SEAGATE

Flash Acceleration of HPC Storage

A flash-based storage tier between the compute cluster and a slower disk-based storage tier can provide a faster storage resource to the supercomputer.

There are, however, many ways to deploy flash storage that are all reliant on intelligent software that manage the I/O workflow. Initially, mid-tier solutions gained first-to-market awareness – however, technology advances with server-side flash storage are now answering the demands of many traditional HPC workloads, especially those with random, mixed I/Os.

This white paper explores those new technology options.

Burst Buffers … Or is there a better way?

There has been a lot of discussion over the last few years about how to use SSDs and NVMe to accelerate HPC workloads. The logic is fairly sound as both SAS-based flash (SSDs) and PCIe- based storage (NVMe) have several performance-related advantages over traditional storage based on spinning disks.

However, in many cases, this is based on technical data such as:

- Much higher IOPS than HDDs
  - 10,000+ vs ~150 – 500 for SAS HDDs
- Much higher throughput
  - 1,000+ MB/s vs 250 MB/s for standard enterprise SAS HDDs
- No moving parts
  - HDDs has spinning disks and an actuator arm moving the heads over the surface of the disks
- Faster connectivity (in the case of NVMe)
  - PCIe direct attach is much faster than SAS
- Expected price parity with HDDs in the near future
  - Based on consumer flash devices

But for the most part, these assumptions are based on individual drives and not as components of a parallel file system which put different requirements on the drives.

All-Flash Storage

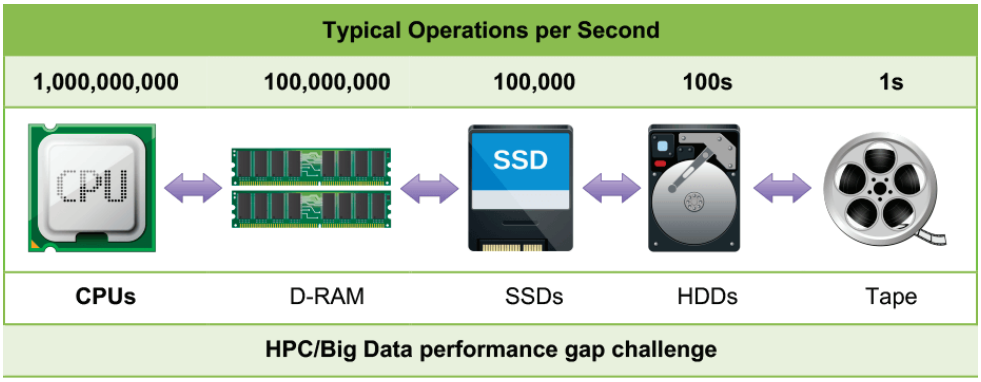


Figure 1 – Relative comparisons of theoretical number of I/O operations for different components.

More and more HPC applications fall into the growing category of “data intensive applications.” Basically, this means that the applications can generate more results with an increase in data availability. For HPC storage systems, this means two things: higher aggregated throughput for streaming data, and ability to deliver more IOPS for random access to smaller amounts of data. While it is possible to deliver HPC data storage systems today that can deliver 2,000 GB/s or more, a number of projects require much higher throughput for deployment in 2018 – 2020.

In addition, where CPUs and computer memory keep increasing their I/O capabilities with every new generation whereas non-volatile storage technologies have focused more on capacity, there’s a significant gap between the different capabilities (see Figure 1).

Current SSD technology does seem to fit well within the data framework often seen in data centers around the world and could easily fill the gap in a perfect world. However, there are a number of issues with this approach:

- Traditional flash caching solutions available today are designed for client-side caching to pool the small blocks of application I/O but not for large sequential I/O to disk-based storage.
- Separate storage tiers mean data migration that is often so costly (from a time point of view) that it totally negates the value.
- The cost is still, and will continue to be, 5-7x higher than spinning drives, making all-flash systems financially difficult to support.
- The data retention capabilities are very low on capacity flash (eMLC has a reliable data retention of 3-4 months and the newer tri-layer cells have even lower retention).
- While flash is capable of much higher throughput performance than HDDs, SSDs do not handle large streaming I/Os well.

While an all-flash-based HPC storage system has a lot of appeal from a performance point of view, it is not cost-efficient, nor has it the reliability to store important data over time. Furthermore, the issues with ingesting large streaming data flows, common in handling big data and HPC, makes an all-flash array less useful. But modern high-performance systems are expected to handle workloads under the control of job schedulers, requiring the ability to handle mixed I/O (the combination of streaming, transactional, small random and large sequential).

The solution to this problem is obviously to use SSDs to handle part of the workload for which it's best suited. But this should be done without having to use separate storage tiers forcing data movement, without requiring users to re-write or re-compile applications, or use complicated policy engines to handle the workflows. Basically, it should be transparent to the user, the application and even the file system of choice.

Introducing Nytro Intelligent I/O Manager

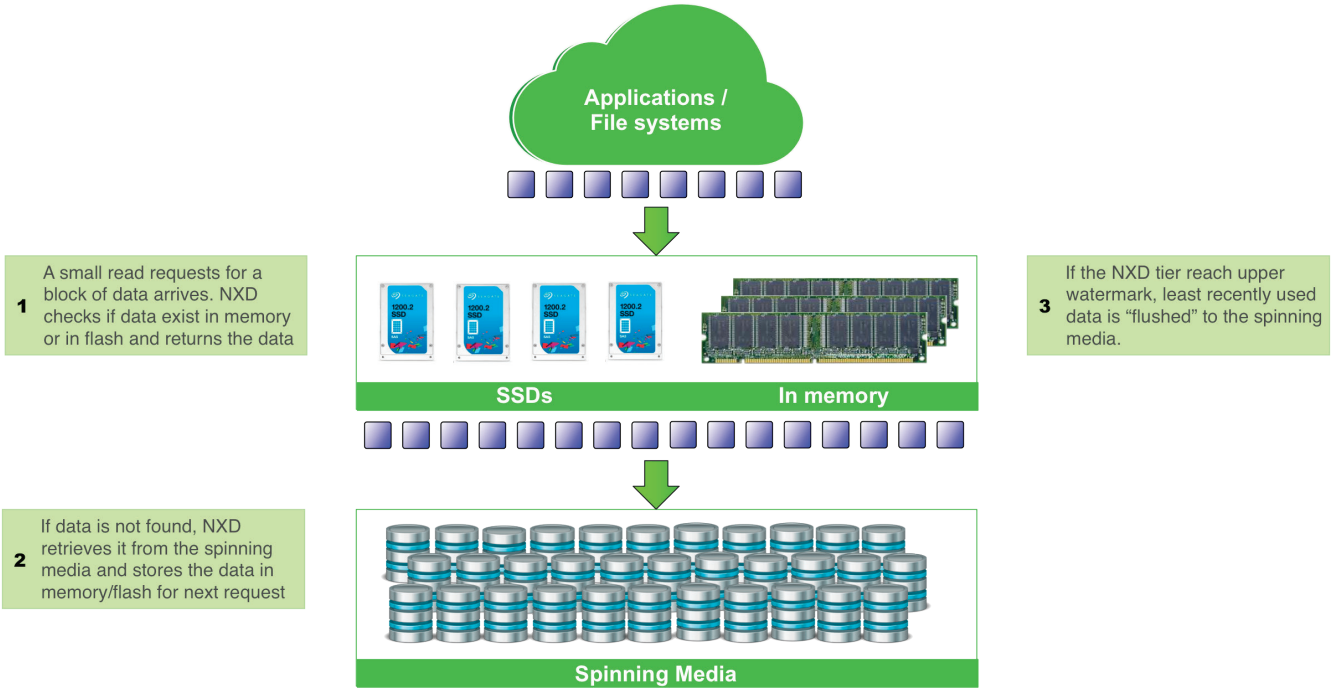
As a leader in storage, Seagate designed the Nytro Intelligent I/O Manager (NytroXD, NXD) to handle small file I/O and large sequential I/O for parallel file systems in a seamless manner. The NXD portfolio of flash acceleration includes a number of unique modes that help accelerate I/O patterns. These modes work in conjunction with Seagate's NytroXD dynamic I/O analyzer, which analyzes workloads in real time to accelerate applications that require storage performance.

NXD is a new ClusterStor feature that combines new hardware configuration to provide an automatic capability to selectively accelerate the performance of specific I/Os usually characterized as "small." While the definition of "small" in this context is user-configurable, the current default sets this limit as equal to or less than 64KB. The NXD software stack contains many functions that addresses different needs of modern HPC applications:

- The Read Persistence feature uses advanced caching technologies to enhance small block random I/O performance by identifying frequently accessed data. This data is copied into low-latency flash storage located in the ClusterStor Storage appliance (specifically the CS L- and G-300N systems).
- Write Back – This feature helps in write-intensive workloads. It allows writes at full speed to the NXD storage tier and then gathers the data sets, then writes it down to disk over period of time.
- I/O Histogram – Profiling specific application workloads from a storage point of view is critical. Understanding where data ultimately lands on the storage is key to set the correct tunable parameters for the NXD.
- Performance Statistics – Real-time update of the cache device performance, including cache hits, cache usage level, cache space consumption, etc., is essential to keep the storage solution working at optimal performance.
- Dynamic Flush – Flushing of dirty data is performed by the flush manager of NXD under different scenarios, such as when the amount of dirty data reaches a certain threshold or when I/O activity drop. The flushing "speed" can be controlled by user-controlled parameter.

The different functionalities are explained in more detail below.

NXD Read Persistencer



NXD Read Persistence is designed to accelerate read operations by responding to applications from the in-memory copy. When a request for a block of data comes in, that block is read from the in-memory copy or flash layer and, if not available, it is read from the spinning media, from where it is served up to the host. This in-memory cache is relatively small and can saturate quickly from read and write operations. As cache becomes full and space for new data becomes necessary, a cache flush operation is performed to make room for new data.

Metadata persistence implemented as part of write-back metadata support ensures that metadata of the dirty region will always be updated on the cache device and kept in-sync with in-memory copy. This is achieved by sending metadata update requests to the cache device for the regions that are modified by the I/O (the same function will be performed for the dirty regions that are flushed by the flush manager). During read-fill operation, only the in-memory metadata will be updated in the context of I/O operation.

NXD will copy the read cache metadata onto the cache device during graceful shutdown of the system. The read cache metadata is not persistent in the case of ungraceful shutdown of the system like system crash, power-recycle etc. After successfully saving the metadata, a bit in the metadata header, called trust bit, will be updated and written to the cache device to indicate that read cache metadata on the cache device is now valid. On subsequent system start, this bit is checked to decide whether the read cache is valid or not

NXD I/O Histogram

NytroXD is capable of analyzing and displaying the I/O of characteristics to provide the user with details of actual block sizes being written and read from the Nytro caching layer. This information can be used to fine tune the parameters of the NytroXD code to improve performance for certain applications or typical workloads.

```
[root@sabre02 nytrocli]# ./nytrocli64 /xd show histogram
**************************************************************************
     Seagate NXD Management Utility
     Version 3.0.10.1 (2017.02.01)
     Copyright (c) 2017 Seagate Technology LLC. All Rights Reserved.
**************************************************************************
Histogram :
==================
Num Reeds < 4K                  = 0
Num Reads 4K                    = 56796
Num Reads 4K+1 – 8K             = 456385
Num Reads 8K+1 – 16K            = 55
Num Reads 16K+1 – 32K           = 31
Num Reads 32K+1 – 64K           = 4
Num Reads 64K+1 – 128K          = 0
Num Reads 128K+1 – 256K         = 0
Num Reads 256K+1 – 512K         = 0
Num Reads 512K+1 – 1M           = 0
Num Reads 1M+1 – 2M             = 0
Num Reads 2M+1 – 4M             = 0
Num Reads 4M+1 – 8M             = 0
Num Reads 8M+1 – 16M            = 0
Num Reads 16M+1 – 32M           = 0
Num Writes < 4K                 = 0
Num Writes 4K                   = 23978
Num Writes 4K+1 – 8K            = 499318
Num Writes 8K+1 – 16K           = 19954
Num Writes 16K+1 – 32K          = 1962
Num Writes 32K+1 – 64K          = 127
Num Writes 64K+1 – 128K         = 39
Num Writes 128K+1 – 256K        = 6
Num Writes 256K+1 – 512K        = 2
Num Writes 512K+1 – 1M          = 0
Num Writes 1M+1 – 2M            = 0
Num Writes 2M+1 – 4M            = 0
Num Writes 4M+1 – 8M            = 0
Num Writes 8M+1 – 16M           = 0
Num Writes 16M+1 – 32M          = 0

[root@sabre02 nytrocli]#
```

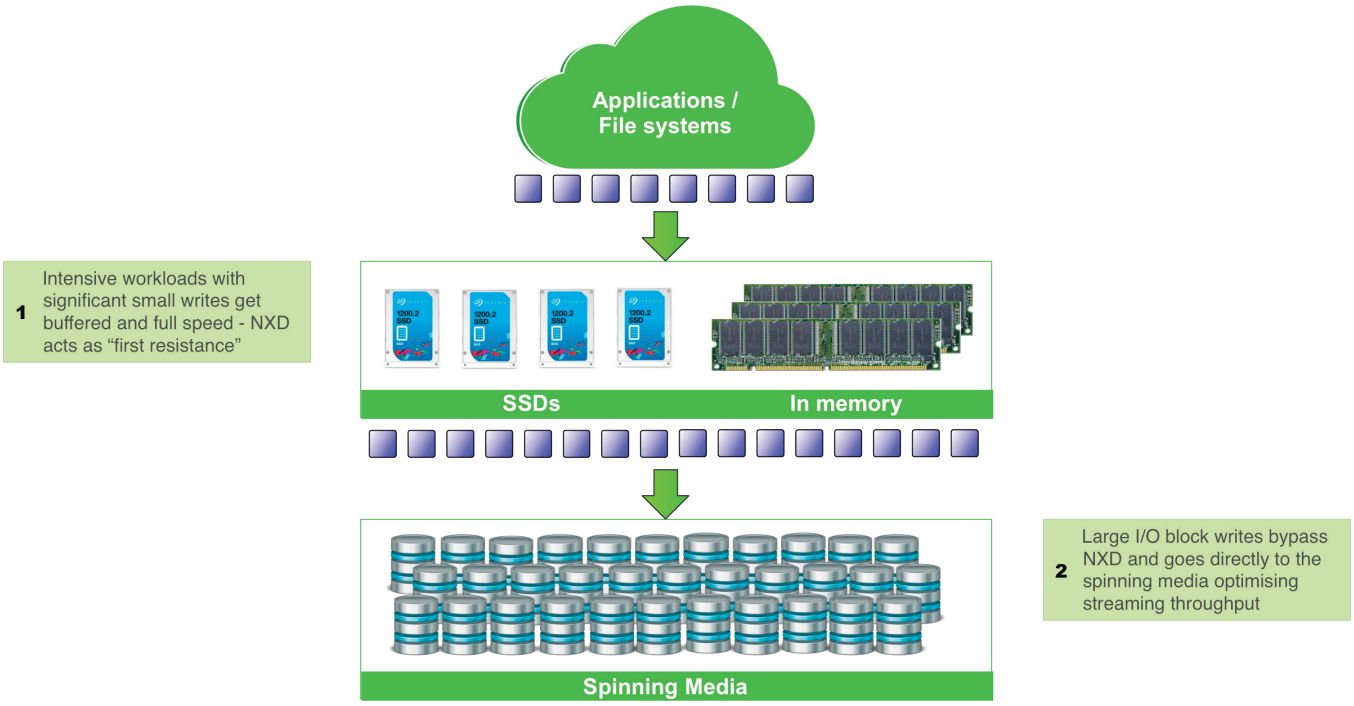
By comparing the Nytro I/O analysis with client side I/O profiling, HPC storage and application administrators can get a clear picture of their HPC application throughput, and potential performance issues can be fixed by tuning the Storage Cluster and/or Clients.

The tools provided also allows users to monitor the actual performance metrics of the NytroXD caching layer in real time.

```
[root@cstor01n04 ~]#/opt/seagate/nytrocli/nytrocli64 /xd show perfmon
*******************************************************************************
  Seagate NXD Management Utility
  Version 3.0.10.1 (2017.01.24)
  Copyright (c) 2017 Seagate Technology LLC. All Rights Reserved.
*******************************************************************************
nxd cache 0 :
============
Name of the Cache Group             = nxd cache 0
Number of VDs                       = 1
Number of Cache Devices             = 1
Queue Depth                         = 4096
Total Cache Size                    = 1.453 TiB
Cache Size in use                   = 14.476 GiB
Cache Block Size                    = 4 KiB
Cache Window Size                   = 64 KiB
Bypass IO Size                      = 32 KiB
Total number of IOs                 = 81507920
Number of reads                     = 38271442
Number of writes                    = 43236478
Total number of bypass IOs          = 7178337
Number of bypass reads              = 4759009
Number of bypass writes             = 2419328
Number of Cache Hits                = 53587873
Number of Cache Misses              = 20741710
Number of dirty CWs                 = 0
Total Cache Blocks flushed          = 35122373
```

Performance details include the total cache size allocated, cache in use and the tunables configured. The performance data also provides the cache statistics, e.g., the reads, the writes and cache hits, the misses as well as cache flushing data; thereby one can see the ingestion ratio vs. the flushing ratio. This ingest / flush ratio helps in tuning flush speed rate, and helps in bursty HPC application workloads. Currently, the only administrator-configurable tunable is the Bypass IO size, but other parameters can be tuned with the help of Seagate Professional Services

NXD Write Back



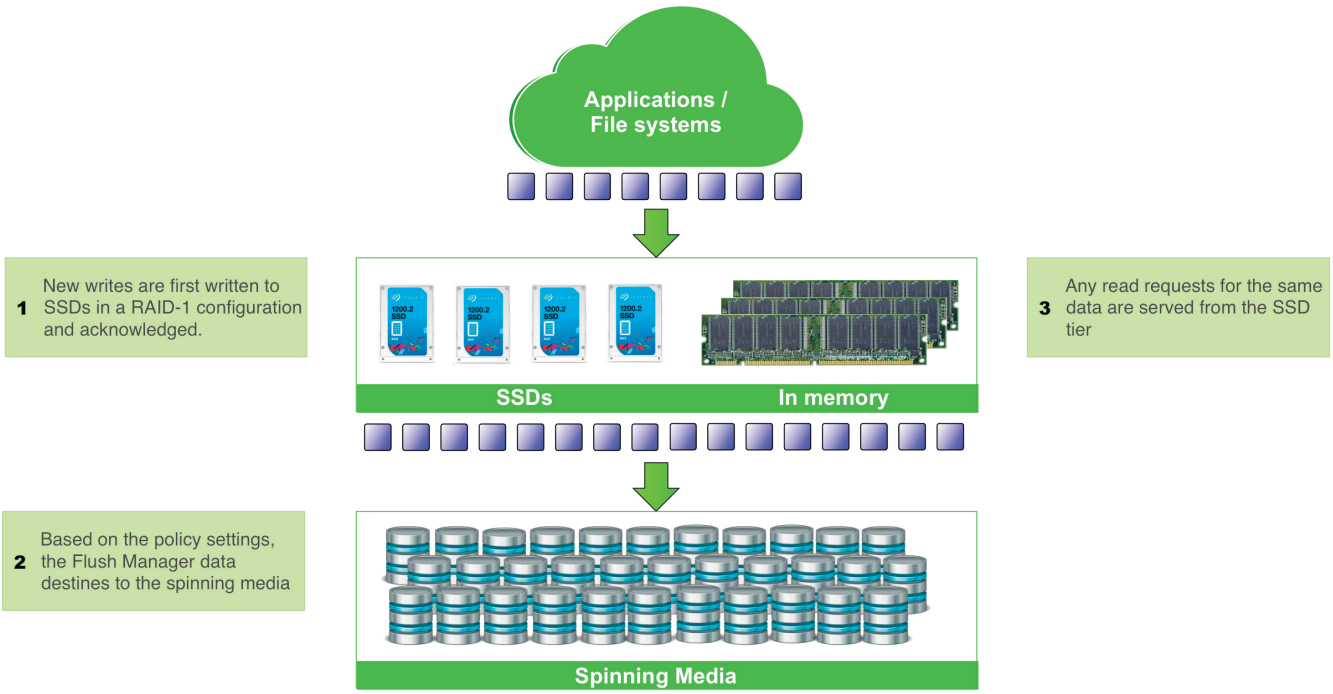
NXD WriteBack is designed for write-intensive workloads and acts as a damper for incoming writes. Unlike other solutions in which the writes land in the flash layer, NXD has “tunable parameters” that detect small writes vs. large writes, and redirects block I/O accordingly. When a small block write request arrives to the storage node, it is streamed onto the NXD SSD-based write pool at a maximum rate while a large block write is “bypassed” and sent to the spinning media in the traditional manner. The analysis of this arbitration is done within the Linux I/O stack and it allows the system to coalesce a number of small I/O blocks before committing the data to the spinning media, thereby reducing the seek operations and achieving the absolute maximum media transfer rate from the rotating media. This caching mode increases the overall effective write bandwidth of the rotating media.

NXD Dynamic Flush

NXD supports write-back caching policy where both read and write requests on hot regions of drive will be cached. With write-back, write requests to the hot regions are acknowledged immediately after

it is written to the cache device and this (dirty) data will be flushed to back-end spinning media in the background. Flushing of dirty data will be performed by the flush manager of NXD software under different scenarios, such as when amount dirty data reaches a threshold, I/O activity during a time interval is low, etc.

Different parameters that are required for the flush logic can be set at the initialization timer. Flush manager will create background threads and gets notified to start the flush of dirty data. Flushing dirty data under different scenario is done by setting the appropriate parameters and initializing the flushing thread.



Scenarios that trigger a flush:

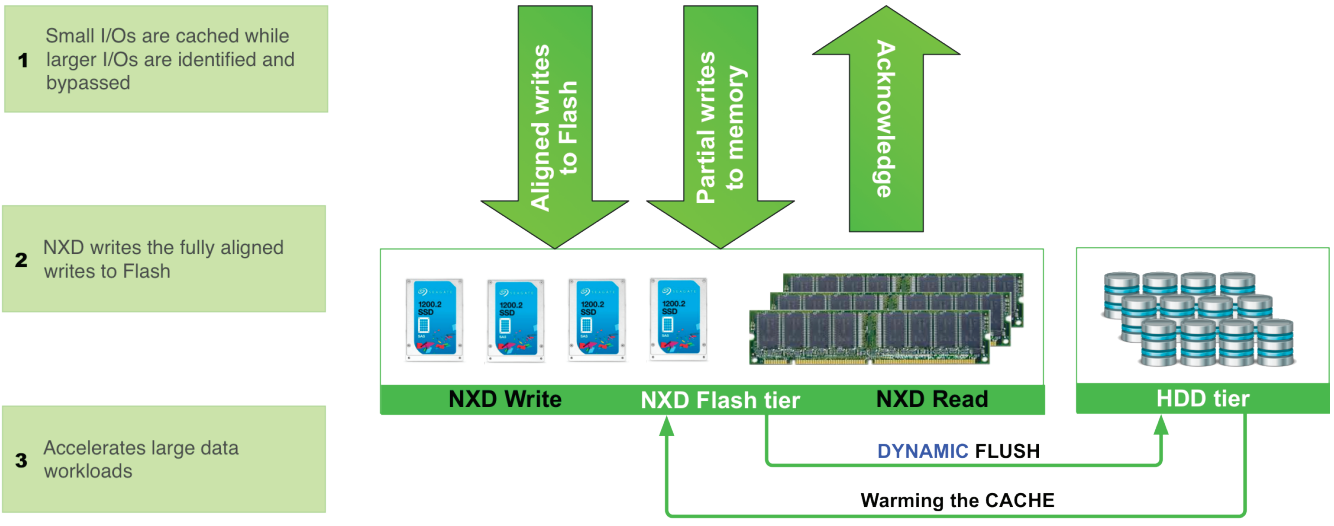
1. Amount of dirty data hits the threshold. This is currently set to 2/3 of total cache capacity.
2. No or very low I/O activity for a time duration.
3. A Virtual Drive with dirty data is being taken out of the cache group.

NXD Flash Caching for Big Data

Big Data workloads are extremely varied and unpredictable, requiring storage systems to deliver extreme levels of performance (in both streaming and random workloads). Most storage systems built with caching technologies that are designed for small block random workloads get congested by mixed workload I/Os competing for cache. This can create application bottlenecks, when exposed to real-world, mixed-application workloads.

NXD is designed to accelerate applications by delivering the highest levels of performance for high-throughput streaming, as well as for random workloads. It does this by analyzing the workload data block size in real time to:

- **Write Back** – where write I/O is directed to cache and completion is immediately confirmed to the host. This results in low latency and high throughput for write-intensive applications, but there is data availability exposure risk because the only copy of the written data is in cache. Write-back cache is the best performing solution for mixed workloads as both read and write I/O have similar response time levels.
- **Resiliency and Mirror** – NXD cache devices use RAID 1 / 1+0 to add resiliency and performance and to preserve precision cache during random, unaligned IOs to accelerate small block, random I/O operations.
- **High Availability** – The NXD cache-RAID uses ClusterStor HA to provide continuous high availability.



These features are all enabled by default in ClusterStor 300N Lustre/GPFS product lines. With ClusterStor-embedded application controllers, all data received in a storage controller and cached by NXD is mirrored on the Seagate flash devices before an acknowledgement is returned to the storage host. NXD with its cache RAID mirroring ensures that data is persisted in the event of a controller failure where the data has not yet been transferred to redundant media. This mechanism is complemented by ClusterStor HA, which enables systems to gracefully persist their cache in the event of failure. When enough data is collected in the cache, based on the flushing policies, it gets staged to be written to redundant media.

Other caching technologies on the market fill the cache with large block bandwidth operations thereby slowing down I/O operations. NXD manages this issue when high-bandwidth operations (full stripes) are written to the storage system, by redirecting the full stripes to redundant media before an acknowledgement is returned. NXD has the intelligence of allowing large block bandwidth data to stream directly to rotating media and bypass the cache – all while caching the small random block I/O to improve performance. Overall, this increases the bandwidth of the controller and reduces traditional I/O bottlenecks.

NXD works in conjunction with its dynamic flush to optimize streaming performance without congesting cache. When a Cache RAID is front-ended with NXD Write Back cache, NXD recognizes this and write operations go directly to the NXD Flash Write Back cache, rather than the rotating media. The write operations are completed, then an acknowledgement is returned to the host system. In the background, NXD Dynamic Flush monitors the policies of cache usage vs. flushing priority and then intelligently groups the small blocks in the NXD Cache RAID tier into a large block and begins highly optimized flushing operations to the redundant rotating media, maximizing the rotating media rate. This has the effect of taking any bursty, high-bandwidth write operations and allowing the storage system to ingest them at full speed, allowing the HDDs to operate at maximum efficiency.

NXD vs. Alternate Flash Caching Solutions

There are multiple flash vendors that offer all-flash and/or hybrid (flash + disk) solutions to accelerate high-performance applications. However, it's important to recognize that the primary market for these solutions are databases, email and other enterprise applications in which the I/O profile is always small block.

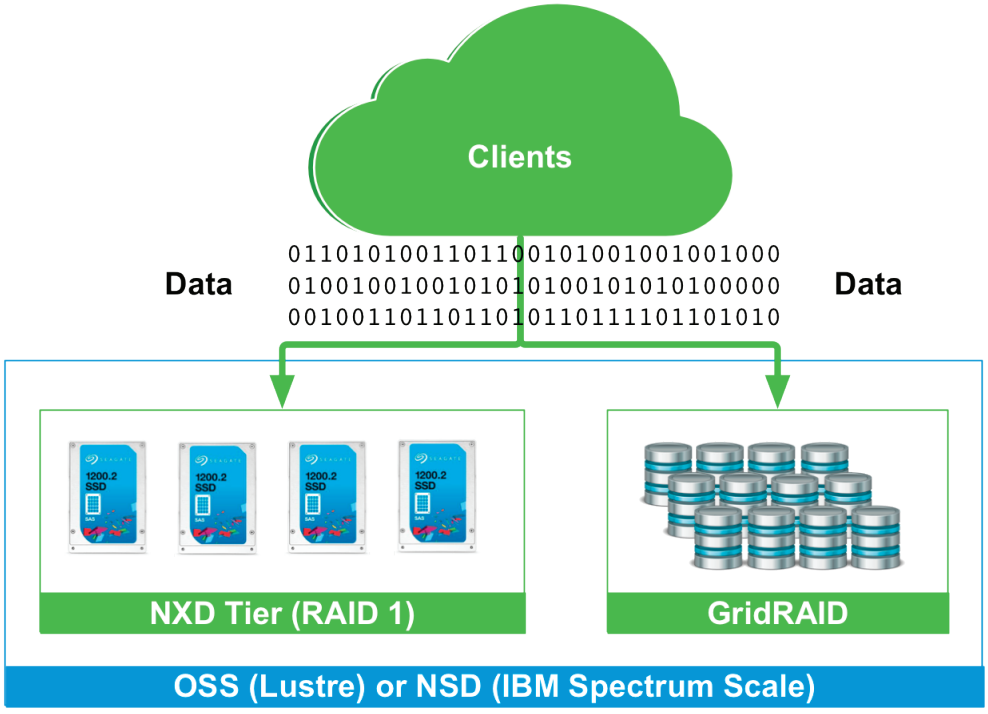
The Seagate ClusterStor L/G-300N HPC storage appliance is specifically designed for different data workloads (big data, random, sequential) and uses a fundamentally different approach to Flash Caching for the parallel workloads. More importantly, the flash caching algorithm is designed to work with many different types of applications from a wide variety of industries, such as Life Sciences, Rich Media, Financial Services, Oil & Gas, Automotive and Aerospace industries.

File System Integration

Seagate ClusterStor NXD is designed to integrate with file systems and applications to truly lower the TCO for customers trying to tackle the challenges of Big Data. The following section describes how NXD integrates with two ClusterStor Solutions (today): L300N and G300N.

CLUSTERSTOR L300N

L300N is a massively scalable, high-performance, open source file storage appliance based on Intel Lustre file system. The L300N system is built by HPC and storage experts, supported by the world's most skilled parallel I/O and storage teams at Seagate, and known worldwide as the standard in HPC storage clustering that powers the largest number of top supercomputing sites worldwide. In a typical L300N implementation, a parity declustered RAID system with 41 HDDs is used for the object storage



target (OST) and each enclosure holds two such volumes as well as two embedded application controllers that access each volume in a fail-over paradigm. The two last slots are occupied by two SSDs (high-performance Seagate SAS SSDs with a capacity of 6.4 TB). These SSDs are partitioned in several RAID 1 partitions used for the NXD cache as well as Linux Journals and Write Intent Bitmaps.

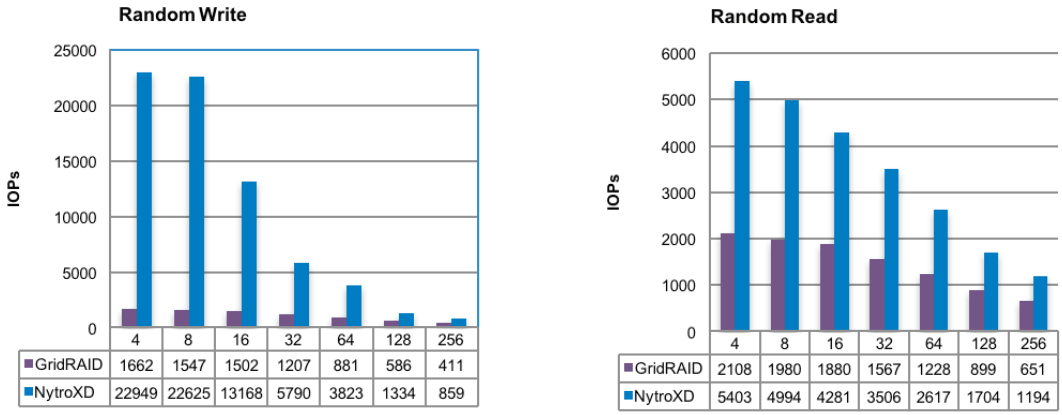
This combination of NXD features (Dynamic Flush, Histogram Analysis, Read Persistence, Scalable Flash Tier) delivers significant performance advantages over traditional pure HDD or SSD implementations. It provides much great performance at a significantly lower TCO. Workloads that get accelerated when NXD is configured with ClusterStor L300N includes small block/file reads/writes/ rewrites operations.

CLUSTERSTOR G300N

A Spectrum Scale-based appliance, G300N is very similar to the L300N Lustre appliance in terms of general design. The main difference is that the SSDs are also used to store distributed metadata in addition to the NXD cache, the WIBs and Linux Journals.

Results

NytroXD can enhance any standard file systems including IBM Spectrum Scale and Lustre. The effectiveness will of course be different due to basic properties of said file system but overall the results are similar. Below is an example of the effect of NytroXD as a function of different block sizes comparing plain GridRAID and NytroXD on a Lustre file system.



Benchmark results comparing the IOPS performance of a NytroXD enabled ClusterStor solution for a number of different block sizes.

For smaller block sizes in the 4-8 KB range, NytroXD is capable of producing almost 15x better performance. As the block size gets closer to 1 MB (optimal block size for disk-based systems) the degree of acceleration diminishes more or less linearly for writes. For read operations, the effect is not as exceptional but still significant. Obviously, benchmarks are at best indicatory and only real-life testing using end-user applications and datasets can produce a clear picture of the general acceleration provided by ClusterStor with Nytro Intelligent I/O Manager.

Summary

Organizations are under more and more pressure to accelerate the performance of their Big Data and HPC workflows, while cutting costs. Solid state devices and other flash-based media have the potential to solve this problem. However, replacing the disk drive or rotating media with solid state is very expensive. Flash caching is one of the more cost-effective ways to deploy solid state storage. ClusterStor NXD is a unique flash caching technology designed for Big Data/HPC workloads that extends the functionality of the ClusterStor storage appliance. This is accomplished by selectively adding some amount of flash storage to the traditional rotating media, thereby accelerating application performance at a much lower acquisition cost and TCO. The NXD portfolio of flash acceleration tools include Read Persistence, Small Block Acceleration, Dynamic Flush with IO Histogram analysis. They work in conjunction with the ClusterStor GridRAID to analyze workloads in real-time and accelerate applications and file systems that require multi-dimensional storage performance.

But while the NytroXD solution provides a more cost-efficient way of accelerating high-performance I/O using flash-based accelerators, the most significant difference and unique value to users is the transparency with which this happens. There are several “Burst Buffer” technologies on the market, but common for most if not all of them is that they constitute a separate storage tier, isolated from the main scratch storage and requiring data movement to secure cached data on a reliable media. And rather than creating a large pool of storage that spend most of its time either sitting idle or migrating data to a secondary tier, having flash acceleration that scales linearly with the growth of the scratch file system seems like a much better idea.

The market for HPC/Big Data storage solutions today is growing exponentially and the need for performance and capacity is growing even faster. And while flash-based storage technology has the potential to deliver the insatiable need for performance it must be applied intelligently. Solutions that require advanced data migration, custom clients, re-compilation of codes or other changes to an established workflow do not scale and will eventually become a bottleneck. Only a truly transparent solution can be expected to deliver the seamless accelerations current workloads require.

CONTACT

As you explore your options, **contact us** to learn more about all the benefits of working with us.