Distributed Storage Trends and Implications for Cloud Storage Planners

Bottom line: Simultaneous transitions in storage, networking, and compute demand mean NAND storage alone may not meet the needs of disaggregated cloud storage – Intel® Optane™ SSDs can close the gap.

Authors

Andrew Ruffin
Strategy & Business Development Manager

Satvik Vyas
Senior Product Manager

Jacek Wysoczynski
Senior Manager of Product Planning

Frank Hady
Intel Fellow, Chief Optane Systems Architect

Table of Contents

Introduction and summary

Cloud service providers (CSPs) must continually find new ways to manage and scale their storage services in order to meet growing customer performance expectations and capacity demand. CSPs offering infrastructure to customers provide two SSD storage options:

- local storage, where SSDs are physically attached to the compute server itself and,
- disaggregated storage, where SSDs are hosted on a separate server and assigned to the compute server on demand over the network. This white paper will focus on disaggregated storage to serve storage traffic between storage and compute.

CSPs manage their disaggregated storage in a variety of ways, with several large CSPs developing their own storage software. Others leverage open source solutions, such as Ceph, or buy purpose-built solutions such as those offered by VAST Data, Pure Storage, and Lightbits Labs. Whether building or buying a storage solution, a key problem must be solved for disaggregated storage – how to handle and protect incoming data writes. Common solutions today include write buffers composed of NVDIMMs or overprovisioned NAND SSDs where data writes land before moving to bulk storage. However, higher throughput storage networking and the transition to PCIe Gen4, along with increasing performance gaps between NAND and Intel® Optane™ SSDs, has exposed a situation where today's solutions to protecting incoming data writes will no longer work.

This white paper will explore why today's solutions will not work in the transitioning cloud storage landscape. It then outlines a solution using next generation Intel Optane SSDs that CSP storage planners should consider.

Handling and protecting incoming data

With several trends pulling CSPs toward higher performance storage offerings, one key problem in distributed storage has become even more difficult to solve: how to handle and protect incoming data writes. Today, NVDIMMs or overprovisioned NAND SSDs typically serve as write buffers in distributed block storage services. However, as network speed increases, and bulk storage moves to denser but slower and more cost-effective technologies, the required capacity of the write buffer may outgrow the available capacities of NVDIMMs. Further, the sheer number of overprovisioned NAND SSDs required to handle increased write traffic quickly outgrows physical server space. Without a write buffer, performance and useful life of bulk storage significantly degrade. CSP storage planners thus must consider new solutions to handling data writes.

Two ways to address this problem include 1) a power loss imminent (PLI) buffer, and 2) a write buffer. For a PLI buffer, as shown in Figure 1, incoming writes land in

DRAM and are written simultaneously to a fast buffer and to slower bulk storage. While data is aggregated in DRAM for improving bulk storage efficiency, they are made durable in a higher performance non-volatile buffer. This significantly improves time-to-durability, and in turn, overall system performance. If the server were to lose power, while data being written to the slower bulk storage may be lost, the non-volatile PLI buffer would protect and restore that data upon restart. In addition, performance and endurance of bulk storage is improved by replacing small block operations with large block operations.

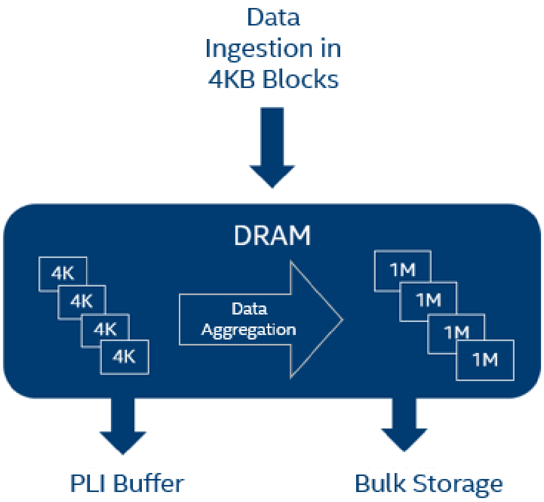


Figure 1. PLI buffer.

As shown in Figure 2, a write buffer is used not only to protect incoming data, but also to aggregate smaller blocks into larger blocks. Once aggregated, data is moved to bulk storage in the background. This process has a potential added benefit: with writes off-loaded to the write buffer, CSPs can optimize bulk storage to service reads more efficiently. This approach enables storage architects to optimize media management policies to offer superior quality of service for reads demanded by attached compute instances. This write buffer improves performance and endurance of bulk storage in the same manner as the PLI buffer. same manner as the PLI buffer.

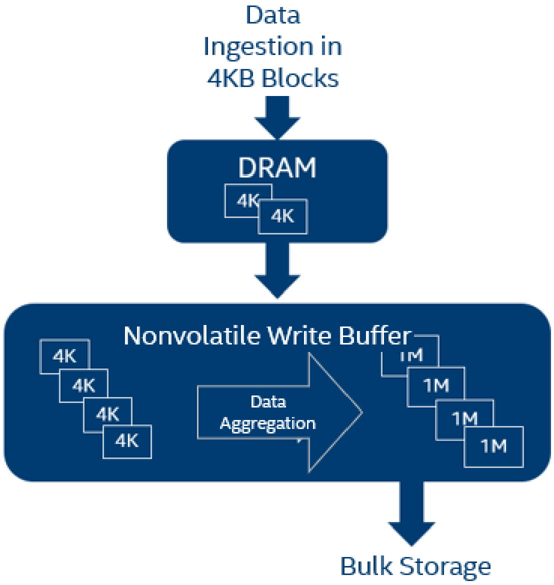


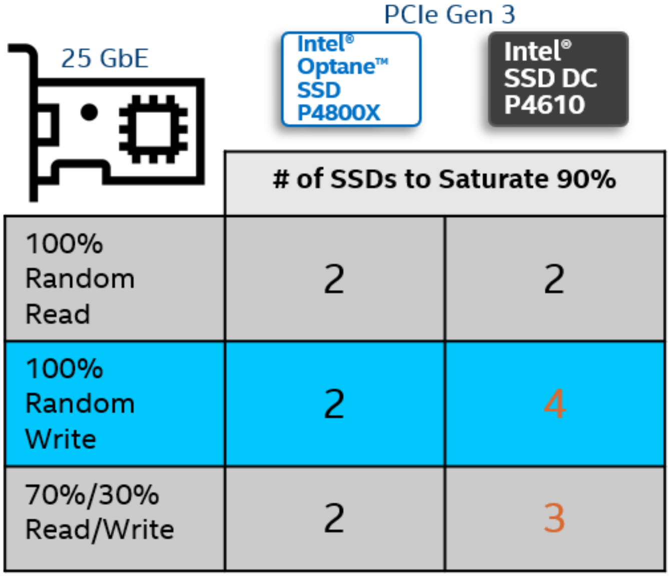
Figure 2. Write buffer.

This discussion will focus on the PLI buffer use case, but the observations are likely applicable to other similar use cases considered by CSPs.

PLI buffer considerations

From a technology perspective, one must narrow down what matters most to solve a problem. For a PLI buffer, random write bandwidth and endurance are most important, with storage capacity necessary but only to the extent required to protect data temporarily housed in system memory. From an operational perspective, one must consider the ability to maintain and service many data center racks full of storage servers, making front-serviceable, hot-swappable form factors (e.g. U.2 or EDSFF SSDs) the most appealing compared to internally housed NVDIMMs.

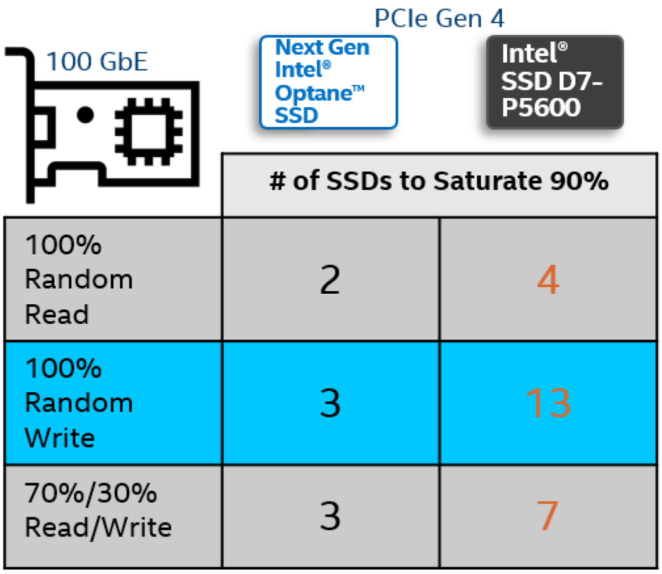
In order to drive operational efficiency at scale, CSPs aim to attain high network utilization,[1] targeting as a best practice up to 90% of their available network bandwidth. As shown in Figure 3,[2] on a 25GbE network, CSPs can handle this 90% threshold with the same number of PCIe Gen3 NAND SSDs or Intel Optane SSDs for 4K random reads. For working data analysis, such as 70/30 read/write workloads, three Gen3 NAND SSDs can fully handle the same network inputs and outputs. Intel Optane SSDs slightly outperform NAND SSDs for 70/30 workloads and halve the number of drives needed for random writes, which are most important to consider for the PLI buffer use case. While a gap does exist between the storage technologies, Intel Optane SSDs were not fully introduced to the market when most decisions were made to build Gen 3 25GbE storage systems.



25 GbE	Intel® Optane™ SSD P4800X	Intel® SSD DC P4610
	# of SSDs to Saturate 90%	
100% Random Read	2	2
100% Random Write	2	4
70%/30% Read/Write	2	3

Figure 3. Intel Optane SSD & NAND SSD 25GbE saturation.

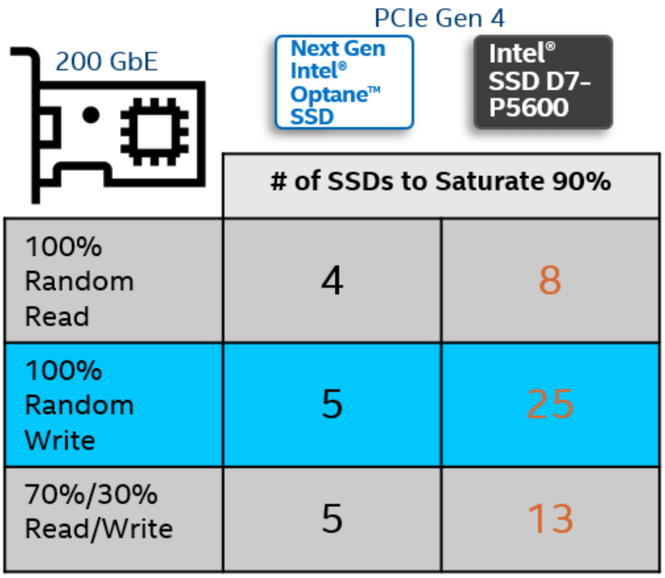
However, with a transition to 100GbE and PCIe Gen4 systems and SSDs, the differentiation between NAND and Intel Optane SSDs becomes clear. On a 100GbE server network, it takes four PCIe Gen4 NAND SSDs and two Gen4 Intel Optane SSDs to reach saturation with 4K random reads. One can see the differentiation for 70/30 workloads below in Figure 4, but the most interesting workload characteristic for the PLI buffer is 100% writes. For this workload, the difference becomes stark, with three Intel Optane SSDs handling 100Gbps of random writes while 13 NAND SSDs are required to do the same.[3]



100 GbE	PCIe Gen 4	
	Next Gen Intel® Optane™ SSD	Intel® SSD D7-P5600
	# of SSDs to Saturate 90%	
100% Random Read	2	4
100% Random Write	3	13
70%/30% Read/Write	3	7

Figure 4. Intel Optane SSDs & NAND SSD 100GbE saturation.

For a CSP considering moving to 200GbE storage server connections in the future, the trend continues. While five Intel Optane SSDs can handle 100% incoming random writes, a CSP would need to deploy 25 TLC NAND SSDs to handle the same workload (see Figure 5).[4] Even if overprovisioning NAND SSDs by 50% (thus increasing performance by 2x), as a PLI buffer, a planner would need to deploy eight fewer Intel Optane SSDs than NAND SSDs. The next section discusses how this setup translates into value from the perspective of a CSP storage planner.



200 GbE	PCIe Gen 4	
	Next Gen Intel® Optane™ SSD	Intel® SSD D7-P5600
	# of SSDs to Saturate 90%	
100% Random Read	4	8
100% Random Write	5	25
70%/30% Read/Write	5	13

Figure 5. Intel Optane SSDs & NAND SSD 200GbE saturation.

Value analysis

While the above technology analysis shows a match between Intel Optane SSDs and PLI buffer usage over alternative technologies, value for performance matters most in making a deployment decision. For CSP planners, overall performance per CapEx and OpEx dollar spent separate one option from another. At scale, typically this analysis is considered at the rack level. In other words, what is the performance per dollar from a data center rack full of storage servers, or even at the full data center level. While cost analyses vary significantly depending on the idiosyncrasies of a specific CSP, this discussion considers a general model that the reader can then modify and apply to their CSP's needs.

Consider a 42U data center rack with a power limit of 24kW and 40U available for 2U storage servers after reserving the top 2U for a top-of-rack switch. Each storage server is dual socket and equipped with dual 100GbE NICs for redundancy (100GbE effective throughput). The goal is to build as dense a storage rack as possible while also satisfying the need to handle a maximum of 90% of 100Gbps of random writes to a PLI buffer in each server. The solution with the best cost per gigabyte of available storage should translate into the best business model for the CSP, adding value for the end user of that disaggregated storage service.

Consider Figure 4 from the prior section of the white paper. In order to handle up to 90% of 100Gbps of incoming random writes in a PLI buffer, a CSP would need to deploy three next generation Intel Optane SSDs compared to 13 TLC NAND SSDs. Assuming seven TLC NAND SSDs can be overprovisioned by 50% to achieve the performance required, one must then consider the amount of bulk storage one can fit in a single server after the PLI buffer requirement is satisfied.

Consider a PCIe Gen4 server with 64 PCIe lanes per socket with a balanced storage and networking configuration as shown above in Figure 6. The NIC would consume 16 lanes, leaving 48 lanes remaining for x4 SSDs, or a total possible 12 NVMe SSDs per socket and 24 per server. Three of the 24 possible SSD slots would be taken by the next generation Intel Optane SSDs, while seven overprovisioned NAND SSDs would be taken in the other option. This leaves room for 21 and 17 bulk storage NVMe SSDs, respectively. In order to reduce the cost of backend storage as compared to TLC NVMe SSDs, QLC NVMe SSDs could be considered, such as the 7.68TB Intel® SSD D5-P4420.

In this case, each server with Intel Optane SSDs for the PLI buffer could incorporate over 161TB of raw storage as compared to just over 130TB of raw storage for the overprovisioned NAND as the PLI buffer. While this approach increases the overall upfront expense per server due to the larger amount of storage density, at the rack level, this translates into a 12.6% improvement in cost per GB of raw storage. This takes into account not just the upfront CapEx but also the OpEx from power usage over 3 years of operation (see the appendix for a detailed description on how this is calculated).

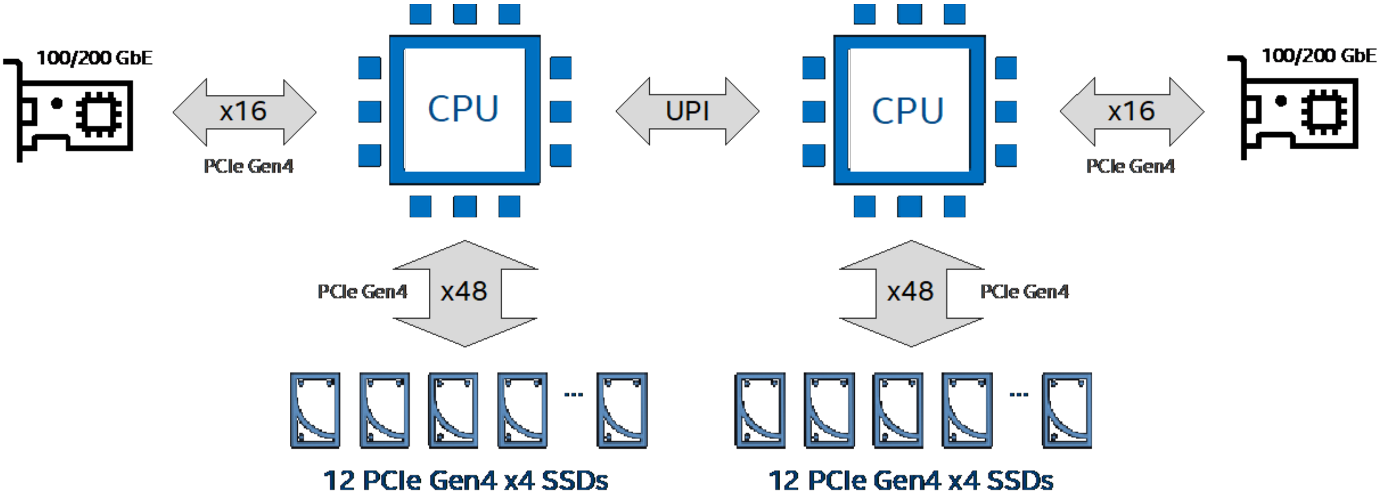


Figure 6. Storage server high level topology.

Some CSPs may choose to not fully populate the PCIe lanes due to blast radius concerns or the addition of other accelerators consuming PCIe lanes. In this scenario, a CSP may want to stay below 100TB of storage per server, deploying 12 QLC SSDs alongside the PLI buffer for a total of just over 92TB of raw storage. Even in this scenario, the Intel Optane SSD PLI buffer option provides a 7.4% improvement in cost per GB of raw storage due to lower CapEx for the PLI buffer as well as lower OpEx with less energy and cooling required over time. Again, see the appendix for the specific configurations and assumptions.

Conclusion

The combination of increased compute demand, denser distributed storage, higher throughput networking, and the transition to PCIe Gen4 with larger gaps in performance between NAND and the Intel Optane SSD P5800X has exposed a situation where previous solutions to protecting incoming data writes may no longer work. This white paper has explored the implications of these transitions for CSP planners to consider as they plan to address customers' demand for storage performance. Further, the paper has conveyed the potential value of next generation Intel Optane SSDs in a distributed storage architecture. Read on in the appendix for further details on the aforementioned trends and for details on the value analysis.

Appendix

Trends dueling demand for cloud storage

CSPs are facing a fourfold, interrelated transition affecting the future of their disaggregated storage services: 1) growing compute demand, 2) growing data storage with CSPs, 3) storage technology differentiation, and 4) customer demand for storage performance.

Compute demand affecting storage requirements

CPU architectures are adapting to demand for increased compute scaling, especially from CSPs that rent CPU cores to their customers. According to Dell EMC, “the era of multi-die scalable server CPUs is here,” which enables “CPU core counts, high-speed I/O lanes, and other features [to grow] faster than historically, fueled by silicon process geometries at 10nm and below.”[5]

With more scaling expected ahead, Intel has worked to meet increased compute demand (and associated memory bandwidth) in the data center specifically for over two decades. Figure 7 below shows the trend in maximum core count for a portion of that history.[6]

Higher CPU core counts and more efficient infrastructure management allow CSPs to monetize more cores per server deployed, and customers will continue to attach remote storage volumes to their compute instances. Growing compute demand necessitates bringing more data to that compute, continuing the ever present need to balance IO traffic to and from compute resources. More instances deployed leads to higher storage and network performance requirements, and the following trends illustrate that direction.

Data growth affecting network requirements

According to IDC, the amount of data created in the next three years will surpass the amount of all data created over the past 30 years, though much of that data is not captured and stored.[7] However, of the data stored, the majority will be in core data centers (both CSPs and enterprise data centers) by this year, growing to 68% by 2024.[8] Of all data storage capacity, IDC estimates CSPs will deploy nearly 40% by 2024, up nearly 2100% from 2015 (see Figure 8 below).[9] Other analysis from Statista states that 48% of corporate data was stored in the cloud as of 2019, up 60% since 2015.[10]

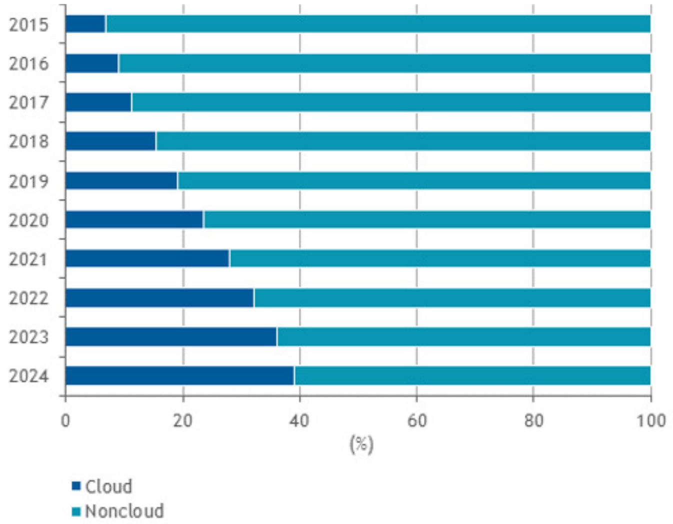


Figure 8. Worldwide global StorageSphere installed base share by cloud versus non-cloud (2015-2024).

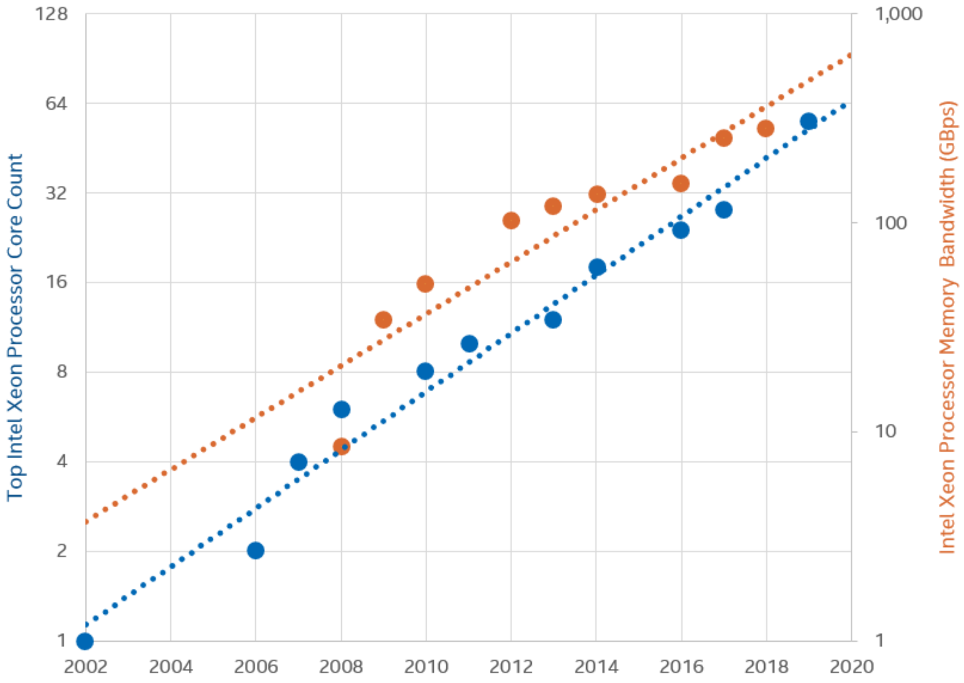


Figure 7. Maximum Intel Xeon CPU core count & memory bandwidth trend over time (log scale).

The increased amount of data stored in disaggregated servers means CSPs must upgrade their networks to handle more traffic between their storage and compute services. Today, CSPs most commonly deploy 25GbE and 50GbE server networking, but the transition to 100GbE and beyond has already begun. According to Dell'Oro Group, several of the largest US-based CSPs are deploying 100GbE networking in 2020 and 2021, with the largest China-based CSPs upgrading their networks in 2021 and 2022.[11] Other CSPs should reach a majority of 100GbE per server speeds by 2024.[12] For CSPs overall, the total server network bandwidth deployed will reach majority 100GbE by 2022, as shown in Figure 9.[13]

Specialized services, such as artificial intelligence and machine learning, high performance computing, and dense compute infrastructure-as-a-service offerings may require even higher throughput disaggregated storage networking of 200GbE and higher.

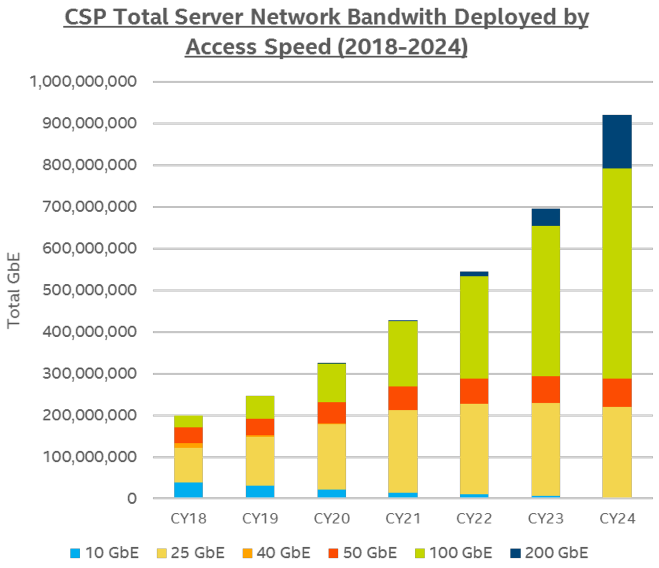


Figure 9. Omdia, total CSP server network bandwidth deployed by access speed (2018-2024).

Storage technology differentiation growing

Intel® Optane™ technology represents the first major memory and storage breakthrough since NAND was first created in 1987. The first storage product based on Intel Optane technology, the Intel® Optane™ SSD DC P4800X, provided significant benefits over NAND-based SSDs. However, the next generation Intel Optane SSD P5800X further widens the gap between NAND-based SSDs and SSDs based on Intel Optane technology. Details on the product's capabilities (i.e. latency, quality of service, endurance, and application performance) were disclosed during the December 2020 Intel Memory & Storage Moment,[14] and in the context of network saturation analysis, one must consider the random IOPS performance of the SSDs involved. Figure 10 details the random read, mixed, and write IOPS possible from PCIe Gen3 and Gen4 NAND and the lowest capacity Intel Optane SSD DC P4800X. While the Intel Optane SSD DC P4800X has approximately 1.5X the mixed IOPS and 2.5X the random write IOPS as compared to a PCIe Gen3 Intel® SSD DC P4610, it lags somewhat at 0.9X on random read performance.[15] Moving ahead, the lowest capacity Intel Optane SSD P5800X shows differentiation across the board, with 2.0X random read, 3.1X mixed IOPS, and 5.0X random write performance as compared to a PCIe Gen4 Intel® SSD D7-P5600.[16] While the lowest capacity Intel Optane SSD P5800X is used in this analysis, the 800GB and 1.6TB versions achieve even higher random performance.

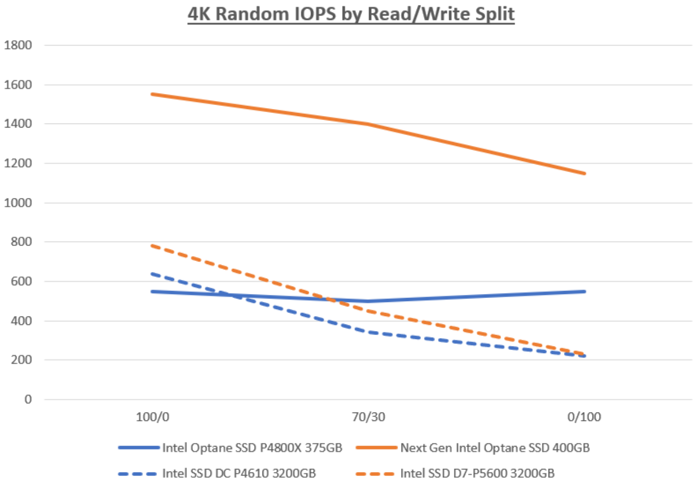


Figure 10. 4K random IOPS by read/write split for PCIe Gen3 and Gen4 NAND & Intel Optane SSDs.

Customers demanding increased storage performance

CSPs have for many years segmented their storage offerings to cater to customers who demand differing levels of storage IO performance, providing a specific service level agreement (SLA) to set expectations. These expectations have grown significantly, most recently exemplified with the addition of a new offering from AWS in their Elastic Block Storage (EBS) service. AWS has previously identified storage IO performance as a pain point for its customers, upgrading the EBS SLA from 30 16KB IOPS/GiB to 50 16KB IOPS/GiB in 2016. In 2020, AWS further increased the SLA by 10x to 500 IOPS/GiB "based on [customer] requests and [an] insatiable desire for more performance."[17]

Value analysis details

In the scenario described in the main section of the white paper, a data center rack has a 24kW power limit with 40U of physical server space with 2U occupied by a top-of-rack switch. Servers are 2-socket 2U each, with up to 24 PCIe U.2 slots available. CPUs are 200W each, with 512GB of DDR4 DRAM deployed consuming 0.28W per GB, and dual 100GbE NICs consuming 15W each. Storage power consumption assumes active power usage per specifications, with a 3200GB Intel SSD D7-P5600 consuming 18W,[18] a 7.68TB Intel® SSD D5-P4420 consuming 15W,[19] and a 400GB Intel Optane SSD P5800X consuming 14W.[20] In both options, 20 servers can fit within the physical space and rack power budget. The next generation Intel Optane SSD buffer servers consume 931W each with 161,280GB of raw storage per server, and the NAND buffer servers consume 955W each, with 130,560GB of raw storage per server. In the scenario with only 12 7680GB Intel SSD D5-P4420 deployed per server, the next generation Intel Optane SSD buffer servers consume 796W each, and the NAND buffer servers consume 880W each, both options with 92,160GB of raw storage per server.

Total power consumption is added across the 20 servers, with the total lifetime kW hours calculated over 3 years, with an assumed power usage effectiveness of 1.2, a $/kWh of $0.10, a cooling factor of 1.5, and utilization of 80%.

Costs for CPU, DRAM, NICs, motherboard, PSUs, and chassis are equal across the options. A single 400GB Optane SSD buffer costs 11% more than a single 3200GB NAND SSD buffer assuming $0.15/GB for the NAND SSD. The 7680GB bulk storage SSDs each cost 140% more than the NAND SSD buffer assuming equal $/GB pricing. Price assumptions are based on Intel projections as of December 2020.

1. Alan Shieh et al., Sharing the Data Center Network, NSDI '11: Proceedings of the 8th USENIX conference on Networked systems design and implementation, March 2011 (online at https://dl.acm.org/doi/10.5555/1972457.1972489).
2. P4800X 375GB 550,000 random read and write 4KB IOPS as per product specifications (online at https://ark.intel.com/content/www/us/en/ark/products/97161/intel-optane-ssd-dc-p4800x-series-375gb-2-5in-pcie-x4-3d-xpoint.html). 500,000 70/30 read/write 4KB IOPS measured – CPU: Intel® Xeon® Gold 6254 3.10GHz 30MB 160W 18 cores per socket, CPU Sockets: 2, BIOS: SE5C620.86B.02.01.0009.092820190230, RAM Capacity: 32G, RAM Model: DDR4, RAM Stuffing: NA, DIMM Slots Populated: 4 slots, PCIe Attach: CPU (not PCH lane attach), Chipset: Intel C610 chipset, Switch/ReTimer Model/Vendor: N/A, OS: CentOS 7.5.1804, Kernel: 4.14.74, FIO version: 3.5; NVMe Driver: Inbox, C-states: Disabled, Hyper Threading: Disabled, CPU Governor (through OS): Performance Mode. Enhanced Intel SpeedStep Technology (Speed Step), Intel Turbo Mode, and P-states = Disabled; IRQ Balancing Services (OS) = Off; SMP Affinity, set in the OS; QD1 utilizes I/O Polling Mode with ioengine=pvsync2/hipri. Test date: December 5, 2019. P4610 3200GB 638,000 random read and 220,000 random write 4KB IOPS as per product specifications (online at https://ark.intel.com/content/www/us/en/ark/products/140104/intel- ssd-dc-p4610-series-3-2tb-2-5in-pcie-3-1-x4-3d2-tlc.html). 340,000 70/30 read/write 4KB IOPS measured – CPU: Intel® Xeon® E5-2699 v4 @ 2.20GHz 55MB 22 Cores, BIOS: SE5C610.8 6B.01.01.0022.062820171903 (Intel Server Board S2600WT), CPU Sockets: 2, RAM Capacity: 32G, RAM Model: DDR4-2137, RAM Stuffing: 1 of 4 channels, DIMM Slots Populated: 4 slots, PCIe Attach: CPU (not PCH lane attach), Chipset: Intel C612 chipset, Switch/ReTimer Model/Vendor: N/A, NVMe Driver: Kernel 4.8.6 (native), C-states: Disabled, Hyper Threading: Disabled, CPU Gover- nor (through OS): Performance Mode, OS: CentOS 7.3, Kernel: 4.8.6. Test date: October 22, 2019.
3. P5800X 400GB 1,550,000 random read,1,150,000 random write, and 1,400,000 70/30 read/write 4KB IOPS as per product introduction (online at https://www.intel.com/content/www/us/en/events/memory-and-storage.html) and as measured – CPU: Intel® Xeon® Gold 6254 3.10GHz 30MB 160W 18 cores per socket, CPU Sockets: 2, BIOS: SE5C620.86B.02.01.0009.092820190230, RAM Capacity: 32G, RAM Model: DDR4, RAM Stuffing: NA, DIMM Slots Populated: 4 slots, PCIe Attach: CPU (not PCH lane attach), Chipset: Intel C610 chipset, Switch/ReTimer Model/Vendor: Intel G4SAC switch (PCIe Gen4), OS: CentOS 7.5.1804, Kernel: 4.14.74, FIO version: 3.5; NVMe Driver: Inbox, C-states: Disabled, Hyper Threading: Disabled, CPU Governor (through OS): Performance Mode. EIST (Speed Step), Intel Turbo Mode, and P-states = Disabled; IRQ Balancing Services (OS) = Off; SMP Affinity, set in the OS; QD1 utilizes I/O Polling Mode with ioengine=pvsync2/hipri. Test date: September 25, 2020. P5600 3200GB 780,000 random read and 230,000 random write 4KB IOPS as per product specifications (online at https://ark.intel.com/content/www/us/en/ark/products/202706/ intel-ssd-d7-p5600-series-3-2tb-2-5in-pcie-4-0-x4-3d3-tlc.html). 450,000 70/30 read/write 4KB IOPS as measured – CPU: Intel® Xeon® Gold 6139 2.30GHz, BIOS: SE5C620.8 6B.00.01.0014.070920180847 (Intel Server Board S2600WFT), CPU Sockets: 2, RAM Capacity: 32G, RAM Model: DDR4-2137, RAM Stuffing: 1 of 4 channels, DIMM Slots Populated: 4 slots, PCIe Attach: CPU (not PCH lane attach), Chipset: Intel C612 chipset, Switch/ReTimer Model/Vendor: Intel G4SAC switch (PCIe Gen4), NVMe Driver: Kernel 4.17.74 (native), C-states: Disabled, Hyper Threading: Disabled, CPU Governor (through OS): Performance Mode, OS: CentOS 7.5, Kernel: 4.14.74, FIO version: 3.5 Test date: March 16, 2020. A 100GbE connection is saturated by 3,125,000 4K IOPS.
4. The same Intel Optane SSD and NAND SSD are used in this analysis as in the 100GbE comparison, with a 200GbE connection saturated by 6,250,000 4K IOPS.
5. Dell EMC, 2020 Server Trends & Observations, January 14, 2020 (online at https://www.dellemc.com/resources/en-us/asset/articles/products/servers/2020-server-trends-and-observations-brief.pdf).
6. Data pulled from ark.intel.com
7. IDC Worldwide Global DataSphere Forecast, 2020–2024: The COVID-19 Data Bump and the Future of Data Growth, Doc # US44797920, April 2020.
8. IDC Worldwide Global StorageSphere Forecast, 2020-2024: Continuing to Store More in the Core, Doc # US46224920, May 2020
9. Ibid. ix. The 2100% metric is based on the StorageSphere installed base growth for cloud from 2015 to 2024 (2015 = 225,809PB; 2024 = 4,982,493PB)
10. Statista, Share of corporate data stored in the cloud in organizations worldwide from 2015 to 2019, April 30, 2020 (online at https://www.statista.com/statistics/1062879/worldwide-cloud-storage-of-corporate-data).
11. Dell'Oro Group, Inc., Controller & Adapter Five Year Forecast Report (2020-2024), January 2020
12. Ibid. xi
13. Omdia, Ethernet Network Adapter Equipment, Quarterly Market Tracker, Q4 2019
14. Intel Memory and Storage Moment, December 15-16, 2020 (online at https://www.intel.com/content/www/us/en/events/memory-and-storage.html).
15. Ibid., ii.
16. Ibid., iii.
17. Jeff Bar, New EBS Volume Type (io2) – 100x Higher Durability and 10x More IOPS/GiB, AWS News Blog, August 24, 2020 (online at https://aws.amazon.com/blogs/aws/new-ebs-volume-type-io2-more-iops-gib-higher-durability/).
18. Intel SSD D7-P5600 3.2TB specifications online at https://ark.intel.com/content/www/us/en/ark/products/202706/intel-ssd-d7-p5600-series-3-2tb-2-5in-pcie-4-0-x4-3d3-tlc.html.
19. Intel SSD D5-P4420 7.68TB specifications online at https://ark.intel.com/content/www/us/en/ark/products/192906/intel-ssd-d5-p4420-series-7-68tb-2-5in-pcie-3-1-x4-3d2-qlc.html.
20. Intel Optane SSD P5800X information provided at 2020 Intel Memory and Storage Moment (online at https://www.intel.com/content/www/us/en/events/memory-and-storage.html). Intel technologies may require enabled hardware, software or service activation.