WHITE PAPER

Data Center
Intel® SSD DC P3700 Series

Hyperscale, Dependable Data Transfer with the Intel® SSD DC P3700 Series

Testing proves high-speed performance under strenuous use

Table of Contents

Executive Summary 1
Requirements for Point-to-Point, Shared, Production Network 2
The Importance of High-Speed Storage to High Data Transfer Rates 3
Long Term Results 4
Monitoring 6
Conclusion 7
Contributors 7

Executive Summary

High-speed data transfer, in the range of 100Gbps and faster, is being explored by the US Department of Energy[1] (DOE) Office of Science[2] (SC)'s SLAC National Accelerator Laboratory[3] (SLAC), in partnership with Zettar Inc.[4] using the Intel® SSD Data Center P3700 Series. The ultimate goal is to ensure that the US maintains a world-leading capability for advanced research in chemistry, materials, biology, and energy.

Test results obtained by SLAC and Zettar have shown that these NVMe-based Intel® SSDs have been integral to helping meet the mission-critical requirements of a top-tier US Exascale Computing[5] preparation project.

Long-term stress testing began in early 2015 when Zettar installed 16 2.5-inch, 1.6TB capacity, NVMe-based Intel® SSD DC P3700 Series drives. Since then, they have been consistently performing under strenuous use.

Beginning in 2014, the US DOE's SC SLAC and Zettar have explored the various aspects of achieving high-speed data transfer rates of 100 Gbps and faster. Their main goal is to support the data transfer requirements of a few major projects hosted at SLAC. The foremost among them is the Linac Coherent Light Source[6] (LCLS) project – the world's first "hard" X-ray, free-electron laser,[7] and its follow-on project, LCLS-II.[8]

Towards the end of 2020, even after significant data reduction, the target data transfer rate still needs to be around 200Gbps.

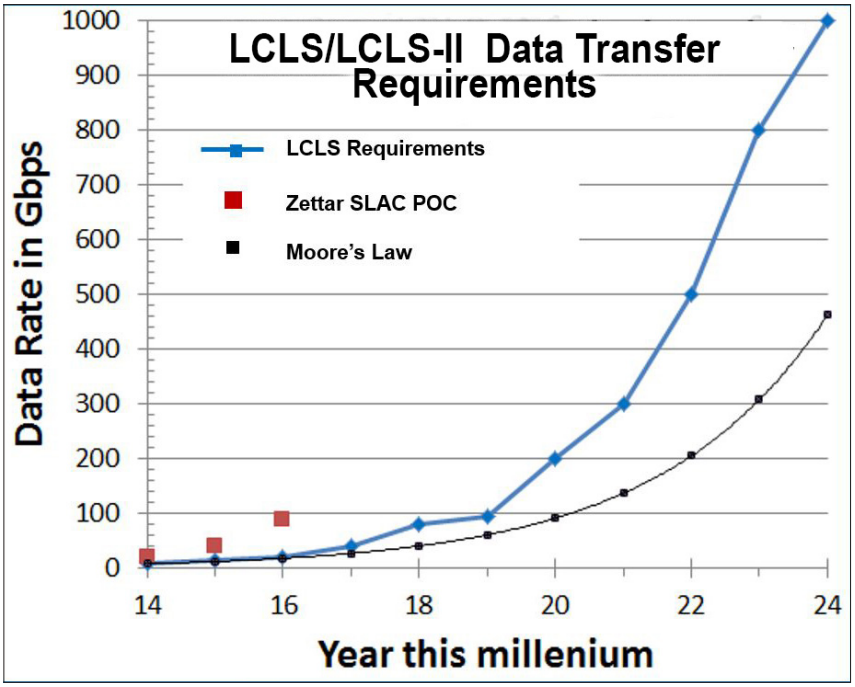
Requirements for Point-to-Point, Shared, Production Network

It is important to note that high data-transfer rates of Tbps, for multiple sources, multiple destinations, and aggregated traffic volume, have already been realized. For example, in 2013 Comcast's* regional backbone was carrying about 1.4Tbps over 15 x 100Gbps connections[9] between New York and Chicago. Also, from 2011-2013, Google carried out its B4 private WAN traffic management project to ensure high bandwidth utilization for its private WAN connecting its data centers.[10] The Google traffic rates were in the Tbps range, as well.

In the LCLS-LCLS-II's cases, the traffic is over a point-to-point, shared, production network connection from a single data source (LCLS I/II at SLAC, Menlo Park, California) to a single data destination (National Energy Research Scientific Computing Center[11] (NERSC), Berkeley, California). Thus, the processing intensity at each infrastructure level is much higher.

SLAC and Zettar, motivated to create a solution capable of addressing these demands, introduced high-performing NVMe-based data center SSDs from Intel's Non-Volatile Memory Solutions Group (NSG), specifically the Intel SSD DC P3700 Series (DC P3700).

In April 2015, Intel NSG provided the SLAC/Zettar effort with 16 1.6TB DC P3700 drives, in the 2.5-inch form factor. These high-performance NVMe SSDs were used in building Zettar's high-speed data transfer test bed.[12] As Zettar pushes the performance envelope of its hyperscale (PB and larger) data distribution platform[13] ever higher, it continues to accumulate a body of stress-testing observations on the performance consistency of these 16 DC P3700 drives.



Graph Credit: Dr. Les Cottrell, US Dept. of Energy, SLAC National Accelerator Laboratory, 2016

Figure 1: Data Transfer Rate Requirements of the LCLS-II Project

The Importance of High-Speed Storage to High Data Transfer Rates

Transferring hyperscale level data at speeds of 100+ Gbps is not dependent on the network alone; adding network bandwidth does not automatically result in higher transfer speed. In fact, as noted in the previously mentioned Google project, most enterprises generally utilize only 30-40% of their private connections. Once the transfer process is established, storage IOPS determines the attainable data transfer rates, as Figure 2 illustrates.[14] Although the figure shows only the sending side, the same principle applies on the receiving side. In particular, for high-speed data transfers, the write performance (or the lack thereof) is almost always the main bottleneck.[15]

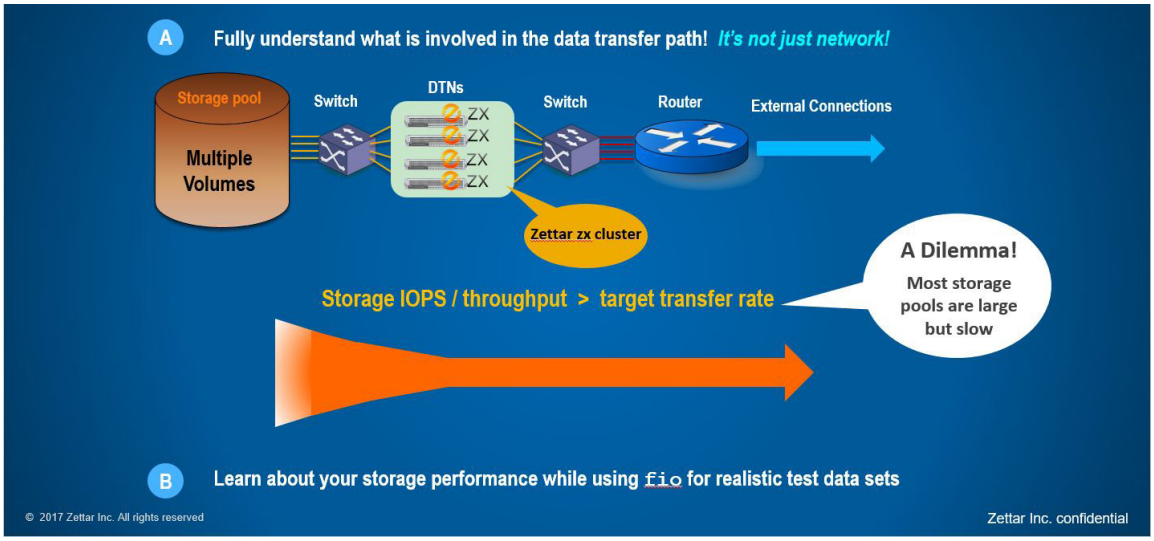


Figure 2: Storage IOPS Determines the Attainable Data Transfer Rates

In the transfer process, data is first extracted from a storage pool on the sending side, and then is written into a storage pool on the destination side. Everything along the data transfer path presents a degree of overhead. Therefore, to meet the desired target data transfer rate, ideally the storage pools on both ends must be able to offer more than enough IOPS to accommodate such overhead. Everything else, e.g. network protocol, is still important, but secondary in nature. Nevertheless, the design of most storage pools focuses on capacity, not high IOPS.

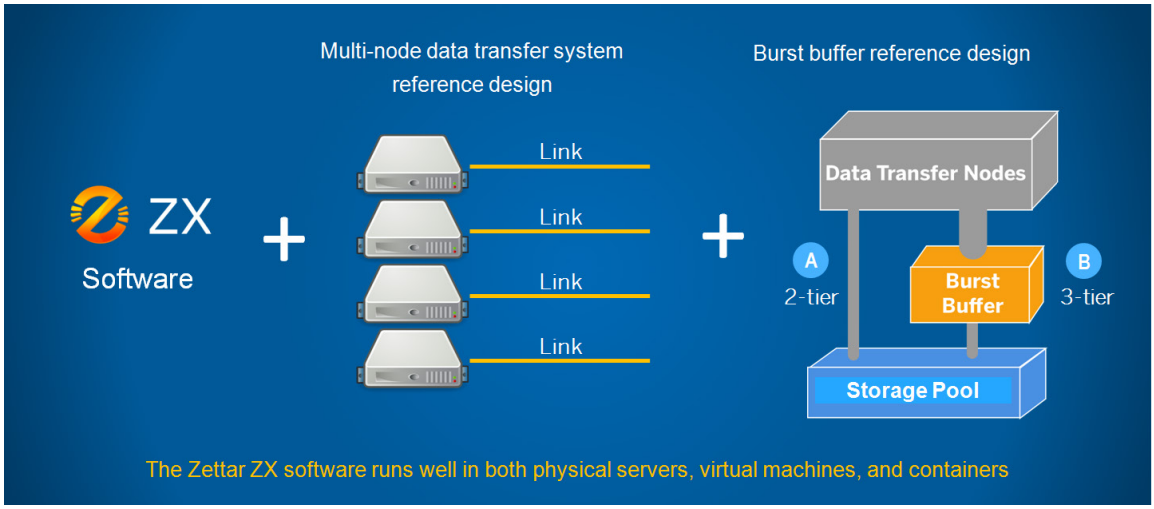


Figure 3: Proposal for the solution using Zettar zx Software

Figure 3 shows how Zettar resolves the bottleneck using one of their two reference designs; in this case the "burst buffer". Both published reference designs can be realized using commercial off-the-shelf (COTS) hardware and depend heavily on the excellent write performance of NVMe-based DC P3700.

The "burst buffer" reference design is based on aggregated NVMe-based DC P3700 drives.[16] This reference design employs the BeeGFS parallel file system[17] from ThinkParQ*[18] as an effective way to aggregate the various distributed NVMe SSDs, although other possibilities exist.

For a description of the multi-node data transfer system reference design, please see SLAC-TN-15-001.[19]

Long Term Results

In 2015, the DOE's Energy Science Network[20] (ESnet) provided a 5000-mile 100Gbps OSCARS[21] loop to the SLAC/Zettar effort. Zettar has done extensive stress testing, using this shared, production 100Gbps network to carry out numerous production preparation runs at very high rates (up to 70+Gbps in 2015 and 100Gbps in 2016), with both regular and TLS encrypted transfers. The OSCARS loop, the major 2015 accomplishments, and the experience with the NVMe-based DC P3700 are described in SLAC-TN-16-001.[22]

The main focus of the SLAC/Zettar effort has always been to demonstrate reliability for deployment in the production environment. Therefore, citing SLAC-TN-16-001, "We want to cover a testing envelope that most users won't exceed for a long time. The Principle of Least Surprise is a critical key in the creation of such software." Since LCLS-II plans to transfer multiple PBs within a few weeks from SLAC to NERSC, the feasibility of fast PB level transfers must be established. Thus, in early May 2017, during the ESnet Site Coordinator Conference (ESCC) Spring 2017,[23] Zettar zx was used to transfer a 1PB dataset in 34 hours over ESnet, and generated more than 1/3 of the live traffic in that timeframe. Figure 4 illustrates the landmark demonstration.

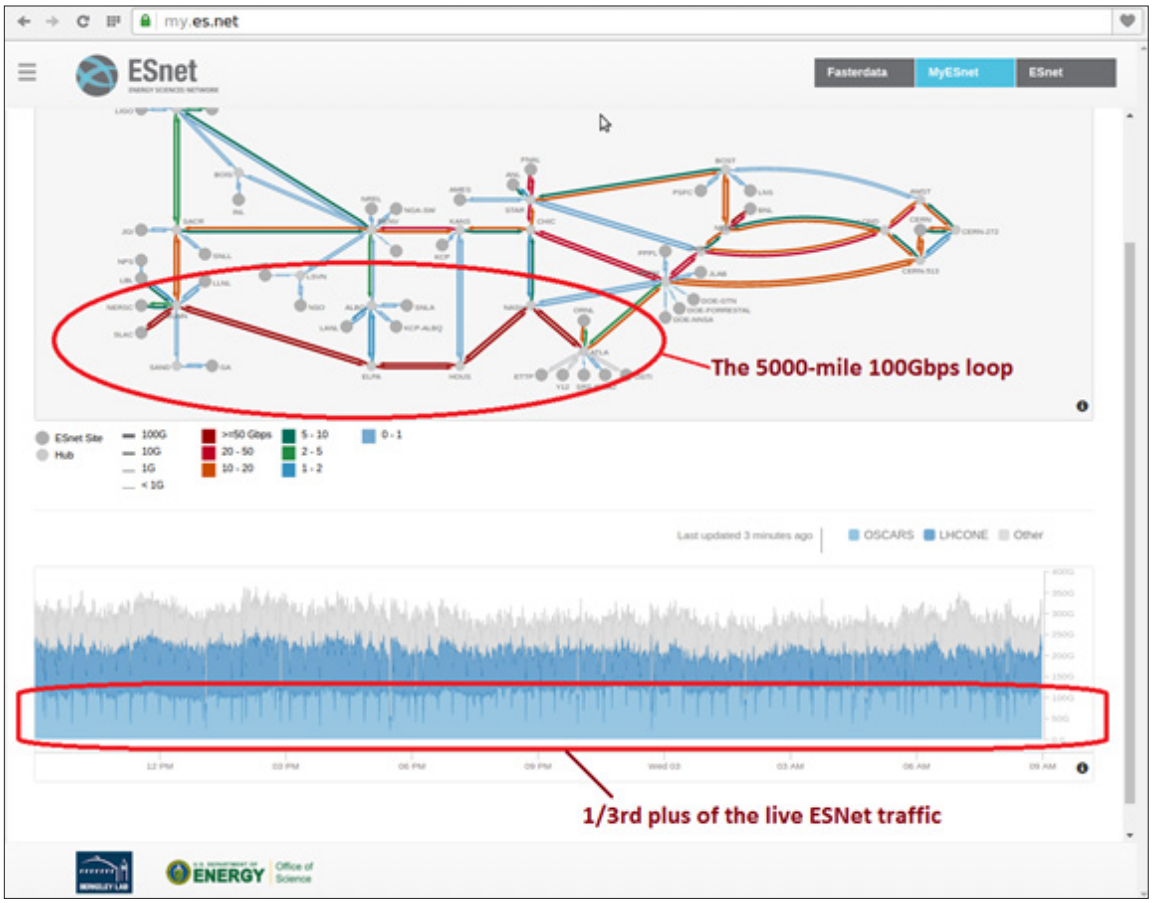


Figure 4: The Landmark 1PB in 34 Hours Capability Demonstration Run (ESCC Spring 2017)

Both ends each employ a modest two-node cluster consisting of two inexpensive 1U commodity servers with 4 x 10Gbps unbonded Ethernet ports (thus 2 x 4 x 10Gbps = 80Gbps - the bandwidth cap). Figure 5 illustrates the overall setup for demonstration.

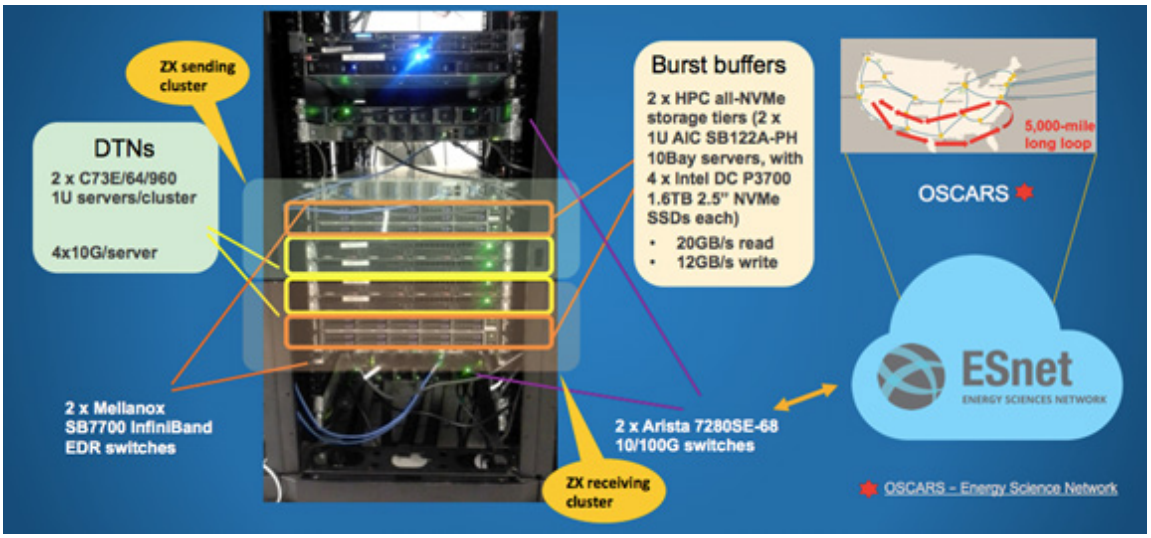


Figure 5: Setup Used for the 1PB in 34 Hours Capability Demonstration Transfer, Including the 5000-Mile ESnet 100Gbps OSCARS Loop

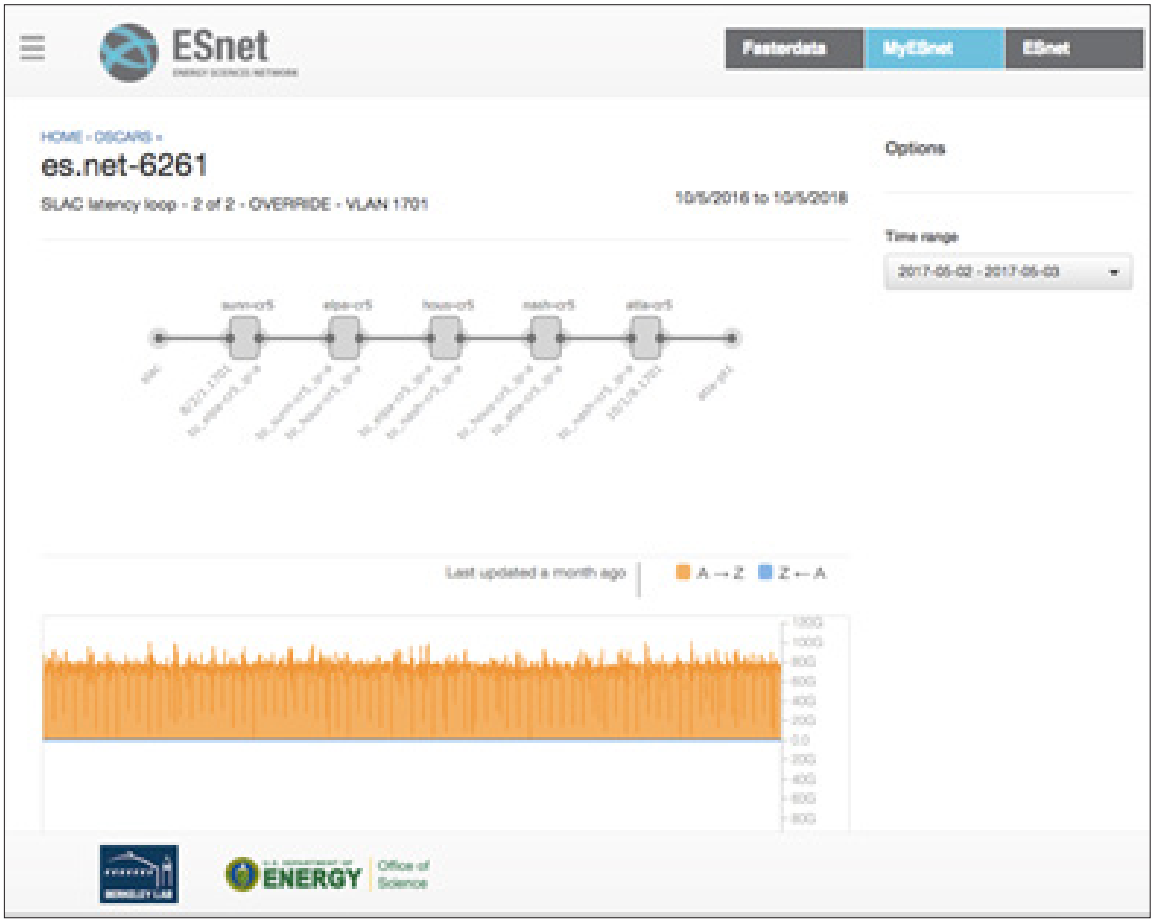


Figure 6: Results of the First Run are Visible Worldwide on the ESnet's Network Portal

The performance consistency of the DC P3700 enables the smoothness of the transfer speed profile. Zettar has been using DC P3700 since early 2015 for building the various test beds for the SLAC/Zettar effort.

Monitoring

All performance critical tasks require a monitoring system in place. During these hyperscale-level transfers, Zettar paid close attention to the various NVMe run-time statistics. Figure 6 shows the results that were captured after the landmark PB transfers.

The burst buffers used in the SLAC/Zettar effort are far more demanding of these NVMe SSDs than most enterprise uses are. The DC P3700 provides the highest write performance among all the volume-produced NVMe SSDs introduced from 2014-2016, and achieves outstanding endurance up to 62.05 PBW (petabytes written). The 34-hour landmark transfer run causes 0.25 PB of "wear" on each NVMe SSD used for the burst buffer on the receiving side, as it has only eight DC P3700 drives, monitoring the remaining endurance of each SSD is very important. Low-endurance NVMe SSDs would not be suitable for such an application. Nevertheless, Zettar has experienced consistent run-time metrics using the DC P3700, even after more than two years of strenuous use.

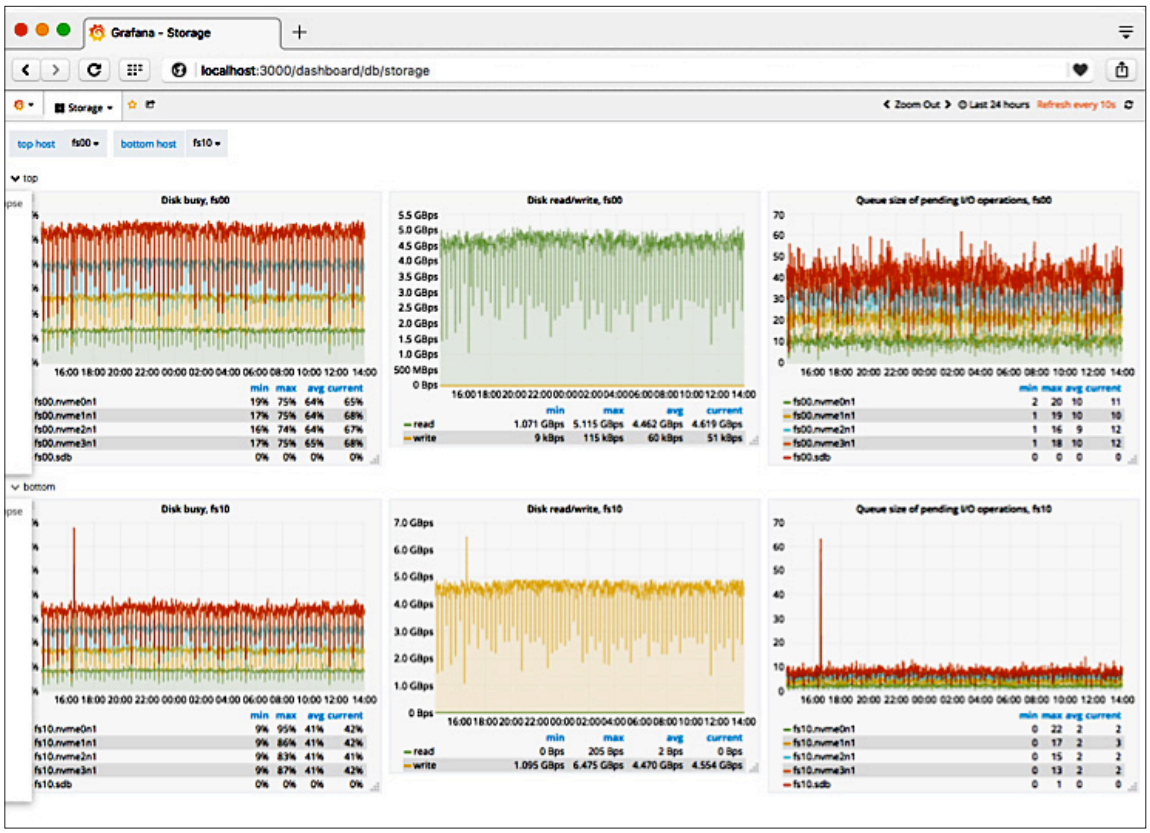


Figure 7: Various Intel® SSD DC P3700 Run-Time Statistics, as Displayed by Zettar's Monitoring System.

Conclusion

Based on more than two years of experience, Zettar deems that these 16 NVMe-based Intel® SSD DC P3700 Series drives deliver exceptional design and performance. Intel NSG also offers a frequently updated Intel® SSD Data Center Tool,[24] with extensive and straightforward documentation. Additionally, Intel is a significant contributor to the default Linux* NVMe driver.[25] These combined benefits make NVMe-based Intel® SSDs a logical and manageable choice for demanding High Performance Computing (HPC) tasks, such as hyperscale data distributions. Having had an excellent experience with the Intel family of DC P3700 drives, Zettar looks forward to working with the new Intel® SSD DC P4500[26]/P4600[27] Series, and expects to be equally delighted by these follow-up products from Intel's Non-Volatile Memory Solutions Group.

Contributors

Chin Fang, Ph.D., Zettar Inc. Founder & CEO

Andrey Kudryavtsev, Intel NSG Solution Architect

Nivedita Singh, Intel NSG Account Manager

Cyndi Peach, Intel NSG Strategic Business Development Manager

Learn more at intel.com/ssd

1. https://energy.gov/
2. https://science.energy.gov/
3. https://www6.slac.stanford.edu/
4. http://www.zettar.com/
5. https://exascaleproject.org/
6. https://portal.slac.stanford.edu/sites/lcls_public/Pages/Default.aspx
7. http://lcls.slac.stanford.edu/Article.aspx?article_id=183
8. https://portal.slac.stanford.edu/sites/lcls_public/lcls_ii/Pages/default.aspx
9. https://www.nanog.org/sites/default/files/mon.general.vik_.comcast.43.pdf
10. https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/41761.pdf
11. http://www.nersc.gov/
12. See FIG. 4 in https://itpeernetwork.intel.com/transferring-data-for-distributed-data-intensive-engineering/ for a functional diagram of the 3rd generation of Zettar test bed (circa 2016)
13. "hyperscale," Zettar zx is designed to handle PB and larger data sizes; "data distribution," zx can handle not only the typical data transfers, but also data replications, migrations, and other more "specialized" data transport tasks. Therefore, the term "data distribution" is more fitting; "platform"; it is feasible to customize zx with Python plugins to extend its capabilities for custom requirements. Thus, zx is not just an application. It is a platform.
14. See also FIG. 1 and 2 in https://itpeernetwork.intel.com/transferring-data-for-distributed-data-intensive-engineering/.
15. See https://indico.cern.ch/event/505613/contributions/2230905/attachments/1346072/2045266/Oral-55.pptx
16. http://www.prnewswire.com/news-releases/aic-announces-release-of-new-white-paper-on-using-nvme-storage-servers-to-build-high-performance-scale-out-storage-for-hpc-applications-300364919.html
17. https://www.beegfs.io/content/
18. http://thinkparq.com/
19. http://www.slac.stanford.edu/cgi-wrap/getdoc/slac-tn-15-001.pdf
20. http://es.net/. It runs a state-of-the-art 100Gbps national backbone connecting the various SC national laboratories, related research institutions, and to other parts of the Internet.
21. http://es.net/engineering-services/oscars/
22. http://slac.stanford.edu/pubs/slactns/tn05/slac-tn-16-001.pdf
23. https://escc.es.net/?q=node/3
24. https://downloadcenter.intel.com/download/26749/Intel-SSD-Data-Center-Tool
25. https://www.linkedin.com/in/keith-busch-25336911/
26. http://www.intel.com/content/www/us/en/solid-state-drives/ssd-dc-p4500-brief.html
27. http://www.intel.com/content/www/us/en/solid-state-drives/ssd-dc-p4600-brief.html

Printed in USA ♻ Please Recycle 0917/AK/TLM 336447-001