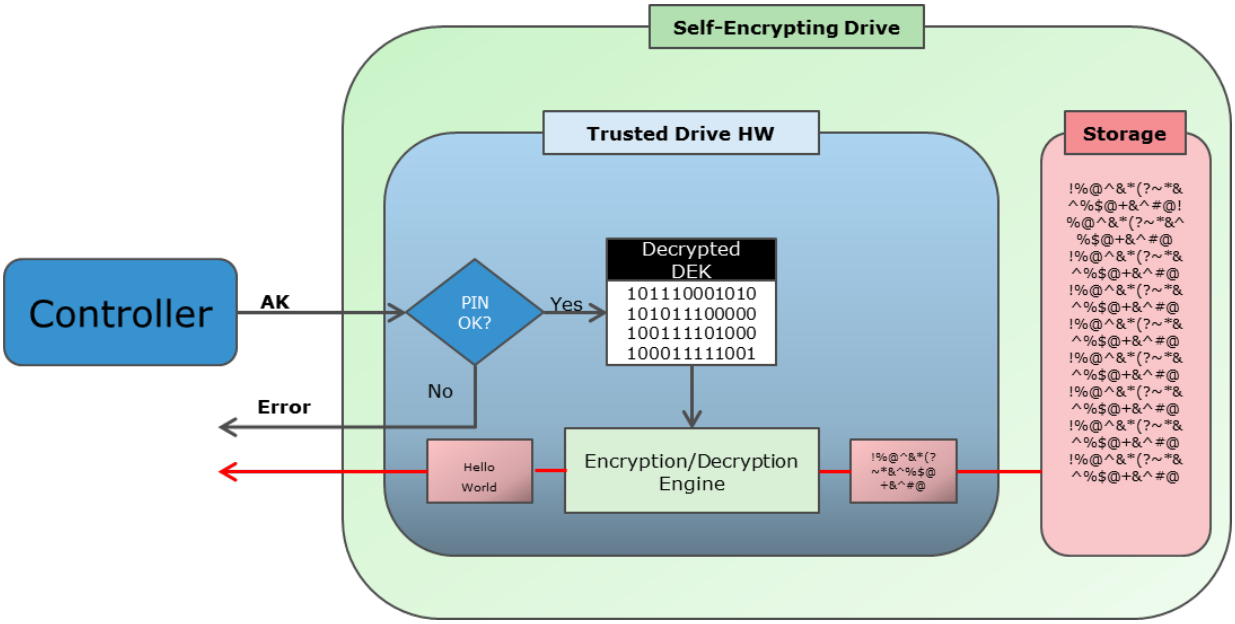


Figure 19. SED Operation Mode

Scalable Performance

XtremIO is designed to scale out to meet future performance and capacity needs, not only for new applications, but also for those already deployed. The XtremIO architecture allows performance and capacity to be increased by adding building blocks (X-Bricks), while maintaining a single point of management and balance of resources across the system.

Note:
- XIOS version 6.2.0 supports only 1, 2, 3, and 4 X-Brick configurations, in both X2-R and X2-S clusters. A future release will also support larger X-Brick configurations in X2-R clusters.
- X2-T supports a single X-Brick. An X2-T cluster type can be upgraded to X2-R.

Scale-Out is an intrinsic part of the XtremIO architecture and can be performed without a forklift upgrade of the existing hardware or any need for prolonged data transfers.

When additional performance is required, the XtremIO Storage System can be scaled-out by adding additional X-Bricks. Multiple X-Bricks are joined together over a redundant, highly available, ultra-low latency InfiniBand network.

When the system expands, resources remain balanced, and data in the array is distributed across all X-Bricks to maintain consistent performance and equivalent flash wear levels.

System expansion is carried out without any need for configuration or manual movement of volumes. XtremIO uses a consistent fingerprinting algorithm that minimizes remappings. A new X-Brick is added to the internal load-balancing scheme and only the relevant existing data is transferred to the new DAE.

The performance scales linearly, such that two X-Bricks supply twice the IOPS, three X-Bricks supply three times the IOPS, four X-Bricks supply four times the IOPS, etc. of the single X-Brick configuration. However, the latency remains consistently low (less than 1 ms[9]) as the system scales out, as shown in Figure 20.

[9] Sub-millisecond latency applies to typical block sizes. Latency for small blocks or large blocks may be higher.

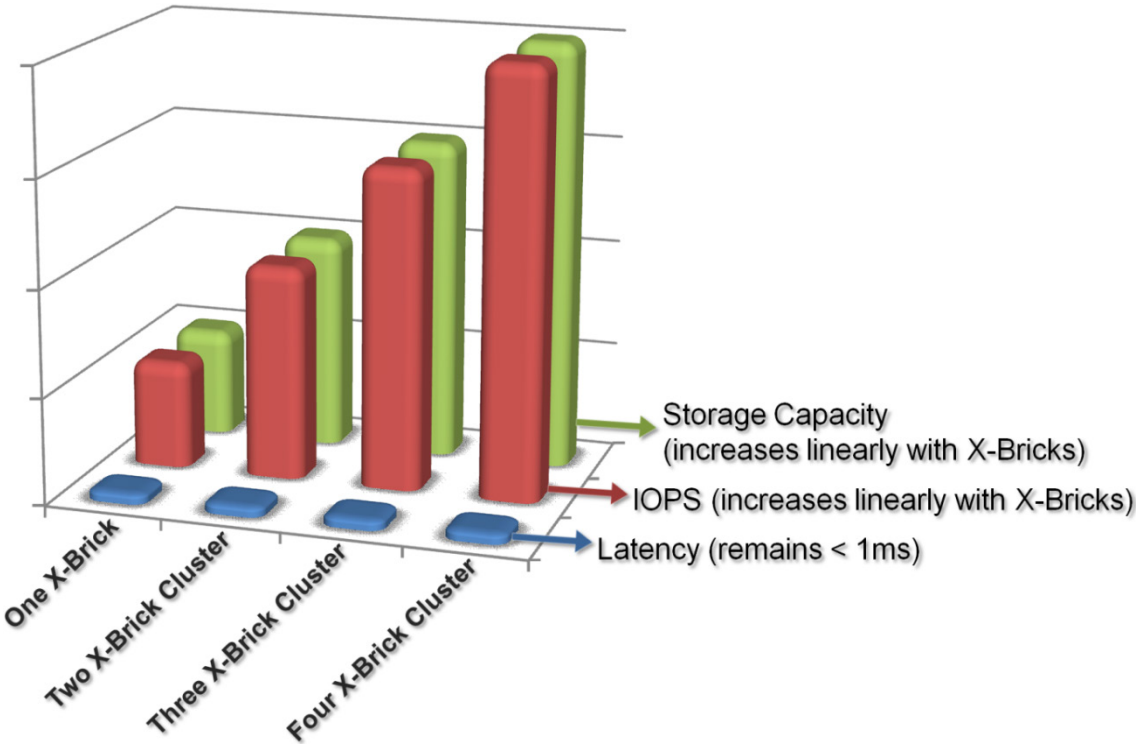


Figure 20. Linear Performance Scalability with Consistent Low Latency

The system architecture also deals with latency in the most effective way. The software design is modular. Every Storage Controller runs a combination of different modules and shares the total load. These distributed software modules (on different Storage Controllers) handle each individual I/O operation, which traverses the cluster. XtremIO handles each I/O request by two software modules (2 hops), no matter if it is a single or a multiple X-Brick cluster. Therefore, the latency always remains consistent, regardless of the cluster size.

Note: The sub-millisecond latency is validated by actual test results, and is determined according to the worst-case scenario.[10]

InfiniBand plays a key role in the XtremIO architecture. XtremIO uses two types of communication over the InfiniBand backplane: Remote Procedure Calls (RPC) for control messages and Remote Direct Memory Access (RDMA) for moving data blocks.

InfiniBand has not only one of the highest bandwidths available in any interconnect technology (56Gb/s [4xFDR]), but also has the lowest latency. The round-trip time for an RDMA transfer of a data block between two XtremIO Storage Controllers is about 7 microseconds, making it almost negligible compared to the XtremIO 500-microsecond latency allowance for each I/O. This enables the software to select any necessary Storage Controller and SSD resources, whether they are local or remote (over InfiniBand) to the Storage Controller that receives the I/O.

All XtremIO enterprise features (including Inline Data Reduction, iCDM, XDP, HA, etc.) have been developed as part of the Scale-Out architecture. All data and metadata are evenly distributed across the entire cluster. I/Os are admitted to the array via all the host ports, utilizing SAN zones and multi-pathing. Therefore, since all the workload is evenly shared among the controllers and SSDs, it is virtually impossible for any performance bottlenecks to occur anywhere in the system.

With XtremIO:

- Processors, RAM, SSDs, and connectivity ports scale together, providing scalable performance with perfect balance.

[10] Sub-millisecond latency applies to typical block sizes. Latency for small blocks or large blocks may be higher.

- The internal communication is carried out via a highly-available 2 X 56Gb/s (4xFDR) per Storage Controller InfiniBand internal fabric.
- The cluster is N-way active, enabling any volume to be reached from any host port on any Storage Controller on any X-Brick with equivalent performance.
- RDMA zero-copy data access makes I/Os to local or remote SSDs equivalent, regardless of the cluster size.
- Data is balanced across all X-Bricks as the system expands.
- There is a higher level of redundancy, and the cluster is more resilient to hardware and software failures. In an N-way Active Scale-Out cluster, if one Storage Controller fails, the system loses only $1/N^{th}$ of the total performance.
- The system is easy to upgrade and, unlike traditional dual-controller systems, the XtremIO Scale-Out model allows customers to start small, and grow both storage capacity and performance as the workload increases.

Even Data Distribution

To external applications, XtremIO appears and behaves like a standard block storage array. However, due to its unique architecture, it takes a fundamentally different approach to internal data organization. Instead of using logical addresses, XtremIO uses the block contents to decide where to place data blocks.

XtremIO uses data blocks internally. In a write operation, any data chunks that are larger than the native block size are broken down into standard blocks when they first enter the array. The system calculates a unique fingerprint for each of the incoming data blocks, using a special mathematical algorithm.

This unique ID is used for two primary purposes:

- To determine where the data block is placed within the array
- Inline Data Reduction

Because of the way the fingerprinting algorithm works, the ID numbers appear completely random and are evenly distributed over the possible range of fingerprint values. This results in an even distribution of data blocks across the entire cluster and all SSDs within the array. In other words, with XtremIO it is neither necessary to check the space utilization levels on different SSDs, nor to actively manage equal data writes to every SSD. XtremIO inherently provides even distribution of data by placing the blocks based on their unique IDs (see Figure 8).

XtremIO maintains the following metadata:

- Logical address (LBA)-to-fingerprint ID mapping
- Fingerprint ID-to-physical location mapping
- Reference count on each fingerprint ID

The system keeps all metadata in the Storage Controllers' memories and protects them by mirroring the change journals among different Storage Controllers, via RDMA. It also saves them periodically to SSDs.

Keeping all metadata in the memory enables XtremIO to provide the following unique benefits:

- **No SSD lookups**

  By avoiding SSD lookups, more I/Os are available to hosts' operations.

- **Instant Virtual Copies**

  Virtual Copy operations are instantaneous, as the process of taking a snap is carried out entirely in the array's memory (see Virtual Copies).

- **Instant VM cloning**

  Inline Data Reduction and VAAI, combined with in-memory metadata, enable XtremIO to clone a VM by memory operations only.

- **Steady performance**

  Physical locations of data, large volumes, and wide LBA ranges have no effect on the system performance.

High Availability

Preventing data loss and maintaining service in case of multiple failures is one of the core features in the architecture of the XtremIO All Flash Storage Array.

From the hardware perspective, no component is a single point of failure. Each Storage Controller, DAE, and InfiniBand Switch in the system is equipped with dual power supplies. The InfiniBand Switches are cross connected and create a dual data fabric. Both the power input and the different data paths are constantly monitored and any failure triggers a recovery attempt or failover.

The software architecture is built in a similar way. Every piece of information that is not committed to SSDs is protected by Non-Volatile Random Access Memory (NVRAM). Each software module has its own local Journal stored on the local NVRAM and mirrored to a remote Journal stored on NVRAM of remote Storage Controller. Both journals can be used to restore data in case of unexpected failure.

In addition, due to its Scale-Out design and the XDP data protection algorithm, each X-Brick DAE may contain up to two Data Protection Groups (DPGs) of up to 36 SSDs each, independent in terms of SSD failure. Each DPG can sustain up to two simultaneous SSD failures. This means, that for a DAE that has two active DPGs (54 or more SSDs) the maximal number of simultaneous failed SSDs is four, but not more than two per DPG.

The XtremIO cluster can handle multiple, non-simultaneous failures within the same DPG. Every DPG holds a reserve capacity equal to one failed SSD. This allows data to be rebuilt without impact on user physical space. The next SSD failure in this particular DPG group without previously failed disks replacement will consume the rebuild capacity from user SSD space.

When no free space is available for a consequent rebuild on a particular DPG, the system may consume free space from the second DPG of the same DAE, assuming it has free space.

If no free space is available for rebuild on both DPGs of same DAE, the system will shut down automatically to protect user data.

The total amount of consequent SSD failures supported per DPG is six for a fully-populated DPG (36 SSDs), and four for a non-fully-populated DPG (30 SSDs and less).

The XtremIO Active-Active architecture is designed to ensure maximum performance and consistent latency. The system includes a self-healing mechanism that attempts to recover from any failure and resume full functionality. An attempt to restart a failed component is performed once before a failover action. Storage Controller failover is carried out as the last resort. Based on the nature of the failure, the system attempts to failover the relevant software component, while maintaining the operation of other components, thus minimizing the performance impact. The whole Storage Controller fails over only if recovery attempts are not successful, or if the system must act in the best interest of protecting against data loss.

When a component that was temporarily unavailable recovers, a failback is initiated. This process is carried out at the software component or Storage Controller level. An anti-bounce mechanism prevents the system from failing back to an unstable component or to a component that is under maintenance.

Built on commodity hardware, XtremIO does not rely solely on hardware-based error detection and includes a proprietary algorithm that ensures detection, correction, and marking of corrupted areas. Any data corruption scenario that is not automatically handled by the SSD hardware is addressed by the XDP mechanism on the array or the multiple copies that are held in the Journals. The content fingerprint is used as a secure and reliable data integrity mechanism during read operations to avoid silent data corruption errors. If there is a mismatch in the expected fingerprint, the array will recover the data, either by reading it again or by reconstructing it from the XDP redundancy group.

Non-Disruptive Upgrade

During Non-Disruptive Upgrades (NDU) of the XtremIO Operating System, the system performs the upgrade procedure on a live cluster, updates all Storage Controllers in the cluster, and then restarts the user I/O flow. Since the underlying Linux Kernel is active throughout the upgrade process, the hosts do not detect any path disconnection during the application restart period.

The NDU procedure is launched from the XtremIO Management Server and upgrades the XtremIO software and the underlying operating system and firmware in a non-disruptive manner.

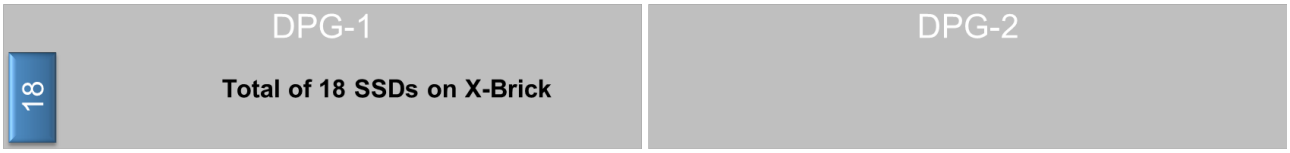
During Linux kernel upgrade (firmware NDU), the system automatically fails over a component and upgrades its software. After completing the upgrade and verifying the component's health, the system fails back to it and the process repeats itself on other components. During the upgrade process, the system is fully accessible, no data is lost, and the performance impact is kept to minimum. The host level path disconnects and failover is managed on host multipath software level.

Scale-Up

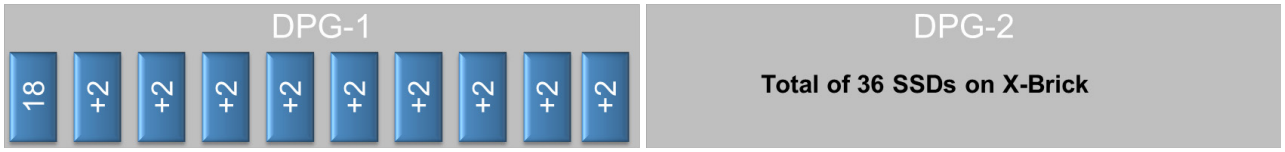
As part of the Scale-Up flow, the new SSDs are added to the existing DAEs and applied on active or new DPGs. No data transfer is required as a part of the Scale-Up process, assuming that the SSD space is balanced across DAE modules. For unbalanced configurations, the relative portion of physical and metadata space are rebalanced between X-Bricks.

Scale-Up Process on a Single X-Brick Cluster

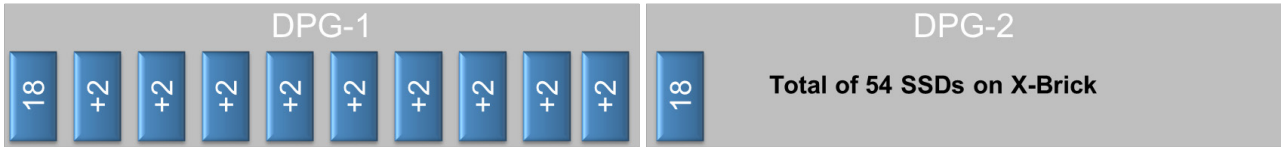
An initial X-Brick configuration of 18 SSDs comprises the first DPG.



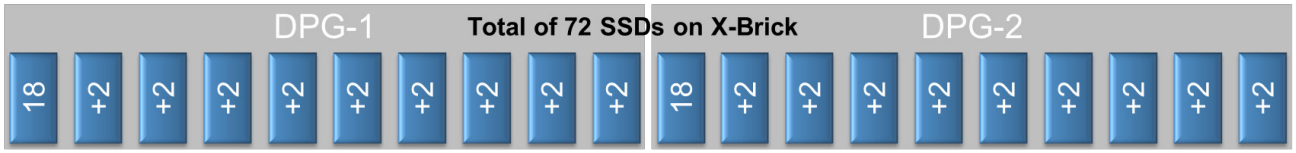
Scale-Up functionality allows any balanced (even) amount of SSDs, between 2 and 18 (2-SSD granularity) growth, up to a full configuration of the first DPG.



The next Scale-Up growth phase requires 18 SSDs as an initial configuration of the second DPG.



Next, scaling up with any balanced (even) amount of SSDs, between 2 and 18 (2-SSD granularity) allows populating the second DPG of 36 SSDs, and reaching a total of 72 SSDs for the X-Brick[11].

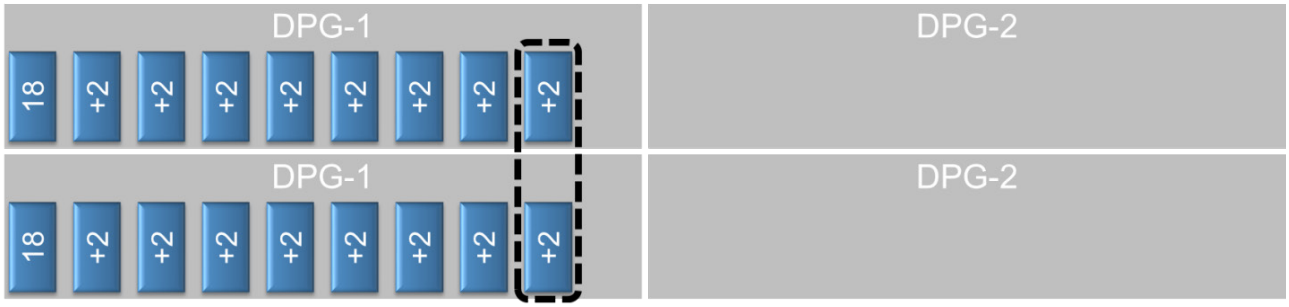


[11] 72 SSDs with 1.92TB drives, 60 SSDs with 3.84TB drives.

Scale-Up Process on a Multiple X-Brick Cluster

The Scale-Up process on a multiple X-Brick might be either:

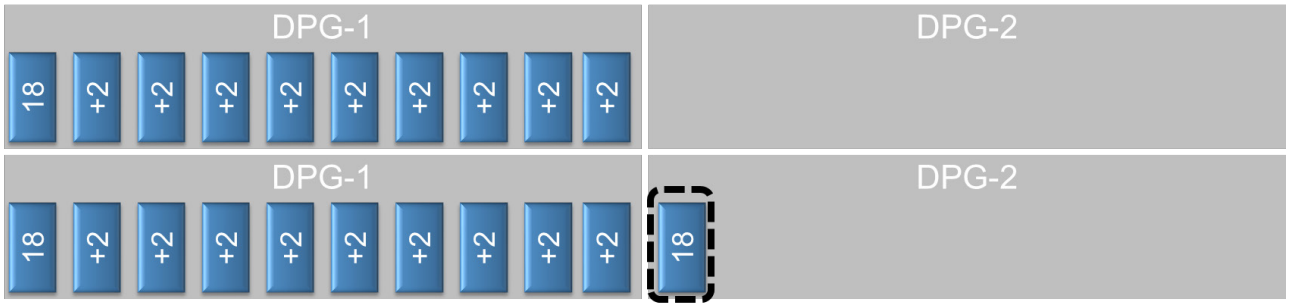
- Balanced growth, when the amount of SSDs is identical between X-Bricks.
- The Scale-Up granularity and "steps" in this case are similar to these of a single X-Brick, but is required to be applied consistently, across all of the X-Bricks within the XtremIO array.



Dual X-Brick Cluster Scale-UP from 64 to 68 SSDs

- Unbalanced growth, when opening the second DPG on a multiple X-Brick.

Initializing a second DPG requires 18 SSDs. In multiple X-Brick clusters, a balanced population of 18 SSDs for each X-Brick may require more SSDs than that required by the customer, from a capacity perspective. To support granular growth, this Scale-Up step is allowed to be sequential, and not concurrent, between X-Bricks. The next Scale-Up iterations requires 18 SSDs to be added to the remaining X-Bricks, reaching the balanced configuration of 54 SSDs across all DAEs.



Dual X-Brick Cluster Scale-UP from 72 to 90 SSDs

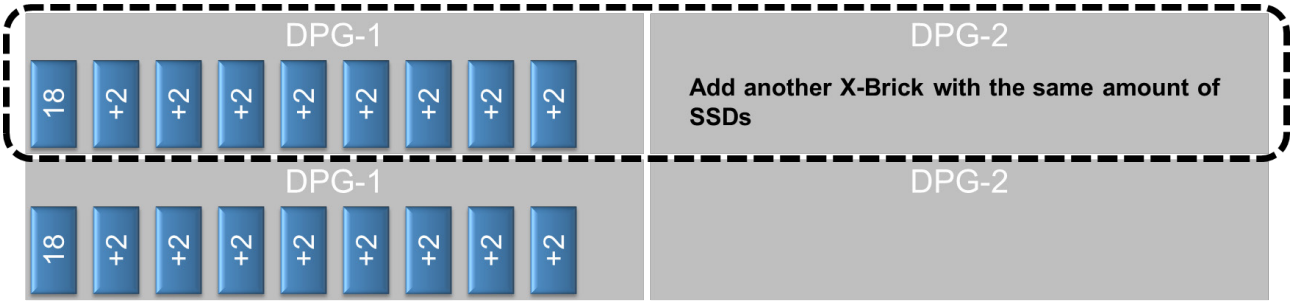
Scale-Out

Non-Disruptive Expansion allows the addition of both compute and storage resources (as described in Scalable Performance on page 33). System expansion is performed without the need for configuration or manual movement of Volumes or data, as part of the Scale-Out process.

As part of the Scale-Out process, the new X-Brick is added to the internal load balancing scheme, and only the relevant existing data is transferred to the new DAE. No user-level planning, migration, rebalance or any other additional management tasks are required.

As with a Scale-Up process, during a Scale-Out, the new resources may be added in the following two ways:

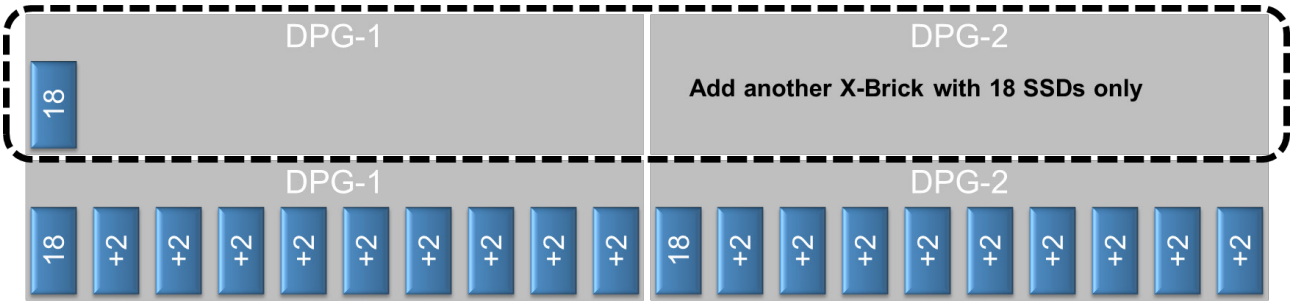
- Balanced Growth



  In this scenario, the amount of SSDs is identical between existing and the new X-Bricks.

- Unbalanced Growth

  When the existing X-Bricks are fully populated with 72 SSDs, balanced growth requires a fully-populated new X-Brick. In order to support more granular growth that fits the customer capacity requirements in this scenario, the new X-Brick may be added with a minimum of 18 SSDs, and with the granularity of 6 or 18 SSDs (18, 24, 30, 36, 54, 60, 66 or balanced 72).



Scale-Out from single X-Brick with 72 SSDs to dual X-Brick with 90 SSDs

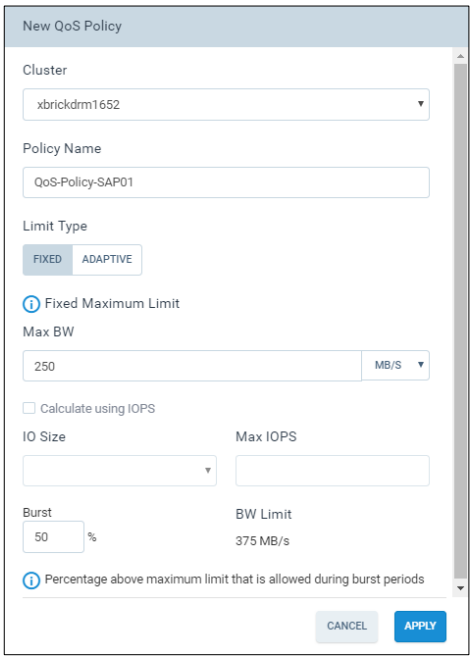
Quality of Service

Enterprises are leveraging Dell EMC XtremIO high performance storage to consolidate multiple applications and environments in a single cluster. In such consolidated deployments there are tens of applications, each with different performance needs and importance to the organization, running dozens of workloads against the storage array. Customers may want to limit some low-priority applications so that they will not use too much of the storage array's resources (noisy neighbors), and in that way to ensure that high-priority workloads will be better serviced by the cluster.

With XtremIO Quality of Service feature, customers can restrict a single Volume, an entire Consistency Group or an Initiator Group, to ensure that they will not pass a certain limit in terms of Bandwidth used or IOPS ran, and in that way allowing other workloads to utilize more of the cluster's resources.

To implement the feature, the XtremIO user is required to create a QoS policy and assign it for a Volume, Initiator Group or Consistency Group. The QoS Policy definition is simple, and requires setting the following three values, per policy:

1. The Limit Type (Fixed / Adaptive)
2. The Max Limit (in MB/s or IOPS)
3. Burst Percentage (from Max Limit)



When the "Adaptive" Limit Type is selected, the user can define the limit in "BW per GB" format, when the limitation is dynamic in accordance to the actual space defined for a particular Volume or CG.

VMware VAAI Integration

VAAI (vSphere Storage APIs for Array Integration) was introduced as an improvement to host-based VM cloning. Without VAAI, to clone a full VM, the host must read each data block and write it to the new address where the cloned VM resides, as shown in Figure 21. This is a costly operation that loads the host, the array, and the storage area network (SAN).

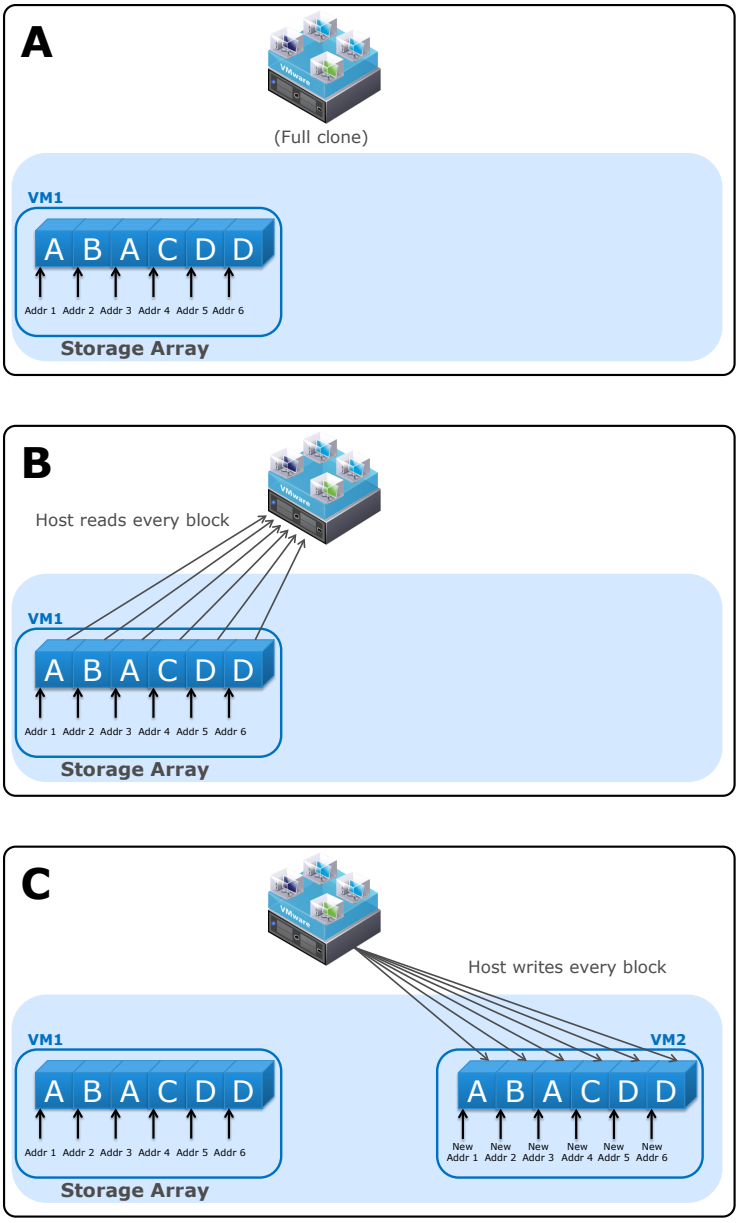


Figure 21. Full Copy without VAAI

With VAAI, the workload of cloning a VM is offloaded to the storage array. The host only needs to issue an X-copy command, and the array copies the data blocks to the new VM address, as shown in Figure 22. This process saves the resources of the host and the network. However, it still consumes storage array resources.

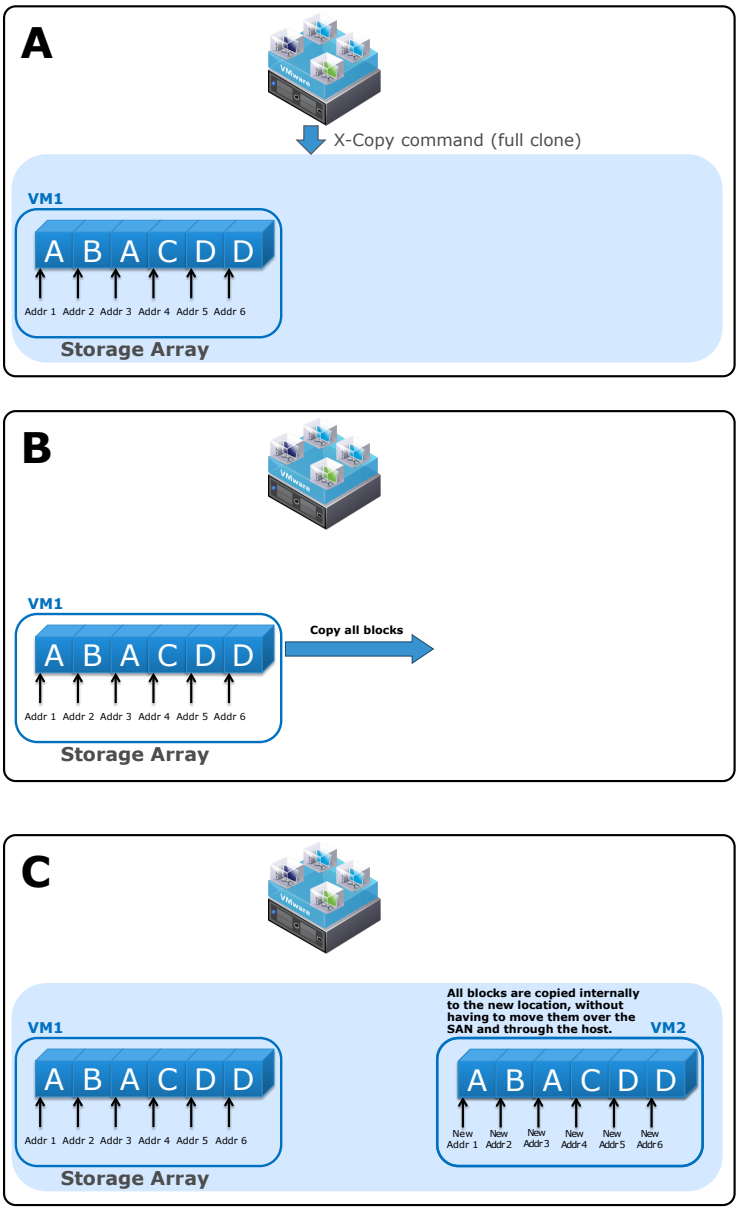


Figure 22. Full Copy with VAAI

XtremIO is fully VAAI compliant, allowing the array to communicate directly with vSphere and provide accelerated storage vMotion, VM provisioning, and thin provisioning functionality.

In addition, XtremIO VAAI integration improves the X-copy efficiency even further by making the whole operation metadata driven. With XtremIO, due to Inline Data Reduction and in-memory metadata, no actual data blocks are copied during the X-copy command. The system only creates new pointers to the existing data, and the entire process is carried out in the Storage Controllers' memory, as shown in Figure 23. Therefore, it does not consume storage array resources and has no impact on system performance.

For example, XtremIO can instantaneously clone a VM image (even multiple times).

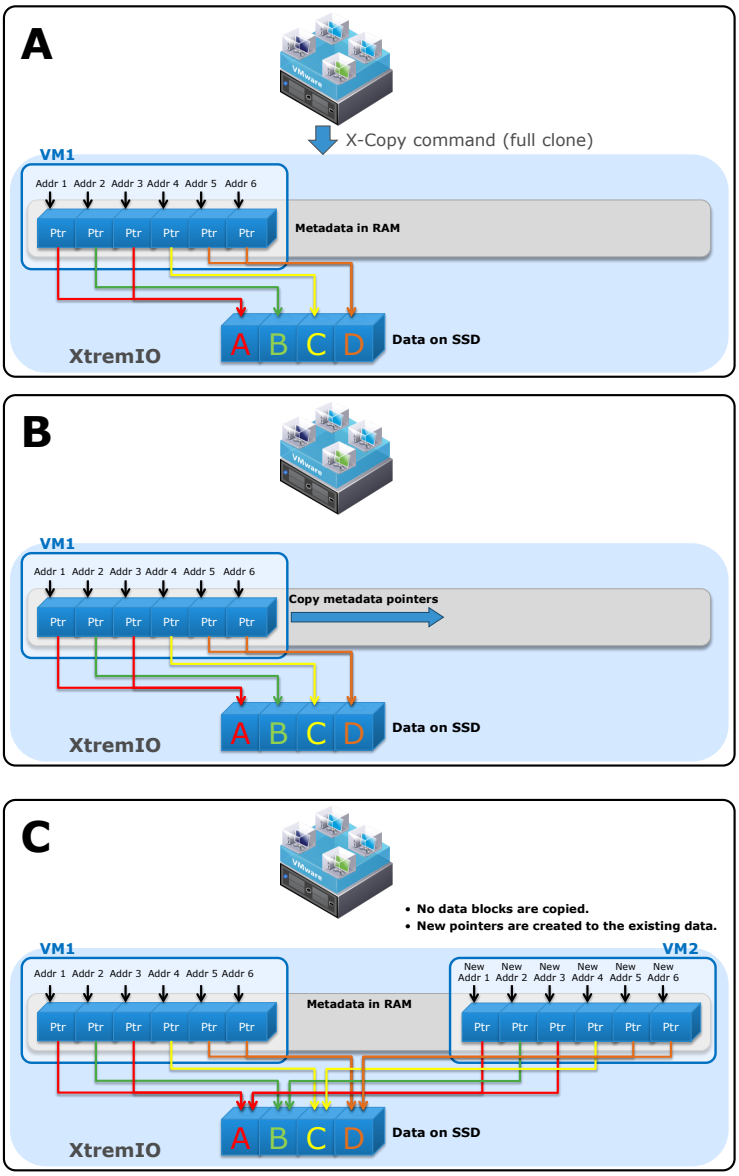


Figure 23. Full Copy with XtremIO

This is only possible with XtremIO in-memory metadata and Inline Data Reduction. Other flash products that implement VAAI but do not have inline deduplication still need to write the X-COPY to flash and deduplicate it later. Arrays that do not have in memory metadata need to carry out lookups on SSDs to perform the X-COPY, which negatively impacts I/O to existing active VMs. Only with XtremIO is this process completed quickly, with no SSD writes and with no impact to I/O on existing VMs.

The XtremIO features for VAAI support include:

- Zero Blocks/Write Same
    - Used for zeroing-out disk regions (VMware term: HardwareAcceleratedInit).
    - This feature provides accelerated volume formatting.
- Clone Blocks/Full Copy/XCOPY
    - Used for copying or migrating data within the same physical array (VMware term: HardwareAcceleratedMove).
    - On XtremIO, this allows VM cloning to take place almost instantaneously, without affecting user I/O on active VMs.
- Record based locking/Atomic Test & Set (ATS)
    - Used during creation and locking of files on a VMFS volume, for example, during powering-down/powering-up of VMs (VMware term: HardwareAcceleratedLocking).
    - This allows larger volumes and ESX clusters without contention.
- Block Delete/UNMAP/TRIM
    - Allows for unused space to be reclaimed, using the SCSI UNMAP feature (VMware term: BlockDelete; vSphere 5.x only).

XtremIO Management Server (XMS)

The XMS controls and manages the system, including:

- Forming, initializing, and formatting new systems
- Monitoring system health and events
- Monitoring system performance
- Maintaining a performance statistics history database (XMS keeps up to two years of historical data, providing rich reporting capabilities)
- Providing GUI and CLI services to clients
- Implementing volume management and data protection groups operation logic
- Maintaining (stopping, starting, and restarting) the system
- Providing intuitive provisioning workflows
- Providing a user-friendly reporting interface

The XMS is preinstalled with the CLI, GUI, and RESTful API interfaces. It can be installed on a dedicated physical server in the data center, or as a virtual machine on VMware.

The XMS must access management ports on both Storage Controllers of the first X-Brick in the cluster, and must be accessible by any GUI/CLI/REST client host machine. Since all communications use standard TCP/IP connections, the XMS can be located anywhere that satisfies the above connectivity requirements.

Since the XMS is not in the data path, it can be disconnected from the XtremIO cluster without affecting the I/O. An XMS failure only affects monitoring and configuration activities, such as creating and deleting volumes. However, when using a virtual XMS topology, it is possible to take advantage of VMware vSphere HA features to easily overcome such failures.

A single XMS can manage multiple clusters.[12] The XMS can manage clusters of varied sizes, models, and XtremIO version numbers. The key benefits of multiple cluster management are:

- From a management perspective, an administrator can manage multiple clusters all from a "single pane of glass."
- From a deployment perspective, only a single XMS server is required to manage multiple clusters.

With time, additional clusters can be added to a deployed XMS. In addition, a cluster can be easily moved from one XMS to another. All management interfaces (GUI/CLI/REST) offer inherent multi-cluster management capabilities.

System GUI

Figure 24 illustrates the relationship between the system GUI and other network components.

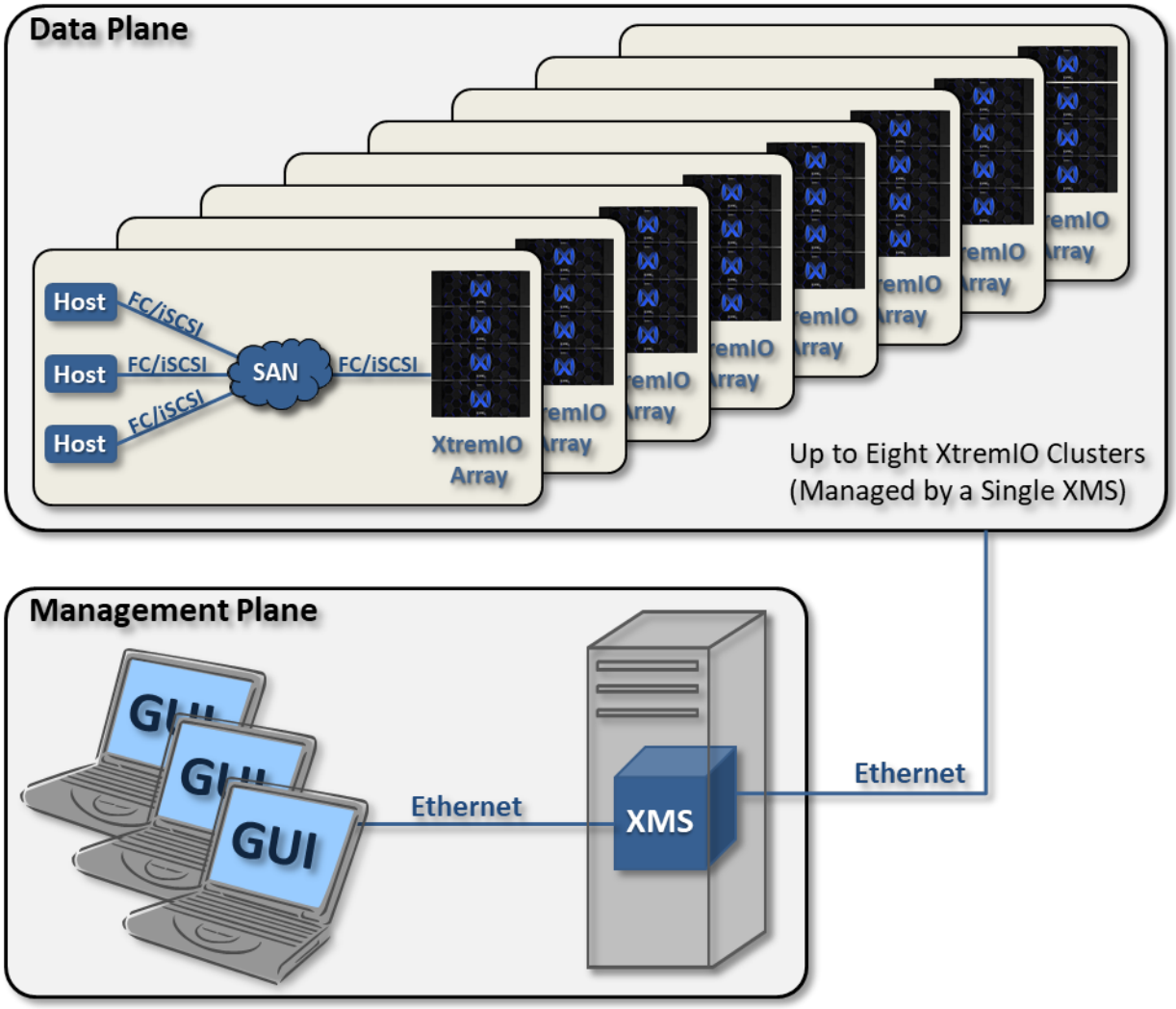


Figure 24. Relationship between GUI and other Network Components

[12] XIOS version 6.1 supports up to eight clusters managed by an XMS in each site. This will continue to increase in subsequent releases of the XtremIO Operating System.

The system GUI is HTML5-based web UI and accessible using a standard browser without additional installations required on the client side. The web UI client communicates with the XMS using standard TCP/IP protocols and can be used in any location that allows the client to access the XMS.

The GUI provides easy-to-use tools for performing most of the system operations (certain management operations must be performed using the CLI).

Figure 25 shows the GUI's Dashboard, enabling the user to monitor the system's storage, performance, alerts, and hardware status.

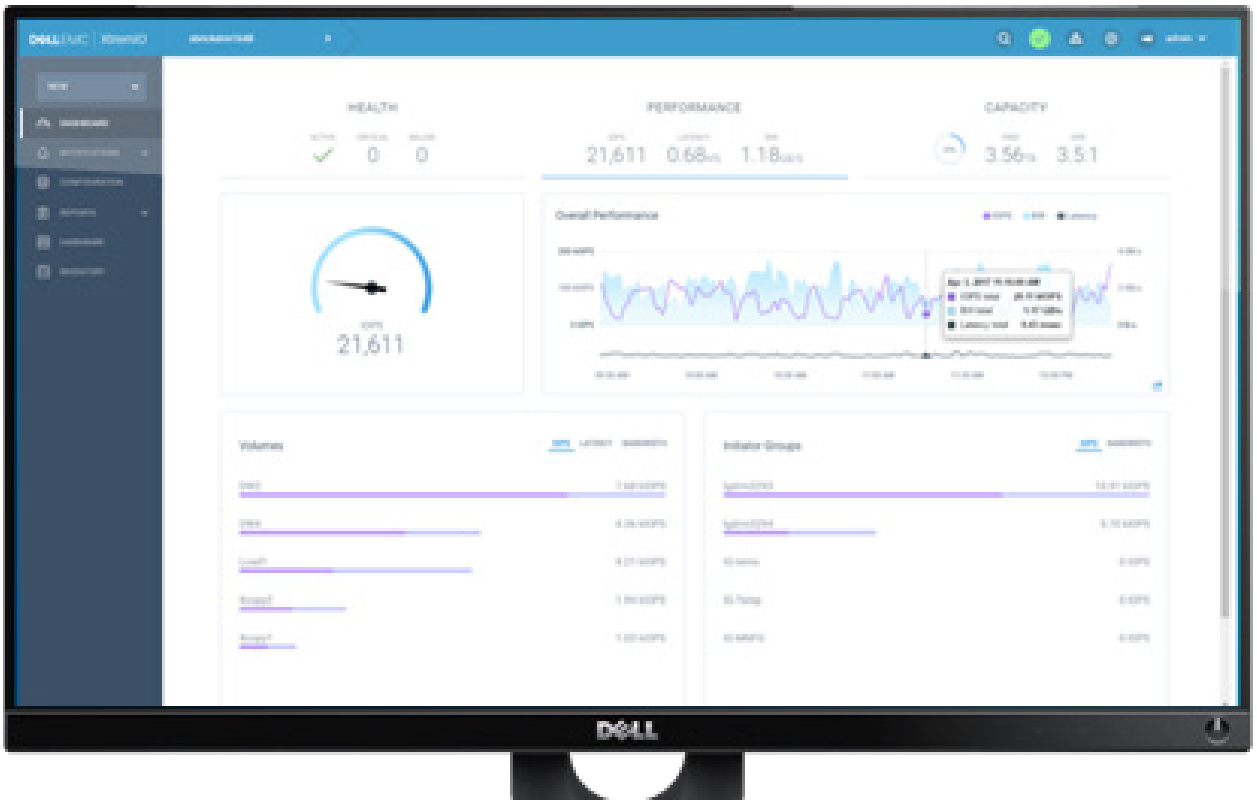


Figure 25. Monitoring the System Using the GUI

Command Line Interface

The system's Command Line Interface (CLI) allows administrators and other system users to perform supported management operations. It is preinstalled on the XMS and can be accessed using the standard SSH protocol.

To facilitate scripting from a remote host, it is possible to define a key-based SSH user access that does not require storing the password in the script and allows remote CLI access.

RESTful API

The XtremIO RESTful API allows HTTPS-based interface for automation, orchestration, query, and provisioning of the system. With the API, third-party applications can be used to control and fully administer the array. Therefore, it allows flexible management solutions to be developed for the XtremIO array.

PowerShell API

The XtremIO PowerShell API Module works in conjunction with Windows PowerShell console or ISE (Integrating Scripting Environment), and is available on computers running Windows.

PowerShell API provides a command line and scripting environment for automation, orchestration, query, and provisioning of a cluster or of multiple clusters. With this API, the customer can use third-party applications to control and fully administer the array. This allows the customer to develop flexible management solutions for the XtremIO Storage Array.

For details on how to install Windows PowerShell, refer to Windows PowerShell online documentation, hosted on Microsoft's Developer Network (https://msdn.microsoft.com/en-us/powershell/mt173057.aspx). The XtremIO PowerShell API Module is available for download on EMC support web site (https://support.emc.com/downloads/31111_XtremIO).

LDAP/LDAPS

The XtremIO X2 Storage Array supports LDAP users' authentication. Once configured for LDAP authentication, the XMS redirects users' authentication to the configured LDAP or Active Directory (AD) servers and allows access to authenticated users only. Users' XMS permissions are defined, based on a mapping between the users' LDAP/AD groups and XMS roles.

The XMS Server LDAP Configuration feature allows using single or multiple servers to authenticate external users for their login to the XMS server.

The LDAP operation is performed once when logging with external user credentials to an XMS server. The XMS server operates as an LDAP client and connects to an LDAP service, running on an external server. The LDAP Search is performed, using the pre-configured LDAP Configuration profile and the external user login credentials.

If the authentication is successful, the external user logs in to the XMS server and accesses the full or limited XMS server functionality (according to the XMS Role that was assigned to the LDAP user's group).

The XtremIO X2 Storage Array also supports LDAPS for secure authentication.

Ease of Management

XtremIO is very simple to configure and manage and there is no need for tuning or extensive planning.

With XtremIO, the user does not need to choose between different RAID options to optimize the system. Once the system is initialized, the XDP [see "XtremIO Data Protection (XDP)"] is already configured. All user data is spread across all X-Bricks. In addition, there is no tiering and performance tuning. All I/Os are treated the same. All volumes, when created, are mapped to all ports (FC and iSCSI) and there is no storage tiering in the array. This eliminates the need for manual performance tuning and optimization settings, and makes the system easy to manage, configure, and use.

XtremIO provides:

- Minimum planning
    - No RAID configuration
    - Minimal sizing effort for cloning/Snapshots
- No tiering
    - Single tier, all-flash array
    - No pools management
- No performance tuning
    - Independent of I/O access pattern, cache hit rates, tiering decisions, etc.

Replication of Dell EMC XtremIO to a Remote Array

Dell EMC RecoverPoint

Dell EMC RecoverPoint provides continuous data protection for comprehensive operational and disaster recovery. It supports Dell EMC XtremIO, Unity, VMAX, VNX, and major third-party arrays via VPLEX.

RecoverPoint benefits include:

- Continuous Data Protection for any Point in time to optimize RPO and TRO
- Recovery consistency for interdependent application
- Snap-Based replication (SBR) for XtremIO and Unity
- Reduction in WAN bandwidth consumption and optimal bandwidth utilization
- Multi-site support with "one to many" fan-out replication for higher protection and test operations, similar to the "many to one" fan-in replication for centralized DR site protecting for multiple branch offices.
- Snapshot-based replication support for XtremIO

  The native replication support for XtremIO is designed for high-performance and low-latency applications that provides a low RPO of one minute or less and immediate RTO.

  The benefits include:

  - Block-level remote or local replication
  - Asynchronous local and remote replication
  - Policy-based replication to enable optimizing storage and network resources while obtaining desired RPO and RTO
  - Application-aware integration
- Splitter-based replication using VPLEX

  RecoverPoint splitter-based replication provides synchronous replication, continuous replication with fine recovery granularity (journal based), and replication for active-active data centers.

  The benefits include:

  - Block-level remote or local replication
  - Dynamic synchronous, synchronous, or asynchronous remote replication
  - Policy-based replication to enable optimizing storage and network resources while obtaining desired RPO and RTO
  - Application-aware integration
- RecoverPoint for VMs

  RecoverPoint for VMs is a fully virtualized hypervisor-based replication solution built on a fully virtualized Dell EMC RecoverPoint engine.

  The benefits include:

  - Optimizing RPO/RTO for VMware environment at lower TCO
  - Streamlining OR and DR and increasing business agility
  - Equipping IT or service providers for cloud-ready data protection to deliver "Disaster Recovery as a Service" for private, public, and hybrid clouds

Dell EMC RecoverPoint Replication for XtremIO

Dell EMC RecoverPoint replication for XtremIO uses the "snap and replicate" option and is a replication solution for high-performance, low-latency environments. It leverages the best from both RecoverPoint and XtremIO, providing replication for heavy workload with a low RPO.

The solution is developed to support all XtremIO workloads, supporting all cluster types from a single X-Brick and up to eight X-Brick clusters, with the ability to scale out with the XtremIO Scale-Out option.

RecoverPoint replication for an XtremIO array is implemented by leveraging the array's content-aware capabilities. This allows efficient replication by only replicating the changes since the last cycle. In addition, it only leverages the mature and efficient BW management of RecoverPoint for maximizing the amount of I/O that the replication can support.

When RecoverPoint replication is initiated, the data is fully replicated to the remote site. RecoverPoint creates a snapshot on the source and transfers it to the remote site. The first replication is done by first matching signatures between the local and remote copies, and only then replicating the required data to the target copy.

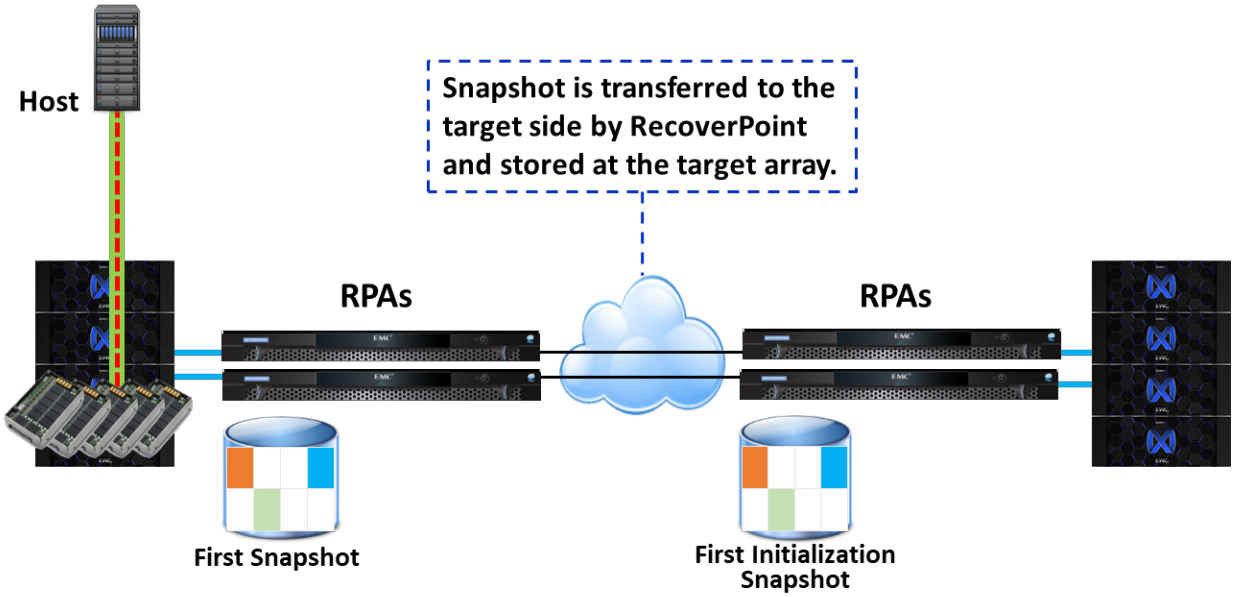


Figure 26. RecoverPoint "Snap and Replicate" Option—Initial Replication

For every subsequent cycle, a new snapshot is created and RecoverPoint replicates just the changes between the snapshots to the target copy and stores the changes to a new snapshot at the target site.

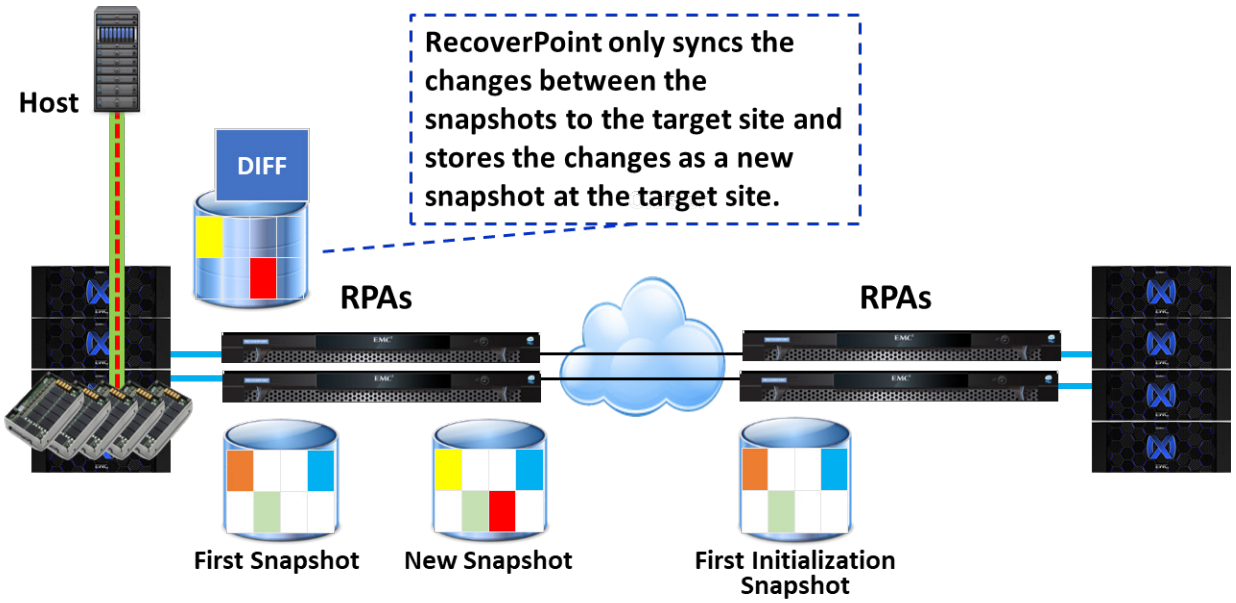


Figure 27. RecoverPoint "Snap and Replicate" Option—Subsequent Replications

The snapshots on the target are kept according to the retention policy and can be used for DR testing and for failover to the target copy.

RecoverPoint replication for XtremIO offers unique and superior values, including:

- Bi-directional replication
- XtremIO to XtremIO replication
- Heterogeneous replication between XtremIO to VPLEX, VMAX, and VNX arrays
- All disaster recovery operations
- Full integration with Dell EMC and VMware Ecosystems
- Support for XtremIO full scale and performance
- Simple management and configuration from a single console
- Ability to failover, recover production, and test copy with immediate RTO
- Space efficiency and fast data synchronization by leveraging XtremIO lightweight Snapshots
- Policy-based local and remote replication management
- Flexible protection window and snapshots retention management
- Automatic provisioning of journal volumes
- Automatic provisioning of replica volumes

Synchronous and CDP Replication for XtremIO

Synchronous replication and CDP is supported with the VPLEX splitter solution.

PowerPath, VPLEX, RecoverPoint, and XtremIO can be integrated together[13] to offer a strong, robust, and high-performing block storage solution.

- **PowerPath**: Installed on hosts to provide path failover, load balancing, and performance optimization of VPLEX engines (or directly to the XtremIO array if VPLEX is not used).
- **VPLEX Metro:** Allows sharing storage services across distributed virtual volumes and enables simultaneous read and write access across metro sites and across array boundaries.
- **VPLEX Local:** Used at the target site, virtualizes both Dell EMC and non-Dell EMC storage devices, leading to better asset utilization.
- **RecoverPoint/EX:** Any device encapsulated by VPLEX (including XtremIO) can use the RecoverPoint services for asynchronous, synchronous, or dynamic synchronous data replication.

For example: An organization has three data centers in New Jersey, New York City, and Iowa, as shown in Figure 28

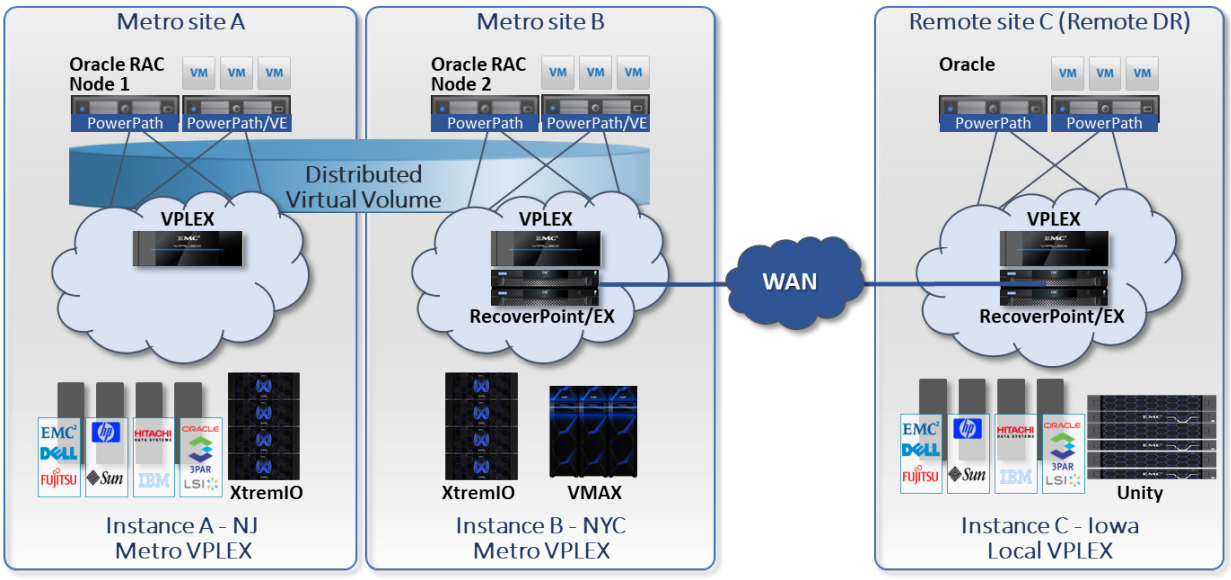


Figure 28. Integrated Solution using XtremIO, PowerPath, VPLEX, and RecoverPoint

Oracle RAC and VMware HA nodes are dispersed between the NJ and NYC sites, and data is moved frequently between all sites.

[13] RPQ approval is required. Please contact your EMC representative.

The organization has adopted multi-vendor strategy for their storage infrastructure:

- XtremIO storage is used for the organization's VDI and other high performing applications.
- VPLEX Metro is used to achieve data mobility and access across both NJ and NYC sites. VPLEX metro provides the organization with Access-Anywhere capabilities, where virtual distributed volumes can be accessed in read/write at both sites.
- Disaster recovery solution is implemented by using RecoverPoint for asynchronous continuous remote replication between the metro site and the Iowa site.
- VPLEX metro is used at the Iowa site to improve assets and resource utilization, while enabling replication from Dell EMC to non-Dell EMC storage.

Dell EMC solutions (such as those in the above example) offer unique and superior values, including:

- High availability and performance optimization of multiple paths in a high performing storage environment
- High-performance content-aware all flash storage that supports hundreds of thousands of IOPS with low latency and high throughput
- Geographically dispersed clusters with zero RPO
- Automated recovery with near-zero RTO
- High availability within and across VPLEX Metro data centers
- Increased performance as workloads can be shared between sites
- Continuous remote replication (or CDP or CLR) of XtremIO systems

Integration with other Dell EMC Products

XtremIO is well integrated with other Dell EMC products. Integration points will continue to expand in subsequent XtremIO releases to offer additional value for Dell EMC customers.

System Integration Solutions

Dell EMC VxBlock

VxBlock Systems provide a wide range of solutions to meet your requirements for size, performance, and scalability. VxBlock Systems are built with industry-leading compute and networking from Cisco, storage from Dell EMC with Dell EMC Unity, XtremIO, and VMAX, and virtual distributed switching from VMware. The VxBlock System 540 is optimized for data reduction and copy-friendly workflows, such as Virtual Desktop Infrastructure (VDI) and test and development environments. Leveraging XtremIO storage, with in-line efficiencies, these systems deliver Scale-Out performance at ultra-low latency.

More information on VxBlock solutions is available at: https://www.emc.com/collateral/data-sheet/vxblock-product-overview.pdf

Dell EMC VSPEX

VSPEX proven infrastructure accelerates the deployment of private cloud and VDI solutions with XtremIO. Built with best-of-breed virtualization, server, network, storage and backup, VSPEX enables faster deployment, more simplicity, greater choice, higher efficiency, and lower risk. Validation by Dell EMC ensures predictable performance and enables customers to select products that leverage their existing IT infrastructure, while eliminating planning, sizing, and configuration burdens.

More information on VSPEX solutions is available at: https://support.emc.com/products/30224_VSPEX

Management and Monitoring Solutions

Dell EMC Storage Analytics (ESA)

ESA links VMware vRealize Operations Manager (vR OPs Manager) for storage with a Dell EMC adapter. The vR OPs Manager displays performance and capacity metrics from storage systems with data that the adapter provides by:

- Connecting to and collecting data from storage system resources
- Converting the data into a format that vR OPs Manager can process
- Passing the data to the vR OPs Manager collector

vR OPs Manager presents the aggregated data through alerts, dashboards, and in predefined reports that end users can easily interpret. Dell EMC adapter is installed with the vR OPs Manager administrative user interface. ESA complies with VMware management pack certification requirements and has received the VMware Ready certification.

More information on ESA is available at: https://support.emc.com/products/30680_Storage-Analytics

Dell EMC Enterprise Storage Integrator (ESI) Plugin for Windows

Dell EMC Enterprise Storage Integrator (ESI) for Windows Suite is a set of tools for Microsoft Windows and Microsoft applications administrators. The ESI plugin consist of an ESI Management Pack for Microsoft System Center Operations Manager (SCOM), an ESI PowerShell Toolkit, and adapters for storage management and application integration.

Advantages of the ESI plugin for Windows include the following:

- The ESI PowerShell Toolkit provides ESI storage provisioning and discovery capabilities with corresponding PowerShell cmdlets, for XtremIO and other Dell EMC storage arrays.
- In addition to supporting physical environments, ESI supports storage provisioning and discovery for Windows virtual machines that are running on Microsoft Hyper-V.
- ESI SCOM Management Packs enable you to monitor the health of your XtremIO system, along with other Dell EMC storage systems, with SCOM, by providing consolidated and simplified dashboard views of storage entities.
- ESI SQL Server adapter allows the viewing of local and remote Microsoft SQL Server instanced/databases, and mapping the databases to Dell EMC storage.

ESI plugin for Windows is a free software and can be downloaded from:
https://support.emc.com/products/17404_ESI-for-Windows-Suite

Dell EMC ViPR Controller

Dell EMC ViPR Controller is a software-defined storage platform that abstracts, pools, and automates a data center's underlying physical storage infrastructure. It provides data center administrators with a single control plane for heterogeneous storage systems.

ViPR enables software-defined data centers by providing the following features:

- Storage automation capabilities for multi-vendor block and file storage environment
- Management of multiple data centers in various locations with single sign-on data access from any data center
- Integration with VMware and Microsoft compute stacks to enable higher levels of compute and network orchestration
- Comprehensive and customizable platform reporting capabilities that include capacity metering, chargeback, and performance monitoring through the included ViPR SolutionPack

More information on ViPR Controller is available at: https://support.emc.com/products/32034_ViPR-Controller

Dell EMC ViPR SRM

Dell EMC ViPR SRM provides comprehensive monitoring, reporting, and analysis for heterogeneous block, file, and virtualized storage environments. It enables the users to visualize applications to storage dependencies, monitor and analyze configurations and capacity growth, and optimize their environment to improve return on investment.

Virtualization enables businesses of all sizes to simplify management, control costs, and guarantee uptime. However, virtualized environments also add layers of complexity to the IT infrastructure that reduce visibility and can complicate the management of storage resources. ViPR SRM addresses these layers by providing visibility into the physical and virtual relationships to ensure consistent service levels.

More information on ViPR SRM is available at: https://support.emc.com/products/34247_ViPR-SRM

Virtual Storage Integrator (VSI) Plugin for VMware vCenter

VSI plugin is a free vSphere web client plugin that enables VMware administrators to view, manage, and optimize storage for their ESX/ESXi servers. It consists of a graphical user interface and the Dell EMC Solutions Integration Service (SIS), which provides communication and access to XtremIO arrays.

The VSI plugin allows the users to interact with their XtremIO array from a vCenter perspective. For example, the user can provision VMFS datastores and RDM volumes, create full clones using XtremIO Snapshots, view properties of datastores and RDM volumes, extend datastore capacity, and do bulk provisioning of datastores and RDM volumes.

In addition, the VSI plugin allows performing the following tasks for XtremIO:

- Setting host parameters to recommended values, including Multipathing, disk queue depth, maximum I/O size, and other settings. If needed, these settings can also be performed at the cluster level.
- Optimizing settings for VAAI and other ESX operations.
- Space reclamation at the datastore level, providing the ability to schedule the space reclamation operations to run at fixed times.
- Integration with VMware Horizon View and Citrix XenDesktop.
- Reporting consumed capacity from a VMware and XtremIO perspective.

The VSI plugin can be downloaded from: https://support.emc.com/products/32161_VSI-Plugin-Series-for-VMware-vCenter

Application Integration Solutions

Dell EMC AppSync

Dell EMC AppSync is a simple, service-level agreement (SLA) driven and self-service data protection and repurposing application for XtremIO environment. With AppSync, it is possible to protect all critical applications in a single click, dial in the correct service level, and enable application owners to drive protection. In addition, AppSync is typically helpful with any copy management activity, like data repurposing for test/dev, backup acceleration using snapshots, or operational recovery, allowing users to easily build up, refresh and restore application instances on the fly.

AppSync allows application administrators to make application copies, using XtremIO Snapshots in an application-consistent manner. It also allows taking, deleting, refreshing, and restoring application-consistent Snapshots according to a pre-defined schedule and to subscribe applications to a "service plan". AppSync provides integration for VMware, Oracle, SQL Server, and Exchange environments.

More information on AppSync is available at: www.emc.com/AppSync

A free Starter License is also included with all X2 arrays.

Business Continuity and High Availability solutions

Dell EMC PowerPath

Dell EMC PowerPath is a host-based software that provides automated data path management and load balancing capabilities for heterogeneous servers, network, and storage deployed in physical and virtual environments. It enables users to meet service levels with high application availability and performance. PowerPath automates path failover and recovery for high availability in case of error or failure and optimizes performance by load balancing I/Os across multiple paths. XtremIO is supported under PowerPath both directly and by virtualizing the XtremIO system using VPLEX.

Dell EMC VPLEX

The Dell EMC VPLEX family is the next-generation solution for data mobility and access within, across, and between data centers. The platform enables local and distributed federation.

- Local federation provides transparent cooperation of physical elements within a site.
- Distributed federation extends access between two locations across distance.

VPLEX removes physical barriers and enables users to access a cache-coherent, consistent copy of data at different geographical locations and to geographically stretch virtual or physical host clusters. This enables transparent load sharing between multiple sites while providing the flexibility of relocating workloads between sites in anticipation of planned events. Furthermore, in case of an unplanned event that could cause disruption at one of the data centers, failed services can be restated at the surviving site.

VPLEX supports two configurations, local and metro. In the case of VPLEX Metro with the optional VPLEX Witness and Cross-Connected configuration, applications continue to operate in the surviving site with no interruption or downtime. Storage resources virtualized by VPLEX cooperate through the stack, with the ability to dynamically move applications and data across geographies and service providers.

XtremIO can be used as a high performing pool within a VPLEX Local or Metro cluster. When used in conjunction with VPLEX, XtremIO benefits from all the VPLEX data services, including host operating system support, data mobility, data protection, replication, and workload relocation.

OpenStack Integration

OpenStack is the open platform for managing private and public clouds. It allows storage resources to be located anywhere in the cloud and available for use upon demand. Cinder is the block storage service for OpenStack.

The XtremIO Cinder driver enables OpenStack clouds to access XtremIO storage. The XtremIO Cinder management driver directs the creation and deletion of volumes on the XtremIO array and attaches/detaches volumes to/from instances/VMs created by OpenStack. The driver automates the creation of initiator mappings to volumes. These mappings allow the running of OpenStack instances to access the XtremIO storage. This is all performed on demand, based on the OpenStack cloud requirements.

The OpenStack XtremIO Cinder driver utilizes the XtremIO RESTful API to communicate OpenStack's management requests to the XtremIO array.

The OpenStack cloud can access XtremIO using either iSCSI or Fibre Channel protocols.

Conclusion

XtremIO offers the market an advanced revolutionary architecture, optimized for all-SSD enterprise storage subsystems. XtremIO offers a rich set of features that leverage and optimize SSD capabilities and have been especially designed to provide unparalleled solutions for enterprise customers' needs and requirements.

XtremIO features include truly scalable and flexible expansion options (buy additional capacity and performance when needed), high performance with hundreds of thousands of IOPS, constant sub-millisecond low latency, content-aware Inline Data Reduction, high availability, thin provisioning, XtremIO Virtual Copies (Snapshots), Metadata-Aware Asynchronous Replication, and VAAI support. XtremIO also offers a unique patent-protected scheme that leverages the SSD characteristics to provide an efficient and powerful data protection mechanism which can protect the data against two simultaneous and multiple consecutive failures. In addition, XtremIO incorporates a comprehensive, intuitive and user-friendly interface which includes GUI, RESTful API, PowerShell API, and command line modes, and is designed for ease-of-use while enabling efficient system management.

XtremIO provides the perfect solution for all-SSD enterprise SAN storage while offering a superior total cost of ownership solution for its customers.

How to Learn More

For a detailed presentation explaining XtremIO X2 Storage Array's capabilities and how XtremIO X2 substantially improves performance, operational efficiency, ease-of-use and total cost of ownership, please contact XtremIO X2 at XtremIO@emc.com. We will schedule a private briefing in person or via a web meeting. XtremIO X2 provides benefits in many environments and mixed workload consolidations, including virtual server, cloud, virtual desktop, database, analytics and business applications.

Learn more about Dell EMC XtremIO

Contact a Dell EMC Expert

View more resources

Join the conversation @DellEMCStorage and #XtremIO