SOLUTION BRIEF

Software-defined Storage
Intel® Optane™ DC SSDs

Accelerate Ceph* Clusters with Intel® Optane™ DC SSDs

Today's data-intensive workloads demand more throughput and lower latency. Intel® Optane™ DC SSDs can reduce overall costs while significantly increasing cluster performance.

This solution brief describes how to solve business challenges through investment in innovative technologies.

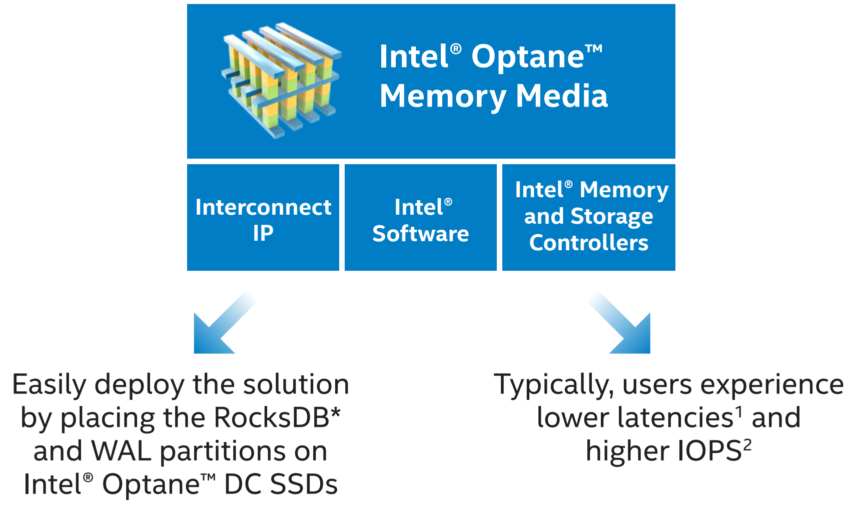
If you are responsible for...

- **Business strategy:** You will better understand how a Ceph* cluster with Intel® Optane™ DC SSDs will enable you to successfully meet your business outcomes.
- **Technology decisions:** You will learn how a Ceph cluster with Intel Optane DC SSDs works to deliver IT and business value.

Executive Summary

With workloads becoming more data-intensive and performance-sensitive, enterprises using Ceph* with BlueStore* are seeking ways to accelerate their clusters while staying cost efficient. High latency was acceptable for historical archival Ceph use cases, but it doesn't meet modern Ceph users' needs.

One way to boost Ceph cluster performance is to add one Intel® Optane™ DC SSD per node to the cluster for RocksDB* and the write-ahead log (WAL) partitions as well as optionally one Intel Optane DC SSD for caching. This approach can accelerate all-flash clusters. With Intel Optane DC SSDs, the cluster's latency[1] and I/O per second (IOPS)[2] are improved. The high write endurance of Intel Optane DC SSDs, combined with excellent performance and low latency, make them a natural choice for Ceph users who want high-performance Ceph clusters. If using Intel Optane DC SSDs for caching, the Intel® Cache Acceleration Software (Intel® CAS) that is available for Intel® SSDs accelerates storage performance by caching frequently accessed data and/or selected I/O classes.


Intel® Optane™ Memory Media
Interconnect IP
Intel® Software
Intel® Memory and Storage Controllers

Easily deploy the solution by placing the RocksDB* and WAL partitions on Intel® Optane™ DC SSDs

Typically, users experience lower latencies[1] and higher IOPS[2]

Figure 1. Overview of Intel® Optane™ DC SSDs in a Ceph* cluster.

Author

Justin Elkow
Solutions Architect
Non-Volatile Storage Group, Intel Corp.

Solution Benefits

- **Improved I/O per second (IOPS)** with Intel® Optane™ DC SSDs[2]
- **Lower read latency** under write load compared to NAND SSDs[3]
- **Higher write endurance** compared to NAND SSDs[4]

Business Challenge: A More Responsive Ceph* Cluster

As more organizations adopt Ceph, performance expectations have increased. Modern database workloads often require single-digit millisecond latency. Other, bursty cloud-based workloads don't need sustained high performance but can require short periods of higher performance. In a recent survey, the Ceph organization found 63 percent of its respondents identified performance as a critical need.[5]

Ceph Block Storage Use Cases on the Rise

Originally, Ceph was used mainly for object storage-based solutions, especially in cold-storage use cases. But more recently, Ceph is being deployed for block-based storage. OpenStack* users deploy Ceph over 4x more than the next most popular solution.[6] Ceph's block storage capabilities are useful to users, enterprises, government agencies, and cloud-based customers. Block storage use cases demand low latency and high performance. These use cases include databases and other workloads, which can benefit by adding Intel® Optane™ DC SSDs to an all-flash cluster.

Solution Value: Have It Your Way

By using Intel® SSDs as part of your solution, you can help increase performance, lower cost, and meet or exceed your organizational service level agreement. By judiciously adding the right kind of Intel SSD to your Ceph cluster, you can accomplish one or several of these goals:

- **Increasing IOPS.** Add Intel Optane DC SSDs to increase IOPS per node[7] and reduce costs through node consolidation[2] while reducing latency. Node consolidation can save on capital expenditures plus power, cooling, and rack space requirements.
- **Reducing latency.** Displace an all-SATA flash array with Intel Optane DC SSDs plus Intel® QLC 3D NAND SSDs. This solution increases IOPS and reduces latency.[2]

In particular, Intel Optane DC SSDs provide the most benefit when used for the metadata tier (RocksDB* and write-ahead log (WAL)), and caching of object storage daemons (OSDs). Below are two examples of the business value Intel Optane DC SSDs can provide when added to a Ceph cluster.

Node Consolidation, Lower Latency, and Lower Cost

On a 19-node Ceph cluster with an all-flash SATA capacity tier, adding a single Intel® Optane™ SSD DC P4800X for RocksDB/WAL/cache reduced the required number of nodes to seven, while reducing latency up to 70 percent.[8] The new configuration has a potential three-year savings of about 60 percent compared to the 19-node SATA all-flash configuration (see Figure 2).[8]

Cost-Neutral Solution with Lower Latency and Higher IOPS

Another approach is to select Intel QLC 3D NAND SSDs for capacity, and instead of using all Intel® 3D NAND SSDs, add Intel Optane DC SSDs for RocksDB and WAL. In this scenario, the upgraded cost per node stays within 2 percent of the original cost, while latency decreases by up to 50 percent, and IOPS increases by about 40 percent.[2]

A Closer Look at Quad-Level Cell NAND

Quad-level cell (QLC) NAND is a recent storage innovation that provides an alternative to its predecessor, tri-level cell (TLC). Compared to TLC, QLC has 33 percent higher density of storage, with four 0 or 1 states in each flash cell.[9] Which generation is best for a particular application depends on what is most important: density or endurance. Because it is denser, QLC can help lower storage costs. On the other hand, QLC has limited write endurance, and is primarily best suited for use cases where read operations far outnumber write operations.[10]

Ceph* Cluster Storage Consolidation with Intel® Optane™ DC SSD vs. SATA All-Flash

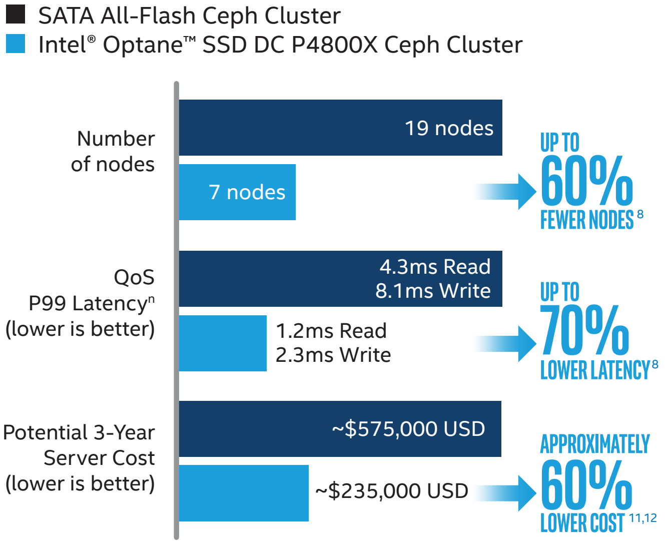


Figure 2. Using Intel® Optane™ DC SSDs for metadata tier RocksDB*, write-ahead log (WAL), and optional object storage daemon (OSD) caching helps Ceph* users consolidate nodes, lower latency, and control costs.

Solution Architecture: Accelerated Ceph Nodes

Intel Optane DC SSDs are a combination of Intel® Optane™ memory media, an advanced system memory controller, interface hardware, and firmware. These SSDs are a new class of storage (that is, they are not NAND-based).

A standard Ceph node uses the same media for all data, both hot and cold (see Figure 3). When the capacity tier uses all-flash SSDs, using Intel Optane DC SSDs for RocksDB, WAL, and optional OSD caching can improve cluster performance and cost efficiency.

Two aspects of Intel Optane DC SSDs enable you to reduce your node count to less than half that of an all-flash solution:

- Consistent lower read latency under increasing write pressure. Intel Optane DC SSDs can maintain up to 63x lower latency than 3D NAND at high write pressure.[13]
- With an endurance rating of up to 60 drive writes per day (DWPD), you can use fewer Intel Optane DC SSDs compared to 3D NAND SSDs (over-provisioned) to handle the high demands of Ceph metadata.[4]

Implementing a cache using Intel Optane SSD DC P4800X Series is easy, because Intel® Cache Acceleration Software (Intel® CAS) has been optimized to unleash Intel® SSD Data Center Family performance. Intel CAS is enterprise-quality software that runs on Linux* or Windows* platforms and accelerates storage performance by caching frequently accessed data and/or selected I/O classes.

CEPH USERS WANT PERFORMANCE[5]

63 percent of Ceph users identified performance as a top need going forward.

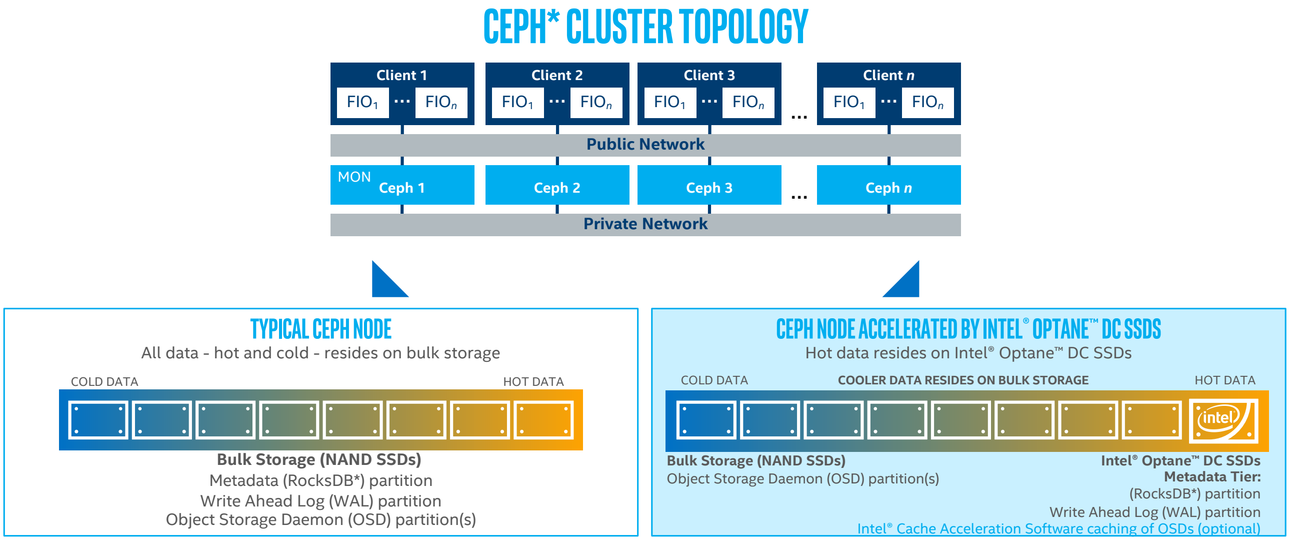
Intelligent Caching Speeds Performance

Intel® Cache Acceleration Software (Intel® CAS) provides a unique, intelligent way to cache based on I/O classification, providing the ability to cache the “hottest” data (such as metadata). Intel CAS-Linux* requires no application or storage system modifications. Intel CAS-Linux is installed as a loadable kernel module, deployed at the block layer, and configured using a user-space administration tool. Features of Intel CAS-Linux include:

- Validated on common Linux distributions and kernels and Windows Server*
- Boosts performance with a very small cache
- Offers several caching modes and cache-cleaning policies, which allows tuning to the user’s workload
- Ability to cache small random blocks/files using I/O classification
- Support for atomic writes and TRIM operations
- Available to download and use with Intel® Optane™ SSDs

Visit intel.com/cas for more information.

Where Intel® Optane™ DC SSDs Fit in a Ceph* Cluster


CEPH* CLUSTER TOPOLOGY
Client 1
Client 2
Client 3
Client *n*
FIO$_1$ ··· FIO$_n$
...
Public Network
MON
Ceph 1
Ceph 2
Ceph 3
Ceph *n*
Private Network
TYPICAL CEPH NODE
All data - hot and cold - resides on bulk storage
COLD DATA
HOT DATA
Bulk Storage (NAND SSDs)
Metadata (RocksDB*) partition
Write Ahead Log (WAL) partition
Object Storage Daemon (OSD) partition(s)
CEPH NODE ACCELERATED BY INTEL® OPTANE™ DC SSDS
Hot data resides on Intel® Optane™ DC SSDs
COLD DATA
COOLER DATA RESIDES ON BULK STORAGE
HOT DATA
intel
Bulk Storage (NAND SSDs)
Object Storage Daemon (OSD) partition(s)
Intel® Optane™ DC SSDs
Metadata Tier:
(RocksDB*) partition
Write Ahead Log (WAL) partition
Intel® Cache Acceleration Software caching of OSDs (optional)

FIO = Flexible I/O*; SSD = Solid State Drive

Figure 3. Ceph* cluster topology with Intel® Optane™ DC SSDs.

Conclusion

Intel Optane DC SSDs can improve performance and reduce cost in any Ceph deployment, especially when they are used for the metadata tier. With a small number of Intel Optane DC SSDs as an accelerator, you can improve the performance of all-flash clusters. Adding Intel Optane DC SSDs results in improved performance and node consolidation, and can reduce overall cost.

Find the solution that is right for your organization. Visit **intel.com/optane** or contact your Intel representative**.**

Solution Provided By:

Learn More

You may also find the following resources useful:

- Intel® Optane™ SSD Data Center P4800X Series
- Intel® Xeon® Scalable processors
- Intel® Cache Acceleration Software (Intel® CAS)
- Using Intel® Optane™ Technology with Ceph* to Build High-Performance OLTP Solutions white paper
- Using Intel® Optane™ Technology with Ceph* to Build High-Performance Cloud Storage Solutions on Intel® Xeon® Scalable Processors white paper
- Use Intel® Optane™ Technology and Intel® 3D NAND SSDs to Build High-Performance Cloud Storage Solutions white paper

[1] Reduce Costs and Optimize Performance with Intel® Optane™ Technology in Your Ceph* Cluster presentation; March 2019; Slide 24, "Improve Performance for Similar Cost & Capacity;" digitallibrary.intel.com/content/solutions/us/en/assetdetail.html/content/dam/solutions/ceph-optanesolutionoverview-final-april-2019.pptx. Intel tested: Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. 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 complete information visit intel.com/content/www/us/en/solid-state-drives/optane-ssd-dc-p4800x-brief.html. Performance results are based on testing as of October 2018 and may not reflect all publicly available security updates. See configuration disclosure for details. No product or component can be absolutely secure. NVMe configuration overview: Intel® Xeon® Gold 6142 Processor, Intel® SSD DC P4510, BIOS: 00.01.0013; ME: .00.04.294; BMC: 1.43.91f76955; Intel® Optane™ SSD config: identical with exception of Intel™ Optane® SS DC P4800X for cache/RocksDB/WAL. See detailed configurations in Appendix B of this presentation.

[2] See endnote 1.

[3] See endnote 1.

[4] Intel: Endurance ratings available at intel.com/content/www/us/en/products/memory-storage/solid-state-drives/data-center-ssds/optane-dc-p4800x-series/p4800x-750gb-2-5-inch.html; and ark.intel.com/content/www/us/en/ark/products/97161/intel-optane-ssd-dc-p4800x-series-375gb-2-5in-pcie-x4-3d-xpoint.html; and ark.intel.com/content/www/us/en/ark/products/122509/intel-ssd-dc-p4600-series-2-0tb-2-5in-pcie-3-1-x4-3d1-tlc.html

[5] Ceph, July 2018, Slide 56, “Where Ceph Community Should Focus Its Efforts;” ceph.com/wp-content/uploads/2018/07/Ceph-User-Survey-2018-Slides.pdf

[6] openstack.org/analytics (Open Stack User Survey 2018; see Deployment Decisions tab; “Which OpenStack Block Storage (Cinder) drivers are you using?” chart).

[7] Intel tested: Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. NVMe configuration overview: Intel® Xeon® Platinum 8180 Processor, Intel® SSD DC S4500, BIOS: 00.01.0013; ME: .00.04.294; BMC: 1.43.91f76955; Intel® Optane™ SSD config: identical with exception of Intel™ Optane® SSD DC P4800X for cache/RocksDB/WAL. Source - Reduce Costs and Optimize Performance with Intel® Optane™ Technology in Your Ceph* Cluster presentation; March 2019; Slide 19, "Optimize IOPS/$ with Intel® Optane™ DC SSDs". See detailed configurations in Appendix A of this presentation. https://digitallibrary.intel.com/content/solutions/us/en/assetdetail.html/content/dam/solutions/ceph-optanesolutionoverview-final-april-2019.pptx

[8] See endnote 7.

[9] Intel, 33 percent more bits per cell. TLC (tri-level cell) contains 3 bits per cell and QLC (quad-level cell) contains 4 bits per cell. Calculated as (4-3)/3 = 33% more bits per cell.

[10] ComputerWeekly.com, May 2018, “Storage 101: The final flash generation? QLC vs MLC, TLC, SLC.” computerweekly.com/feature/Storage-101-The-final-flash-generation-QLC-vs-MLC-TLC-SLC

[11] Source – Intel. Estimated HW, MEDIA, MAINT costs = $575,000; Estimated power & infrastructure costs = $35,000.

[12] Source – Intel. Estimated HW, MEDIA, MAINT costs = $235,000; Estimated power & infrastructure costs = $20,000.

[13] See endnote 7.

Intel may make changes to specifications and product descriptions at any time, without notice. Designers must not rely on the absence or characteristics of any features or instructions marked "reserved" or "undefined". Intel reserves these for future definition and shall have no responsibility whatsoever for conflicts or incompatibilities arising from future changes to them. The information here is subject to change without notice. Do not finalize a design with this information.

The products described in this document may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

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.

Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade.

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. 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 complete information visit intel.com/benchmarks.

Cost reduction scenarios described are intended as examples of how a given Intel-based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.

Intel does not control or audit the design or implementation of third party benchmark data or web sites referenced in this document. Intel encourages all of its customers to visit the referenced web sites or others where similar performance benchmark data are reported and confirm whether the referenced benchmark data are accurate and reflect performance of systems available for purchase.

0419/JGAL/KC/PDF 338287-001US