A SOLIDFIRE COMPETITIVE COMPARISON

NetApp SolidFire and EMC XtremIO Architectural Comparison

This document provides an overview of EMC's XtremIO architecture as it compares to NetApp SolidFire. Not intended to be exhaustive, this overview covers select elements where the solutions differ and presents their impact to overall suitability for data center needs.

Overview

EMC XtremIO is an all-flash, scale-out storage solution that utilizes traditional SAN controllers (nodes) in an active/active dual-controller design, connected to a 40Gb InfiniBand fabric. EMC combines into a single building block called an XtremIO X-Brick with the following components: both active/active controllers, a shared drive array enclosure (DAE) containing 25 enterprise multi-layer cell (eMLC) SSDs, and, depending on the configuration, either one or more battery backup units and/or InfiniBand fabric switches. X-Bricks come in 10TB, 20TB, or 40TB capacities depending on the drive size configuration. XtremIO clusters are scaled out by deploying pairs of X-Bricks onto an InfiniBand connected cluster.

The choice of an InfiniBand fabric and active/active controller pairs enables an XtremIO cluster to follow traditional block-based SAN design parameters, access the maximum amount of performance from each controller pair, and still scale out both performance and capacity across the entire cluster. Running across the entire cluster is XtremIO's XIOS operating system, which handles all cluster operations and functionality from a software level.

In addition to scale-out, XtremIO all-flash clusters offer low-latency and strong overall performance. Like most performance-centric designs, trade-offs resulting in some limitations have to be expected in the areas of scale, drive selection, and cluster functionality.

This document compares the architectural elements of the EMC XtremIO solution with NetApp SolidFire all-flash solutions and assesses their suitability for the needs of the Next Generation Data Center (NGDC). NetApp recommends evaluating all-flash storage solutions based on application requirements and offers a portfolio of solutions tailored to different environments, including: SolidFire SF-series systems with Element OS software; All Flash FAS systems with ONTAP software; and EF-Series systems with SANtricity software.

Findings

- **Scale-Out** - XtremIO clusters offer the ability to scale out both performance and capacity together, but XtremIO's granularity of scale is much heavier than SolidFire's. A new XtremIO deployment can be deployed as a single X-Brick and subsequently scaled out to a second X-Brick; however, once two X-Bricks are deployed, additional scale must occur in pairs of X-Bricks (four controllers and 50 SSDs) meaning the cost to scale is very high. With the recently announced XIOS v4.0, scale is currently limited to eight X-Bricks (16 controllers) per cluster. Each X-Brick in the cluster must be the same size and generation.

  Conversely, SolidFire's distributed architecture allows for linear scale-out of performance and capacity into distinct virtual pools that can be provisioned separately to workloads, thanks to SolidFire Quality of Service (QoS). In addition, through SolidFire's single-node granular scalability, mixed-node support, and the ability to add or remove nodes to/from a cluster, SolidFire enables enterprises to cost-effectively support specific solutions and adapt on the fly to multiple workload environments without affecting the performance of existing applications.

- **Guaranteed Performance** - A key requirement of the Next Generation Data Center is to have an environment with predictable performance and the ability to ensure that performance is always available to tens, hundreds, or thousands of applications. XtremIO solutions offer good speed with low latency; however, there is no ability to control performance or specify QoS for individual volumes. Lack of QoS limits the number of workloads an XtremIO can reliably support and sets up the likelihood that applications and users will experience inconsistent performance in multiple-mixed-workload environments.

  Uniquely, SolidFire enables enterprises to specify and guarantee minimum, maximum, and burst IOPS for individual storage volumes dynamically and independent of capacity. SolidFire's architecture-specific QoS eliminates the "noisy neighbor" problem in multiple workload environments, guarantees performance to all applications, and enables support for thousands of concurrent mixed workloads.

- **Automation** - Both SolidFire and XtremIO have REST APIs for automating storage management. Only SolidFire offers the ability to automate every storage function of the array through the API including disaster recovery and failover with SolidFire Helix.

- **Data Assurance** - To ensure data availability and durability, the XtremIO cluster utilizes traditional dual-parity RAID (called XDP), which stripes data across a 25-drive RAID group. This results in a very low 8% RAID overhead but significantly impacts cluster performance when a drive fails; rebuild times are much longer than SolidFire Helix rebuild times. XDP does not provide for node failure recovery. In the event of a node failure, the XtremIO cluster is in a single point of failure and a significantly reduced performance state until the failed node is physically replaced.

  In place of traditional RAID protection, SolidFire employs patent-pending SolidFire Helix to provide protection against data loss resulting from a hardware failure. Rather than using a traditional RAID stripe across multiple drives, Helix ensures each unique block has a redundant copy stored on a separate drive in a separate node within the cluster. Helix provides data durability without impeding the linear scale out of capacity and performance and enables automated self-healing in the event of a drive or node failure. With RAID architectures, adding storage nodes or rebuilds following a failure are typically measured in hours or days, whereas with SolidFire, expansions or rebuilds take place in minutes and actually occur faster as clusters grow.

SolidFire vs XtremIO similarities

Data addressing

Both XtremIO and SolidFire use content addressing, where a given block of data is addressed in metadata by its content rather than a LUN and LBA. Because of this, a piece of data can be managed and moved freely throughout the system with significantly less metadata management overhead. Because of its foundational ties to the content of the blocks, the content addressing architecture naturally supports global deduplication.

Scale

Both SolidFire and XtremIO are designed for scaling out rather than scaling up. While still utilizing controllers, XtremIO's XIOS enables the performance and capacity of each controller to be shared across the cluster. With XIOS v4, XtremIO can scale out to 16 storage controllers (eight X-Bricks) in a single cluster. However, scale must occur in increments of two X-Bricks at a time, which equates to a granularity of four controllers and 50 SSDs. Scale at this granularity results in large CAPEX acquisitions and often much more infrastructure than what is actually needed. In addition, X-Bricks of different capacities, different generations of hardware, and even different operating system versions cannot be deployed together in a single cluster. This can result in stranded resources, silos of storage, and costly data migrations.

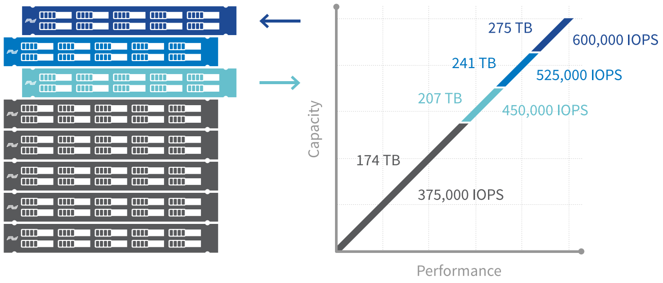


Figure 1: SolidFire Mixed-Node Scale-Out
At any point during or after deployment, nodes can be added, removed, or replaced to increase capacity and/or performance, without impacting existing workloads. As nodes are added, their capacity and IOPS are aggregated into the total provisionable capacity and performance available for assignment to any existing or new volumes.

SolidFire's distributed architecture and Helix technology enables the linear scaling of available capacity and performance as granularly as a single (1U) node at a time. As nodes are added, their capacity and IOPS are aggregated into the total provisionable capacity and performance available for assignment to any existing or new volumes. SolidFire completely supports mixed-node (size and/or generation) cluster interoperability, and at any point during or after deployment, nodes can be added, removed, or replaced to increase capacity and/or performance without impacting existing workloads. Conversely, if a node fails or is removed from the cluster, Helix automatically initiates a rebuild, restoring redundancy for any copies that were residing on the unavailable node.

SolidFire vs XtremIO primary differences

Data durability

Unlike SolidFire, XtremIO arrays utilize more expensive eMLC SSDs and rely on a dual-parity RAID protection scheme called XDP for data durability that employs very wide 23+2 RAID stripes. The wide RAID stripe that XtremIO utilizes enables EMC to advertise an industry-leading space efficiency of 8% RAID overhead for data durability. XtremIO's data durability overhead may be low, but drive rebuild times in the event of a failure will be very high compared to SolidFire. XDP provides no automated recovery in the event of a node failure. In both failure scenarios, performance of the XtremIO cluster is reduced significantly until the failed drive is rebuilt or the failed controller is physically replaced. Details of why XtremIO uses eMLC drives will be explained in more depth in the next section, but for now, understand that eMLC drives in an XtremIO cluster are not any more reliable than the cMLC drives in a SolidFire cluster.

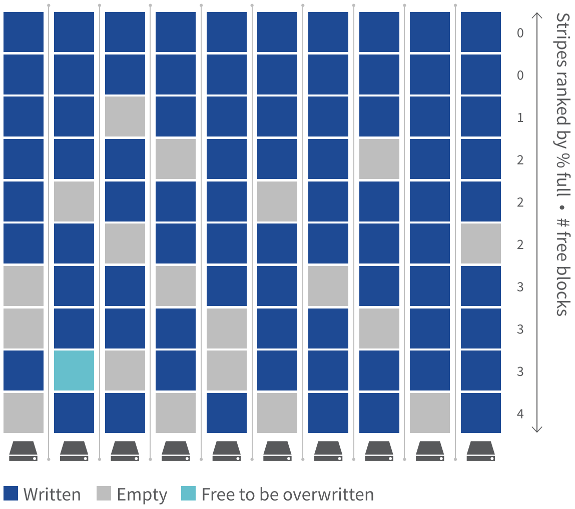


Figure 2: XtremIO Fixed-Block Approach
Once the drive is fully written, stripes with available space are identified, and old data that is no longer valid is read along with the associated parity blocks. The new data is written into that space along with an update to the parity blocks.

Data Durability within a SolidFire cluster starts with patent-pending SolidFire Helix, which automatically spreads a copy of all data across the global cluster. The advantage of Helix is that in the event of a drive failure, only the data that was written to the drive is rebuilt. The rebuild process is completely automated and is as simple as (1) finding the redundant copy of the data from the failed drive on the drives in the cluster, (2) making a copy of the (now primary) data, and (3) distributing the copies across the rest of the drives in the cluster.

SolidFire's Helix method of recovery reduces rebuild times from hours to minutes, does not put undue stress on the other drives in the cluster, incurs very little performance impact to the cluster, and works not only with drives but for nodes as well. If a SolidFire node were to fail, the exact same process outlined for a drive failure occurs on a larger scale for the affected node. The node rebuild process, just like the drive rebuild, is completely automated, happens in approximately an hour, and does not significantly impact performance.

In fact, the more a cluster scales out, the faster the rebuild process completes. This is because an equal number of CPU cycles are borrowed from each node in the cluster to help with the rebuild process, and more nodes results in fewer CPU cycles per node.

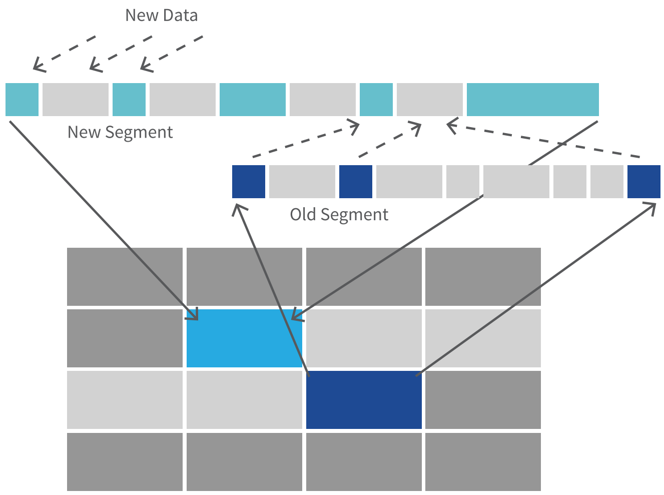


Figure 3: SolidFire Log-Structured Approach
Datasets of varying sizes are aggregated into larger segments and written down in a continuous linear fashion, much like a log file. When a drive has been fully written, the process of recycling the disk space is done in a similar fashion: The valid data on disk is read in from partially empty segments (or segments that contain old, nonvalid data). The good data then combines with the incoming data stream, rewriting itself into a new segment.

Fixed-block vs. Log-structured

XtremIO's architecture utilizes a fixed-block approach, which assigns each block of data a unique fingerprint that determines where that data will be written on the array and ensures that writes are spread evenly across the cluster.

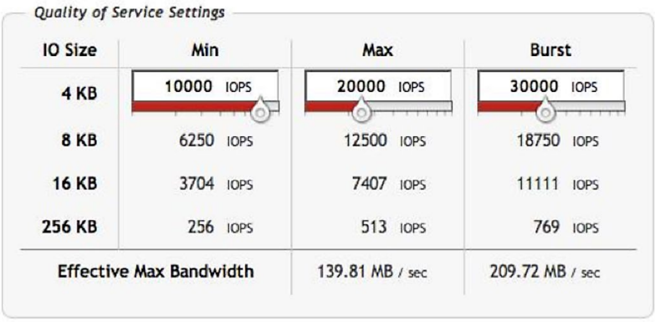
Using the fixed-block approach, once the array begins to fill up and fewer full stripes are available, the stripes with the most available space are identified and ranked. New data is then written into the stripe with the most available space. In this architecture, existing data is not typically moved; old data that is no longer valid is simply overwritten with new data in place. This approach of building stripes with data coming in from anywhere in the cluster and writing to the most lightly used stripe is the key to the architecture's optimization. It is also what enables XtremIO to minimize read/write amplification typically encountered with RAID; however, there is still an amplification factor of 1.22 reads and 1.22 writes to update a given block of data. To mitigate the performance impact of the fixed-block approach and the recycling of empty space across drives, XtremIO arrays must use more expensive enterprise-grade eMLC drives with significantly more capacity overprovisioning (28% vs. 7%) than the cMLC drives SolidFire uses. Also important to note is that the 28% eMLC capacity overhead is used strictly for garbage collection purposes and is unavailable to customer applications.

In SolidFire's log-structured approach, the initial write to disk is done in a fashion where datasets of varying sizes are aggregated into larger segments and written down in a continuous linear fashion, much like a log file. When a drive has been fully written, the process of recycling the disk space is done in a similar fashion: The valid data on disk is read in from partially empty segments (or segments that contain old, nonvalid data). The good data combines with the incoming data, stream rewriting itself into a new segment. In this way, SolidFire is able to efficiently manage the writes to the SSD and utilize less expensive consumer-grade MLC drives, achieving enterprise-class data durability without the added cost

Power loss handling

Power loss handling is another area where SolidFire and XtremIO differ, primarily stemming from the type of media utilized as an intermediary prior to the final write of data to disk.

In XtremIO's case, DRAM is used to store incoming writes, combined with dual battery backup units. Incoming writes and metadata updates are stored in DRAM before replication (very quickly over InfiniBand) to a controller on a second node. Data is then aggregated in memory into RAID stripes and written to disk in the background.



Quality of Service Settings

IO Size	Min	Max	Burst
4 KB	10000 IOPS	20000 IOPS	30000 IOPS
8 KB	6250 IOPS	12500 IOPS	18750 IOPS
16 KB	3704 IOPS	7407 IOPS	11111 IOPS
256 KB	256 IOPS	513 IOPS	769 IOPS
Effective Max Bandwidth		139.81 MB / sec	209.72 MB / sec

Figure 4: SolidFire QoS
SolidFire architecture allows users to set minimum, maximum, and burst IOPS on a per-volume basis.

Each controller is connected to a UPS with dual battery backups that flushes DRAM to the array's drives in the event of a power failure. This method of flushing metadata from DRAM to the array's standard drives rather than from dedicated metadata drives, results in an additional 10% capacity overhead, for a total of 18% capacity overhead between metadata and RAID.

In SolidFire's approach, each node has 8 GB of PCIe NVRAM. The card includes a super capacitor that keeps the DRAM powered, and a small amount of SLC flash. In the event the PCI card loses power, the data gets flushed into the flash and loads back into DRAM upon the restoration of power, Command data and metadata updates are then written in NVRAM on two nodes before acknowledgment. Consequently, SolidFire is able to achieve redundancy across the NVRAM and nodes and maintain power-loss protection in the architecture.

Comparing the approaches, DRAM has the highest throughput and lowest latency. PCIe is fast because it is going into DRAM after traversing the PCI bus. So it still provides very high throughput and low latency but does not rely on the UPS for data integrity in the event of a power loss.

QoS

XtremIO all-flash systems are clearly fast, with consistently low latency like most tightly coupled architectures. However, XtremIO provides no QoS provisioning or performance control to ensure applications in mixed workload deployments consistently get the IOPS they need.

To deliver predictable and guaranteed storage performance, SolidFire leverages QoS performance virtualization of resources.

Patented by SolidFire, this technology permits the management of storage performance independent from storage capacity. SolidFire's architecture allows users to set minimum, maximum, and burst IOPS on a per-volume basis. Because performance and capacity are managed independently, SolidFire clusters are able to deliver predictable storage performance to thousands of applications within a shared infrastructure.

The bottom line

XtremIO offers an all-flash scale-out solution focused on simplicity of development and latency performance at deployment. While, by definition, XtremIO is a scale-out architecture, it is built upon the controller-centric model which introduces complications to scale, limits the array to a more traditional RAID architecture, and requires the use of more expensive eMLC drives.

XtremIO systems are well-suited for point-solution environments, but due to the inability to provision QoS, they're less than optimal in the areas of scale, automation, guaranteed performance, and agility for next generation data center-type applications, including large-scale, multiple/mixed workload, and infrastructure as a service (ITaaS) deployments.