Deliver business insights sooner so you can innovate and grow

28% less time to complete big data workloads*

40% more throughput*

Accelerate performance on Apache Hadoop workloads with Intel Optane DC SSDs and HPE ProLiant DL380 Gen10 servers

An HPE solution with Intel Optane DC SSDs completed a big data workload in 28% less time and provided 40% more throughput than a configuration with only SATA SSDs

To succeed as a data-driven business, you don't just need to mine the huge amounts of data at your disposal—you need to do it as efficiently as possible. Putting high-performing storage to work on more active subsets of data can yield important business insights while helping you maximize the value of your hardware. SATA SSDs have long been a primary option for affordable flash storage. We put them to the test against Intel® Optane™ DC SSDs in HPE ProLiant DL380 Gen10 servers.

In the Principled Technologies data center, we deployed an Apache® Hadoop® cluster running on HPE ProLiant DL380 Gen10 servers. We used the TeraSort benchmark to measure the performance of two different storage configurations: the first using only SATA SSDs, and the second using both SATA and Intel Optane DC SSDs. The configuration with Intel Optane DC SSDs finished the workload in 28 percent less time and provided 40 percent greater throughput compared to the SATA-only configuration. With Intel Optane DC SSDs and HPE ProLiant DL380 Gen10 servers, you could get more out of your hardware with faster job completion times and more quickly gain the data insights you need to continue innovating.

*according to tests we conducted on the TeraSort benchmark, using a Hadoop cluster with HPE ProLiant DL380 Gen10 servers and Intel Optane DC SSDs, compared to the same server clusters using only SATA SSDs

Getting granular with big data

By now, many businesses have recognized the benefits of big data analytics—in a 2019 survey conducted by NewVantage Partners, 92 percent of respondents said they were increasing the pace of their investments in big data and AI.[1] In the same survey, 79 percent of respondents who felt they had achieved "measurable results" with big data pointed to advanced analytics as an area that had helped them do so.[2] For many organizations, taking a more targeted approach to data analytics makes the most business sense.

Say your organization wants to run an ad campaign for a new product. You have access to a decade's worth of consumer data, covering past purchases and online interactions. However, to increase your campaign's effectiveness, you only want to target users who have been active in the last year. Using Hadoop, you run a series of selective jobs based on the subset of last year's data, then use your new data insights to refine your campaign. In our testing, we targeted smaller sets of data to test how adding Intel Optane DC SSDs to a server would affect these analysis jobs. We found that an HPE ProLiant DL380 Gen10 server cluster using Intel Optane DC SSDs completed the workload in 28 percent less time while increasing throughput by 40 percent compared to a configuration using only SATA SSDs. These results suggest that Intel Optane DC SSDs can accelerate performance on subsets of Hadoop data, enabling your business to complete more work and reach important insights sooner.

The benefits of using Intel Optane DC SSDs and HPE ProLiant DL380 Gen10 servers

Intel Optane DC SSDs

The Intel Optane DC SSD is designed to deliver high throughput, low latency, predictably fast service, and high endurance.[3] Intel Optane DC SSDs offer greater endurance than NAND flash technology: up to 60 drive writes per day (DWPD)[4] versus the NAND flash-based Intel SSD D3-S4510 we tested, which offer a maximum of two DWPD.[5] Intel Optane DC SSDs also range in capacity from 375 GB to 1.5 TB, enabling organizations to customize their storage to meet their needs. In our testing, adding 2x 375GB Intel Optane DC SSDs to each node in our Hadoop cluster increased throughput and job completion speeds. For businesses using traditional storage to process big data subsets, swapping out SATA drives for Intel Optane DC SSDs in the cache/performance tier could yield a performance boost.

HPE ProLiant DL380 Gen10 servers

According to HPE, the ProLiant DL380 Gen10 server "delivers the latest in security, performance and expandability."[6] The two-socket server has an adaptable chassis with modular drive bay configuration options. Featuring processors from the Intel Xeon® Scalable processor family, the ProLiant DL380 Gen10 is compatible with Intel Optane SSDs and supports up to 20 NVMe drives. Its 24 DIMM slots can support anywhere from 128 GB to 3 TB of traditional memory, and up to 6 TB of Intel Optane DC Persistent Memory. The HPE ProLiant DL380 Gen10 server also allows the CPU direct access to NVMe storage. HPE designed this direct connection to increase bandwidth and reduce latency, which could lead to faster response times.

Faster performance and higher durability could enable your business to:

- Support demanding, storage-intensive environments for longer, potentially resulting in fewer storage hardware replacements
- Serve more customers, enabling your business to expand its customer base
- Satisfy your users with faster performance even during periods of high use, minimizing customer drop-off

Testing the speed of this Intel and HPE solution

Apache Hadoop uses the Hadoop Distributed File System (HDFS) to store data, which supports tiered data strategies though its storage policies.[7] Our testing explored the impact of using Intel Optane DC SSDs instead of SATA SSDs in the cache/performance tier. Specifically, we measured the performance difference between two configurations where the cache/performance tier consisted either entirely of SATA SSDs or entirely of Intel Optane DC SSDs.

About Apache Hadoop

Developed as a tool for handling big data, Apache Hadoop provides a scalable and efficient platform for large-scale data processing. Using Hadoop, organizations can engage multiple servers to run many concurrent tasks or jobs. The resulting data can provide businesses with important insights that could enable cost reductions, time reductions, new product development and optimized offerings, and smart decision-making.

Storage tier	Usage	Configuration with SATA SSDs	Configuration with Intel Optane DC SSDs
Cache/performance	HDFS hot tier + temporary files	2 SATA SSDs	2 Intel Optane DC SSDs
Capacity	HDFS cold tier	2 SATA SSDs	2 SATA SSDs

In addition to distributed storage provided by HDFS, Hadoop Yarn applications may also use temporary files on local storage. Moving these files to high-performance storage can improve performance in some scenarios. We configured temporary file storage to use the same disks as the HDFS hot tier. The performance benefit of doing this will vary based on cluster geometry, workload, and other variables. Jobs that require at least some temporary storage may benefit from higher bandwidth storage, such as Intel Optane DC SSDs.

To measure the speed at which these Hadoop clusters could sort a dataset, we used the TeraSort workload, which is part of the Intel HiBench suite of benchmarking software. Combining network, I/O, and compute tasks, TeraSort provides a useful indication of how a cluster might handle general-purpose jobs on Hadoop. We ran three tests and report the median results of those three runs. For more details about our configurations and methodologies, see the science behind the report.

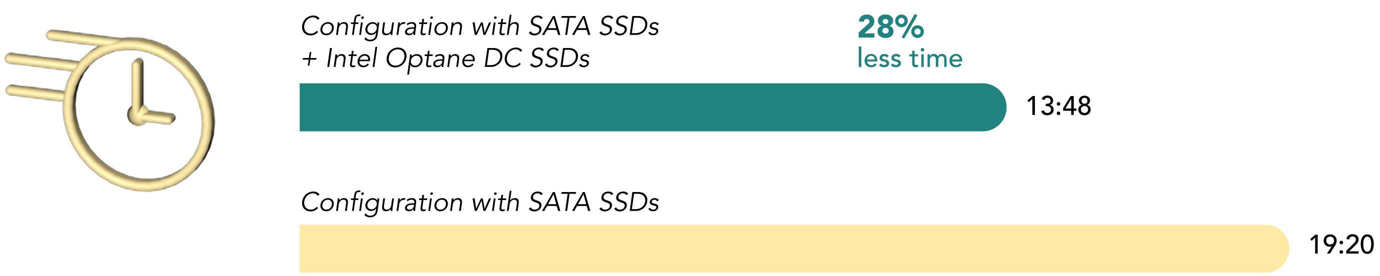
Deliver business insights sooner so you can innovate and grow

28% less time to complete big data workloads

In our testing, the HPE ProLiant DL380 Gen10 server cluster with Intel Optane DC SSDs saved just over five and a half minutes compared to the configuration with only SATA SSDs, completing the TeraSort workload in under 14 minutes. When you have many smaller jobs to run, time savings like this can add up, helping you wrap up projects faster, get the most out of your equipment, and ultimately accelerate the rate at which your business can make strategic, data-driven decisions.

Time to complete TeraSort workload (min:sec)

Lower is better

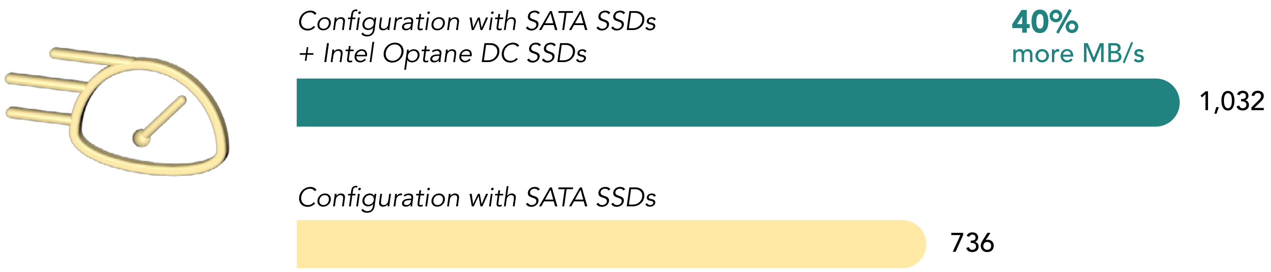


40% more throughput

In addition to completing the workload faster, the configuration with Intel Optane DC SSDs also achieved 40 percent higher throughput than the SATA-only configuration, providing almost 300 more megabytes of data per second. With a solution that can process more data per second, your business could access results in less time, enabling you to apply those insights sooner.

Throughput on TeraSort workload (MB/s)

Higher is better



Conclusion

Proactive businesses are constantly looking for ways to advance project timelines and maximize the value of their data center hardware. Choosing the right server and storage solutions for your big data strategy can help accelerate analysis run times and increase the amount of data processing your teams can achieve. In our hands-on testing of a Hadoop cluster with HPE ProLiant DL380 Gen10 servers, the configuration equipped with Intel Optane DC SSDs processed a TeraSort workload in 28 percent less time and provided 40 percent higher throughput than the same server cluster with only SATA SSDs. With ProLiant DL380 Gen10 servers and Intel Optane DC SSDs, companies could get more out of their active Hadoop datasets.

[1] NewVantage Partners, "Big Data and AI Executive Survey 2019," accessed December 18, 2019, https://newvantage.com/wp-content/uploads/2018/12/Big-Data-Executive-Survey-2019-Findings-Updated-010219-1.pdf.

[2] NewVantage Partners, "Big Data and AI Executive Survey 2019."

[3] Intel, "Breakthrough Performance Expands Datasets, Eliminates Bottlenecks," accessed December 18, 2019, https://www.intel.com/content/www/us/en/products/docs/memory-storage/solid-state-drives/data-center-ssds/optane-ssd-dc-p4800x-p4801x-brief.html.

[4] Intel, "Breakthrough Performance Expands Datasets, Eliminates Bottlenecks."

[5] Intel, "Intel® SSD D3-S4510 and D3-S4610 Series Product Brief," accessed December 18, 2019, https://www.intel.com/content/www/us/en/products/docs/memory-storage/solid-state-drives/data-center-ssds/dc-d3-s4510-s4610-series-brief.html.

[6] HPE, "HPE ProLiant DL380 Gen10 Server," accessed December 18, 2019, https://www.hpe.com/us/en/product-catalog/servers/proliant-servers/pip.hpe-proliant-dl380-gen10-server.1010026818.html.

[7] Apache Hadoop, "Archival Storage, SSD & Memory," accessed December 18, 2019, https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-hdfs/ArchivalStorage.html#Storage_Policies:_Hot.2C_Warm.2C_Cold.2C_All_SSD.2C_One_SSD.2C_Lazy_Persist_and_Provided.

Read the science behind this report at http://facts.pt/nnmqwtj ▶

Principled Technologies®

Facts matter.®

For additional information, review the science behind this report.

This project was commissioned by HPE.