Achieving NVMe Shared Storage with E8’s HA NVMe Enclosure and Intel Datacenter SSD

High Availability, Low Latency Storage at the Line Rate of 24 NVMe Drives

Abstract

NVMe drives have been introduced to the world by Intel in 2014, heralding a new era in PCIe-attached flash. But until now, NVMe drives could only be used as local drives. To be used as shared storage, a new and critical requirement for NVMe connected drives would be high availability (HA). E8 Storage has produced the world’s first centralized NVMe storage solution, which can unlock the economics and architectural advantages of centralized storage with the high I/O and low latency advantages of PCIe NVMe SSDs. E8’s unique and patented distributed software architecture is able to extract the full performance of remote NVMe drives, specifically using the new dual-ported Intel® DC D3700 and D3600 NVMe SSDs. These are the first such dual-ported NVMe SSDs in the world – critical to enabling NVMe to be used as centralized storage.

Why is NVMe such a big deal? While most technology users acknowledge that SSDs have performance and I/O advantages over spinning magnetic storage media, they are quite likely unaware that the benefit of SSD capabilities have been restricted by legacy (spinning) disk protocols – ones that place arbitrary limits on system capabilities built around the physical limitations to spin a metal platter at thousands of RPMs. Instead, NVMe is built from the ground up for SSDs. Using SSDs without the PCIe NVMe protocols is like driving a car down a wagon trail instead of using a highway.

E8 Storage’s NVMe HA enclosure with its distributed software stack, combined with Intel’s dual-ported NVMe SSDs, allows storage professionals to choose a new breed of storage: one combining the benefits of local flash and NVMe, with the benefits of centralized storage and high availability. This is especially suitable for hyperscale data centers and large data centers that are deploying NVMe today, as well as enterprise and private cloud efforts seeking next-generation all-flash-arrays.

"From server-scale to rack-scale: Share the economics and performance benefits across your data center with the HA NVMe Enclosure from E8 Storage"

Zivan Ori, CEO & Co-Founder E8 Storage
Introduction

NVMe SSDs have been revolutionizing the data center, bringing x10 performance and x10 density compared to SATA and SAS devices. This comes from advancing new standards for SSDs through nvmexpress.org, with many key contributions from Intel. This performance and density explosion has pushed data centers and enterprises to use NVMe only as local drives within the server. Connecting them remotely from the server as JBOF (Just-a-Bunch-Of-Flash) will impact their latency, and there are no products or solutions before today that can extract the full bandwidth and throughput of remote NVMe SSDs.

High Availability

To be used as centralized storage, the requirement for high availability becomes critical. Without it, a single failure in any of the storage arrays components, e.g. front-end network ports, internal CPU/RAM, etc., will result in the loss of a large and expensive group of SSDs. For most customers this is unacceptable. A solution of maintaining 2 or 3 replicas instead of relying on centralized storage high availability, while common in mechanical hard drives, becomes wasteful and unaffordable for NVMe SSDs.

Therefore to allow for centralized NVMe, solutions for high availability are required.

HA Enclosures

In general, HA enclosures are common for SAS and SATA drives - for SSDs as well as HDDs. An HA enclosure contains 2 front-end canisters that are redundant and hot swappable for repair. It contains a passive mid-plane that wires the 2 front-end canisters to all the drives. Furthermore, the drives in this configuration need to be dual-ported (a common SAS drive feature) or contain a SATA dual-port interposer to allow usage of SATA drives in an HA enclosure.

The Challenge

Using local NVMe drives introduces a problem that shared storage has solved long ago: local SSDs are islands of storage and their capacity must be determined up front when buying a SSD for the lifetime duration of the server. Customers normally purchase large amounts of SSDs to prevent locking out the server as they get filled up. This entails a lot of stranded capacity; on average up to 70% of the SSDs capacity is unallocated.

Moving to shared NVMe has several advantages. First of all, it eliminates the need to determine up front how much SSD capacity is required when building a new server farm. SSDs can be procured when the capacity of the existing SSDs runs out. This means that customers can benefit from the declining price of flash at the time of purchase rather than paying up front. Secondly, servers are maintained separately from storage, as was customary until now. This means that server downtime does not imply storage downtime and vice versa. Thirdly, the lifecycle of servers and storage in the data center is kept separated, e.g., servers need to be replaced every 2 years due to TCO demands. However, since the NVMe SSDs reside inside them, either they are not replaced at all (leading to inefficient data center), or the SSDs need to be migrated from server to server (which is expensive).

In order for NVMe drives to be used as shared storage, an NVMe enclosure and a software stack are required. The storage technology used must enable full performance (throughput and bandwidth) without impacting latency. Shared storage also requires redundancy - not only in connectivity, e.g., via dual ports, but also redundancy around SSD failures, e.g., via RAID.
The Solution

Intel® Datacenter D3700/D3600 dual-port NVMe SSD with E8 Storage HA NVMe enclosure achieves NVMe shared storage. The distributed architecture and software stack allow customers to achieve full NVMe performance of their remote drives (both throughput and bandwidth) and with minimal impact on NVMe latency (as compared to local NVMe drive usage). The high performance storage solution scales with capacity while maintaining the same level of performance expected from local storage inside the servers. The disaggregated architecture allows storage capacity to be dynamically allocated, augmented or replaced, without impacting the rack’s performance or requiring maintenance downtime.

Benefits include:

• TCO is lowered by deploying NVMe storage when needed, without a need to predict storage consumption and pay for its maintenance before it is used.
• In hyperscale deployment, data is persistent (even in the event of compute node failure) achieving 100% uptime without performance degradation due to service rebalance in the cluster.
• Storage investment is gradual instead of upfront. Customers can now buy NVMe storage when they need it, instead of when they start building the data center.
• Supply chain simplification – all servers are the same: any server in the rack can be provisioned with as much storage capacity and performance as needed, removing the need of multiple configurations and inefficiency in the supply chain.

Features

• Hardware design without a single point of failure:
  » 2 redundant front-end canisters, that are hot swappable
  » Passive mid-plane design
  » 2 PSUs
  » RAID on NVMe SSDs, dual-parity support
  » Dual-port SSDs support
  » Power failure protection

Benefits

• Intel D3700/D3600 dual-port NVMe SSD with E8 Storage HA NVMe enclosure addresses the storage needs in enterprise and hyperscale data centers by providing NVMe shared storage that has high performance, availability and reliability.

Specifications*

<table>
<thead>
<tr>
<th>Specification</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Density</td>
<td>2U24</td>
</tr>
<tr>
<td>Capacity (raw)</td>
<td>24*2TB = 48TB raw</td>
</tr>
<tr>
<td>Disk redundancy</td>
<td>Double parity</td>
</tr>
<tr>
<td>Connectivity</td>
<td>4-8 ports of 40GE/50GE or 4 ports of 100GE</td>
</tr>
<tr>
<td>Performance</td>
<td>R W</td>
</tr>
<tr>
<td>Latency (4KB, QD=1)</td>
<td>100us 40us</td>
</tr>
<tr>
<td>IOPS (4KB)</td>
<td>10M 2M</td>
</tr>
<tr>
<td>Throughput (GB/s)</td>
<td>40 20</td>
</tr>
<tr>
<td>Power consumption</td>
<td>1200W max; 800W typical</td>
</tr>
</tbody>
</table>

* Please note that the performance numbers are not verified yet and might change later. The current data show high performance / low latency on par with local NVMe.

Conclusion

Advances in computing and storage technologies have been accelerating – and they also accelerate the expectations of what enterprises expect from IT regarding computing, networking, and storing of data. The NVM express protocol has engineered SSDs from the ground up to unlock the technological improvements of SSDs – and now the richness of E8 Storage and Intel® Datacenter D3700/D3600 SSDs have been combined to share the NVMe advances across the data center.

To learn more about the HA NVM Enclosure, visit www.e8storage.com
To learn more about Intel’s contributions to storage technologies, visit www.intel.com/storage
Test Environment
The test environment consisted of deploying 2 supermicro* 4-node servers connected through a top of the rack 50GE/100GE Mellanox* switch to an E8-D24 controller populated with 24 Intel D3700 800GB SSDs. Various FIO workloads were used to measure performance.

Hardware Setup
Host (Total 8 Nodes): 2 x supermicro* SuperServer* 2028TR-HTR - 2U 4 nodes (Specification sheet)
- Per node:
  » 2 Intel Xeon E5-2630v3 processors per node
  » 64G DDR4 RAM
  » SATA boot device
  » 1 Mellanox ConnectX®-3 dual-port 40GE NIC
  » CentOS* 7.1 Linux operating system

Network Switch
Mellanox SN2700 - 32-port Non-blocking 100GbE Open Ethernet Switch System

Controller
E8-D24 Controller: populated with 24 Intel D3700 800GB dual-port NVMe SSDs, 4x Mellanox CX4 50GE dual-port NICs

Tests
FIO tests (Based on FIO Ver2.2.11):
- 8 Threads per host running 4KB, 8KB, 32KB 100% RR, QD 64
- 8 Threads per host running 4KB, 8KB, 32KB 70%/30% RW, QD 64
- 1 Threads from a single host running 4KB 100% RR, QD 1
- 1 Threads from a single host running 4KB 100% RW, QD 1
- 8 Threads per host running 128KB Sequential Read, QD 64
- 8 Threads per host running 128KB Sequential Write, QD 6

“NVMe - the next frontier - Expand your SSD capacity as you need with the HA NVMe Enclosure from E8 Storage”

Disclaimers
Intel technologies’ features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at intel.com/storage. Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors and products. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more information go to http://www.intel.com/performance. © 2016 Intel Corporation © 2016 E8 Storage All Rights Reserved. *Other names and brands may be claimed as the property of others.