In-Memory Computing SUMMIT NORTH AMERICA 2018

Expansion of System Memory using Intel Memory Drive Technology

Ravi Durgavajhala

Intel

Ravikanth.Durgavajhala@Intel.com

@ravykanth

www.linkedin.com/in/ravikanth-durgavajhala-b651892

Legal Disclaimer

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 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. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest Intel product specifications and roadmaps.

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors.

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.

Contact your local Intel sales office or your distributor to obtain the latest specifications and before placing your product order.

This document contains information on products in the design phase of development.

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase.

Results have been estimated or simulated based on internal Intel analysis and are provided for informational purposes only. Any difference in system hardware or software design or configuration may affect actual performance.

Performance results are based on testing as of Jul 31, 2018 and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure.

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.

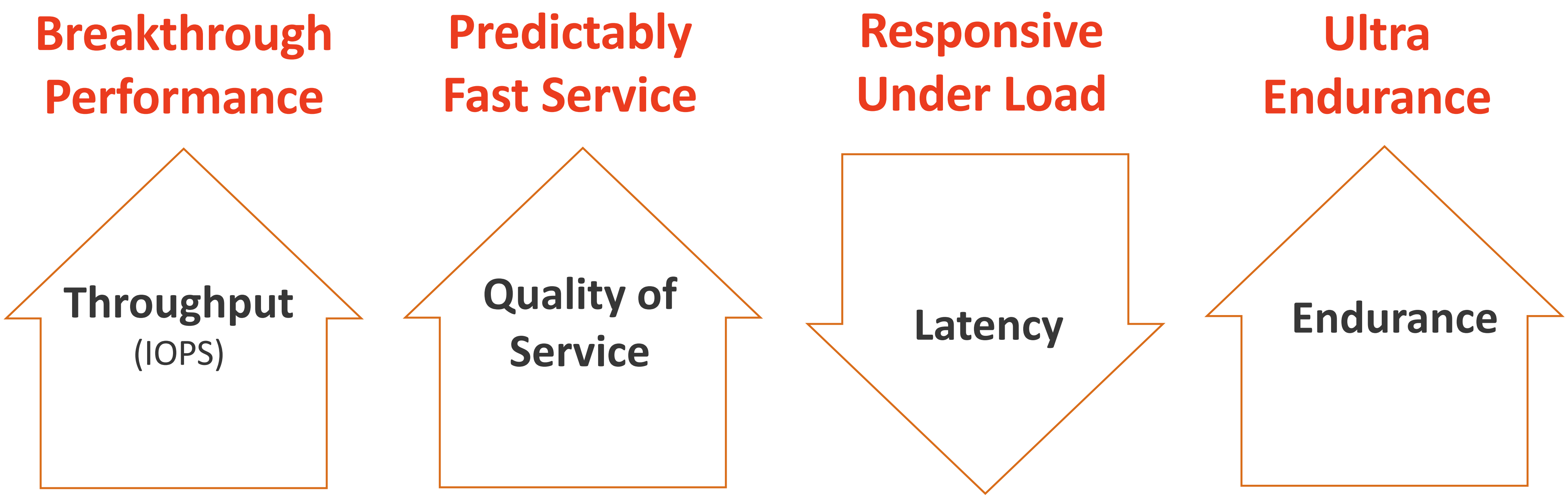
Intel does not control or audit the design or implementation of third party benchmark data or Web sites referenced in this document. Intel encourages all of its customers to visit the referenced Web sites or others where similar performance benchmark data are reported and confirm whether the referenced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

World's Most Responsive Data Center SSD[1]

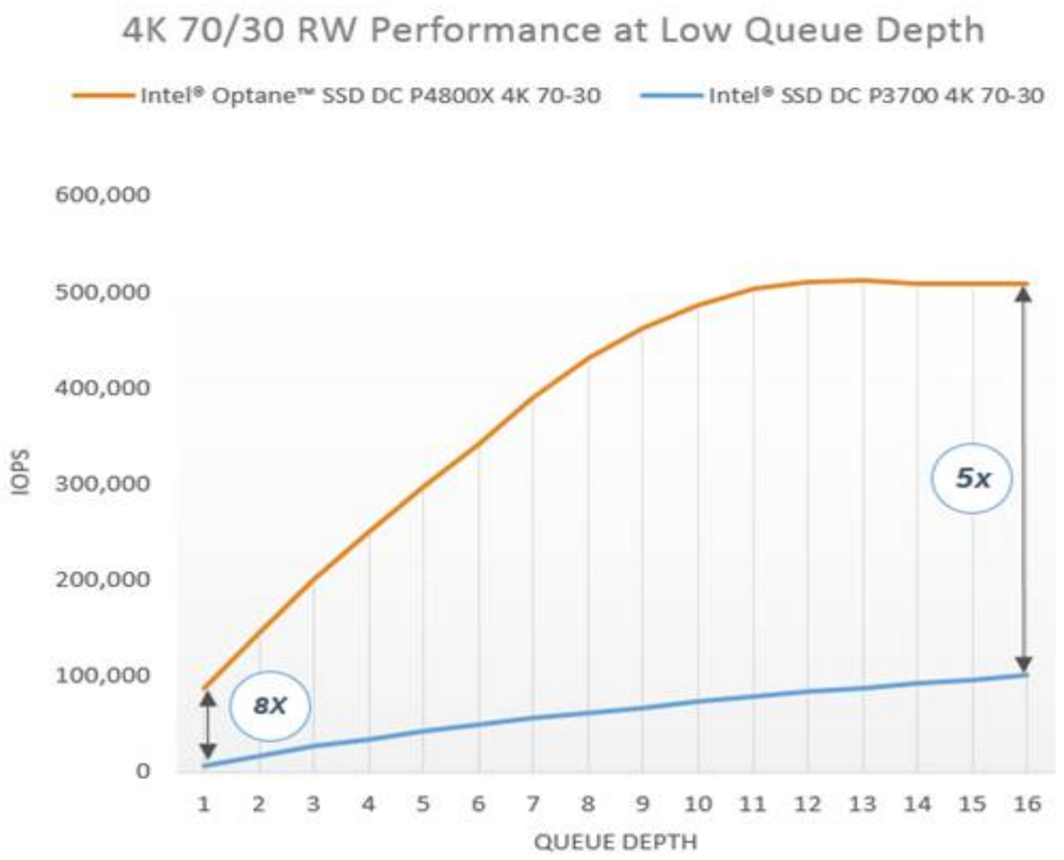
Delivering an **industry leading combination of low latency, high endurance, QoS and high throughput**, the Intel® Optane™ SSD is the first solution to **combine the attributes of memory and storage**. This innovative solution is optimized to **break through storage bottlenecks** by providing a new data tier. It accelerates applications for **fast caching and storage, increasing scale per server** and reducing transaction cost. Data centers based on the latest Intel® Xeon® processors can now also **deploy bigger and more affordable datasets** to gain new insights from larger memory pools.

1. Responsiveness defined as average read latency measured at queue depth 1 during 4k random write workload. Measured using FIO 2.15. Common configuration - Intel 2U PCSD Server ("Wildcat Pass"), OS CentOS 7.2, kernel 3.10.0-327.el7.x86_64, CPU 2 x Intel® Xeon® E5-2699 v4 @ 2.20GHz (22 cores), RAM 396GB DDR @ 2133MHz. Intel drives evaluated - Intel® Optane™ SSD DC P4800X 375GB and Intel® SSD DC P3700 1600GB. Samsung drives evaluated – Samsung® SSD PM1725a, Samsung® SSD PM1725, Samsung® PM963, Samsung® PM953. Micron drive evaluated – Micron® 9100 PCIe® NVMe™ SSD. Toshiba drives evaluated – Toshiba® ZD6300. Test – QD1 Random Read 4K latency, QD1 Random RW 4K 70% Read latency, QD1 Random Write 4K latency using fio-2.15.

Intel® Optane™ SSD DC P4800X



Breakthrough Performance

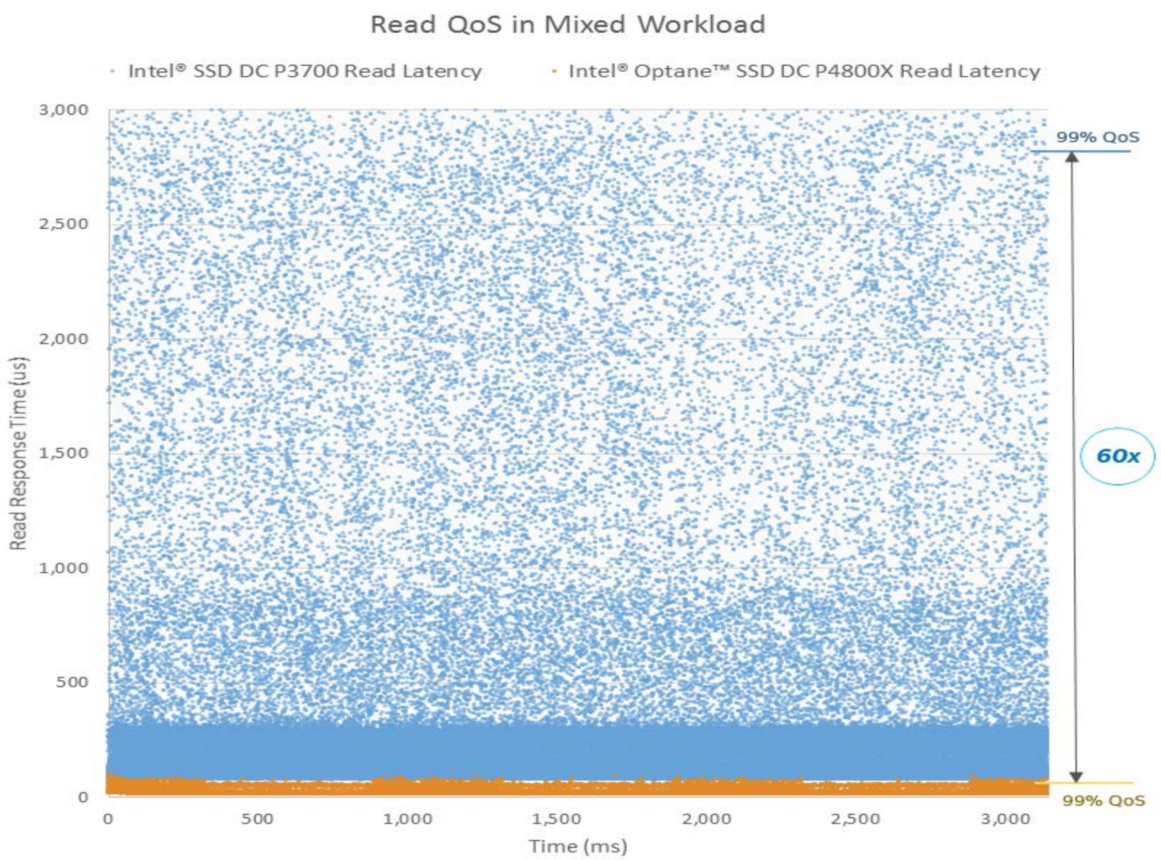


- **5-8x faster** at low Queue Depths[1]
- Vast majority of **applications generate low QD** storage workloads

1. Common Configuration - Intel 2U PCSD Server ("Wildcat Pass"), OS CentOS 7.2, kernel 3.10.0-327.el7.x86_64, CPU 2 x Intel® Xeon® E5-2699 v4 @ 2.20GHz (22 cores), RAM 396GB DDR @ 2133MHz. Configuration – Intel® Optane™ SSD DC P4800X 375GB and Intel® SSD DC P3700 1600GB. Performance – measured under 4K 70-30 workload at QD1-16 using fio-2.15.

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance.

Predictably Fast Service



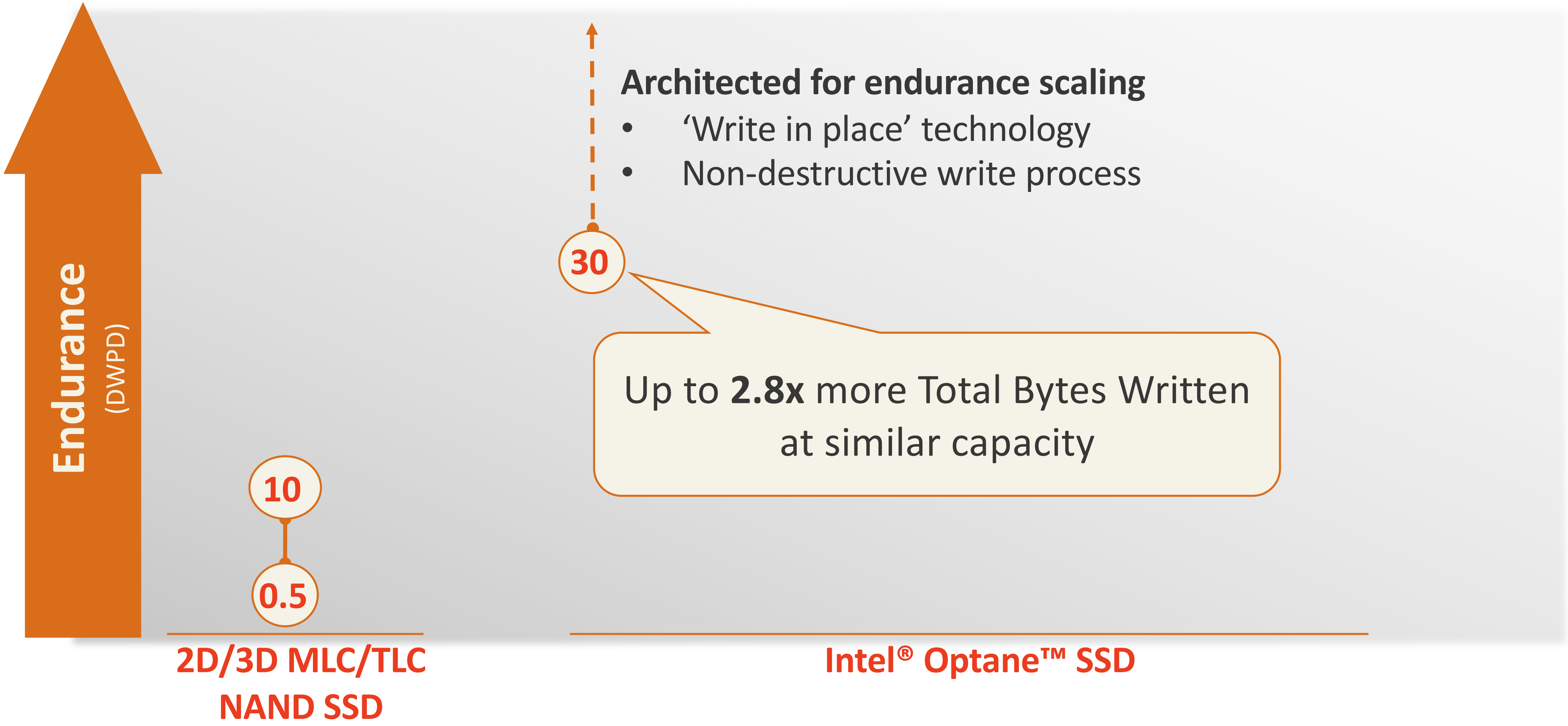
Intel XEON inside + intel OPTANE™

- up to **60X** better at 99% QoS[1]
- Ideal for critical applications with aggressive latency requirements

1. Common Configuration - Intel 2U PCSD Server ("Wildcat Pass"), OS CentOS 7.2, kernel 3.10.0-327.el7.x86_64, CPU 2 x Intel® Xeon® E5-2699 v4 @ 2.20GHz (22 cores), RAM 396GB DDR @ 2133MHz. Configuration – Intel® Optane™ SSD DC P4800X 375GB and Intel® SSD DC P3700 1600GB. QoS – measures 99% QoS under 4K 70-30 workload at QD1 using fio-2.15.

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance.

Ultra Endurance

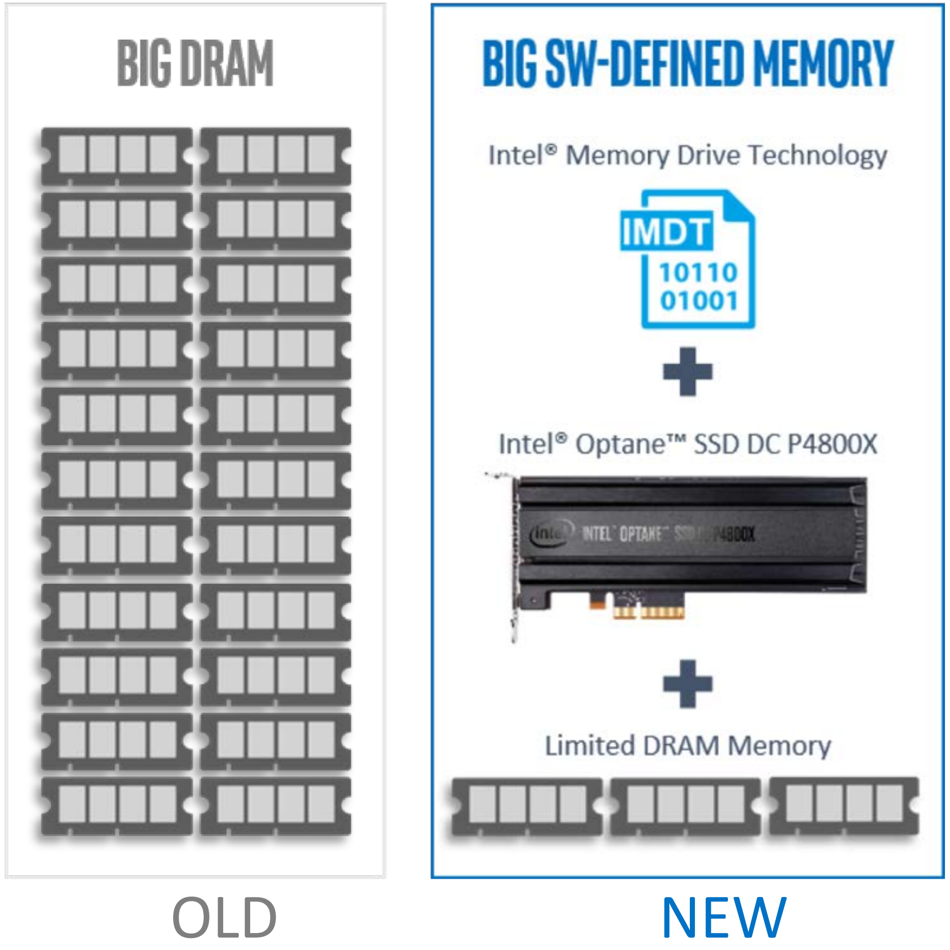


1. Comparing projected Intel® Optane™ SSD 750GB specifications to actual Intel® SSD DC P3700 800GB specifications.
   Total Bytes Written (TBW) calculated by multiplying specified or projected DWPD x specified or projected warranty duration x 365 days/year.

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance.

Introducing Intel® Memory Drive Technology (IMDT)

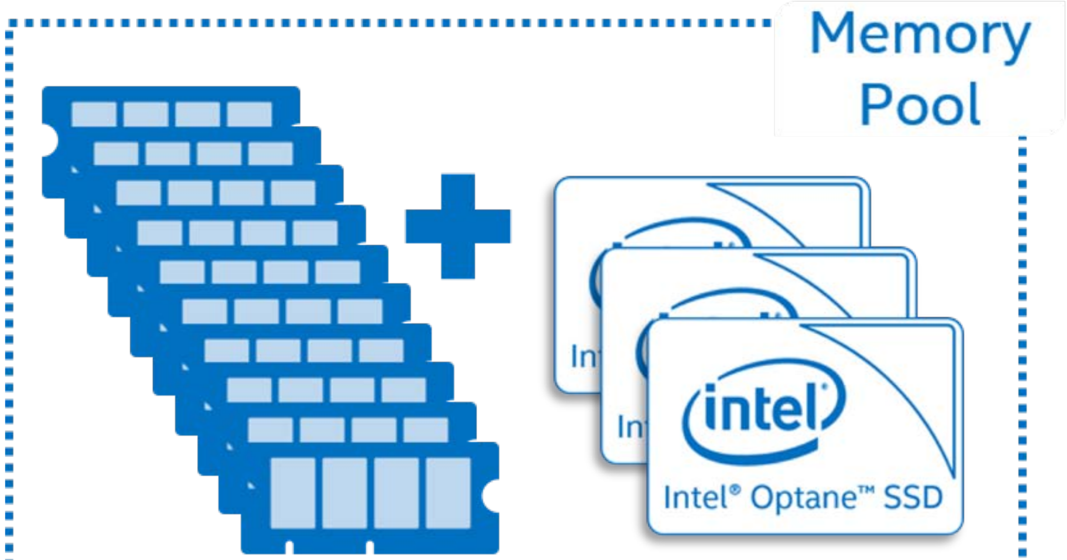
- **Intel® Optane™ Technology** - Write in place, Bit addressable, Low latency
- Use Intel® Optane™ SSD DC P4800X **transparently as memory**
- Grow **beyond system DRAM capacity**, or **replace high-capacity DIMMs** for lower-cost alternative, with **similar performance**
- **Leverage storage**-class memory today!
  - ***No change to software*** *stack: unmodified Linux* OS, applications, and programming*
  - ***No change to hardware****: runs bare-metal, loaded before OS from BIOS or UEFI*



*Other names and brands may be claimed as the property of others

Intel® Memory Drive Technology Delivers Big, Affordable Memory

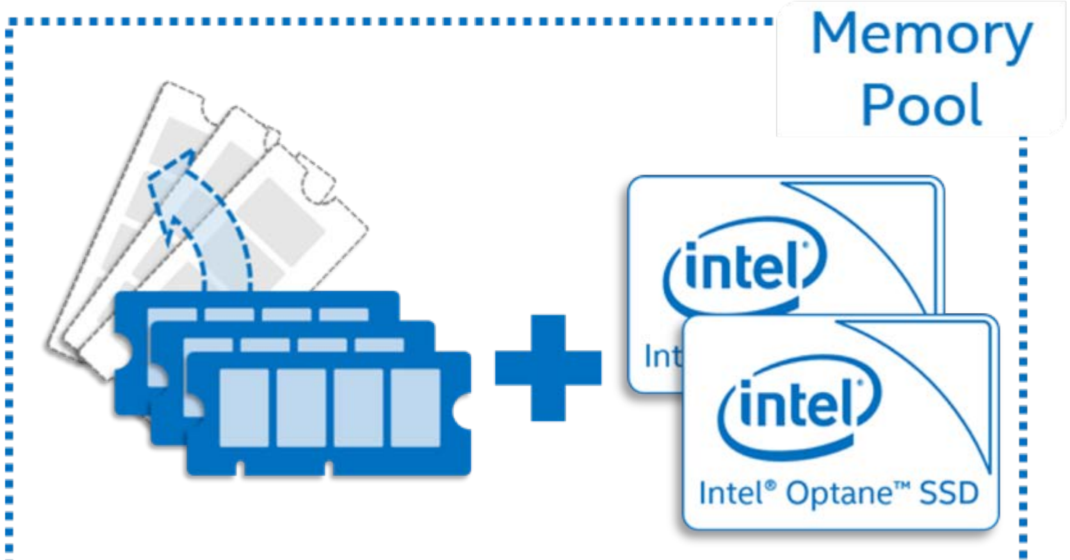
use case 1 — **EXPAND beyond limited DRAM** CAPACITY



INTEL® MEMORY DRIVE TECHNOLOGY

Expand Insights with Massive Data Pools

use case 2 — **Displace dram with** Affordable SSDs



INTEL® MEMORY DRIVE TECHNOLOGY

Reduce High-capacity DRAM CAPEX Expenditures

Note: Intel® Memory Drive Technology supports Linux* x86_64 (64-bit), kernels 2.6.32 or newer.

*Other names and brands may be claimed as the property of others

Problem Statement

Optimize the performance of Spark* and get more out of my infrastructure while operating within the budget.

Assumptions

- Extrapolate overall infrastructure set up.
- Match the individual system resources to that of real-world production, as much as possible.
- Come up with a representative workload.
- Identify a solution along with alternatives.

A quick overview of the K-Means workload

“**Definition**: K-Means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean.

Standard Algorithm: “Given an initial set of k means m1(1),…,mk(1), the algorithm proceeds by alternating between two steps:

1. **Assignment step**: Assign each observation to the cluster whose mean has the least squared **Euclidean distance**, this is intuitively the "nearest" mean.
2. **Update step**: Calculate the new means to be the centroids of the observations in the new clusters.

The algorithm has converged when the assignments no longer change.”

https://en.wikipedia.org/wiki/K-means_clustering

Hardware Configuration

	Master Node	Data Node (x3)
CPU	Intel® Xeon® Gold 6140 CPU @ 2.30GHz	Intel® Xeon® Gold 6140 CPU @ 2.30GHz
Cores per Socket	18	18
Sockets	2	2
Threads per Core	2	2
Total vcores	72	72
Memory	192GB	192GB
SSD	None	3.7TB Intel® SSD DC P4500 (x2)
		375GB Intel® Optane™ SSD DC P4800X (x2)
Network	10Gbps	

* Other names and brands may be claimed as the property of others

Software Configuration

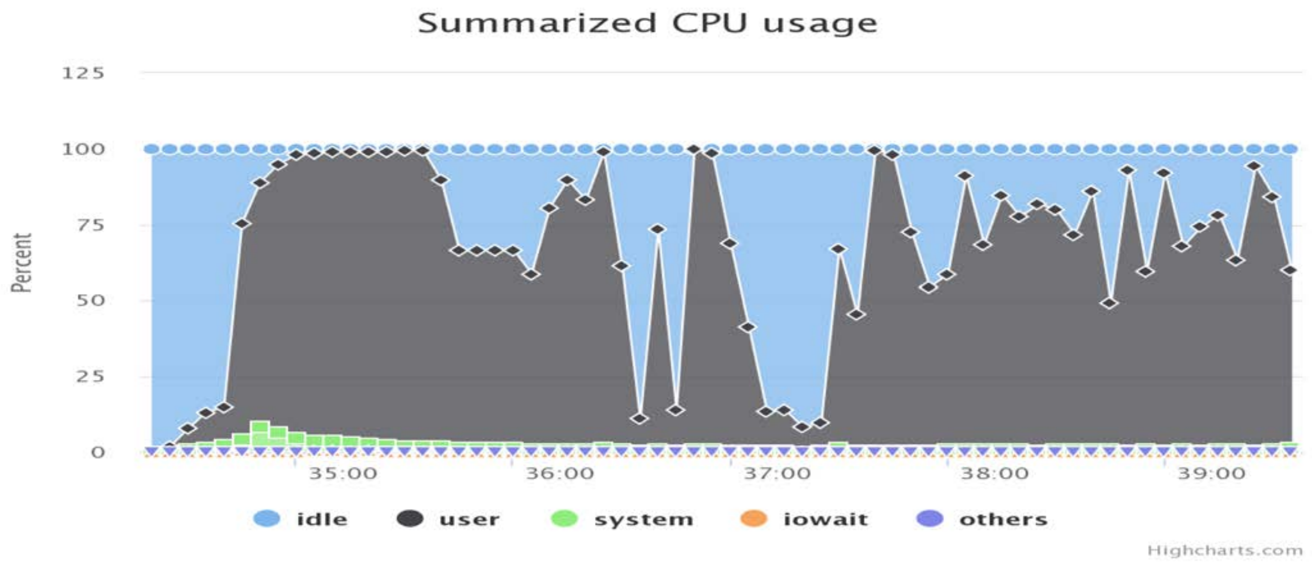
Stack	Version
Distribution	HDP 2.6.4.0
HDFS*	2.7.3
YARN*	2.7.3
Spark*	2.2.0
OS	CentOS 7.4*
Kernel	4.15.12

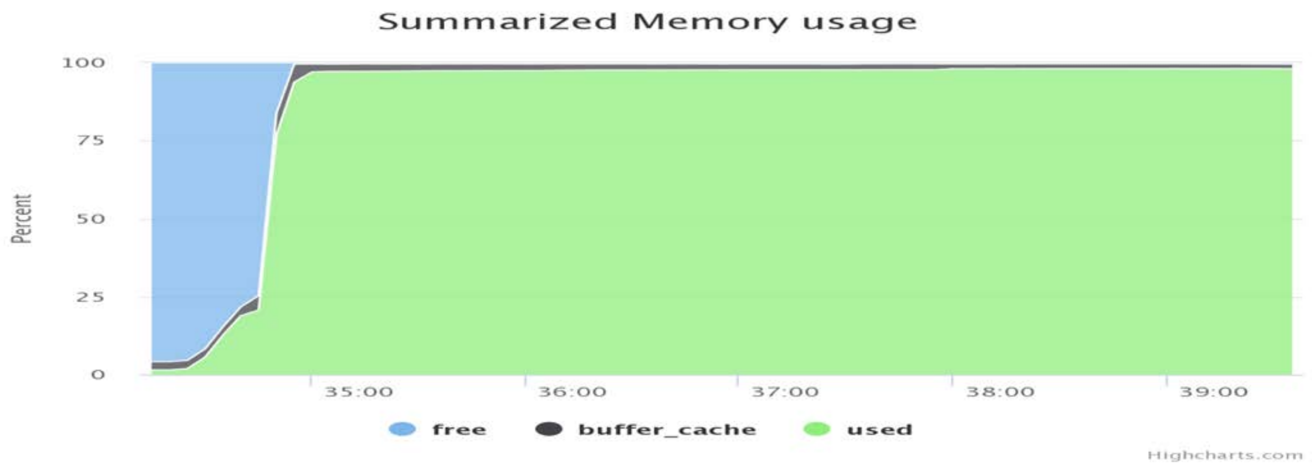
Workload that fits entirely into DRAM

Spark* Workload Configuration	
# of Executors across all Nodes	42
# of Cores per Executor	5
Memory per Executor	12 GiB
Memory Overhead per Executor	3 GiB
Driver Memory	1 GiB
Driver Memory Overhead	1 GiB
K-Means workload Scale Factor	1.2 Billion samples

Time taken to run the workload is **5.3** min[1]

- Spark* configuration is based on generally understood guidelines.
- Data set fits entirely into memory, without any spill.
- The objective is to utilize maximum available resources on the system to get best possible run-time.





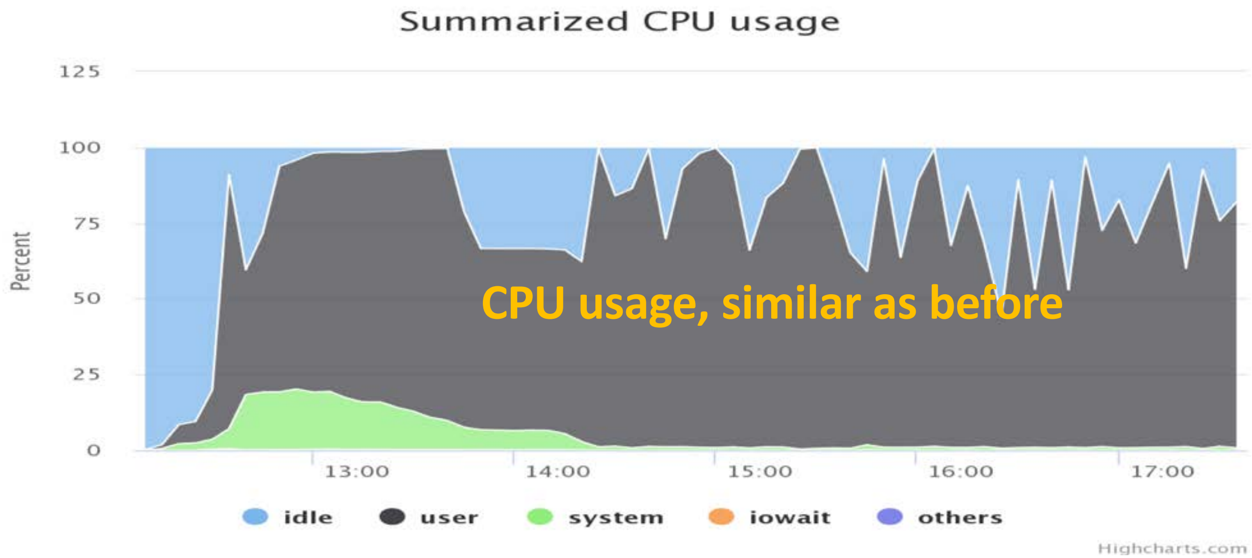
[1] For system configuration details, please refer to Slide #5. Performance results are based on testing as of Jul 31, 2018 and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure.

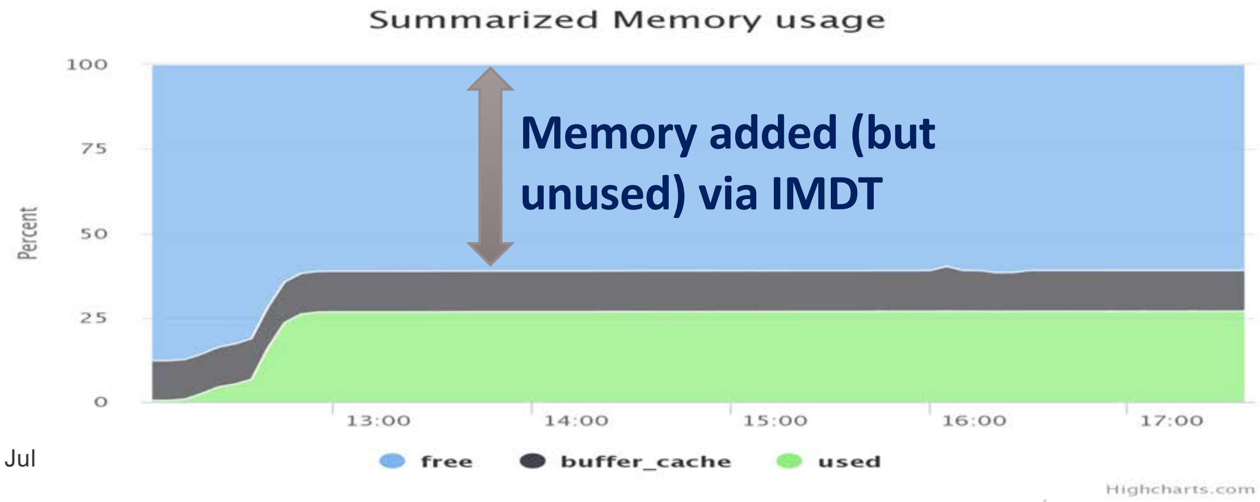
*Other names and brands may be claimed as the property of others

Workload that fits entirely into DRAM (+IMDT)

Spark* Workload Configuration	
# of Executors across all Nodes	42
# of Cores per Executor	5
Memory per Executor	12 GiB
Memory Overhead per Executor	3 GiB
Driver Memory	1 GiB
Driver Memory Overhead	1 GiB
K-Means workload Scale Factor	1.2 Billion samples
Time taken to run the workload is **5.4** min[1]	

- Objective is to ensure performance did not get impacted when running the same workload using same resource configuration, except for memory expansion using IMDT.





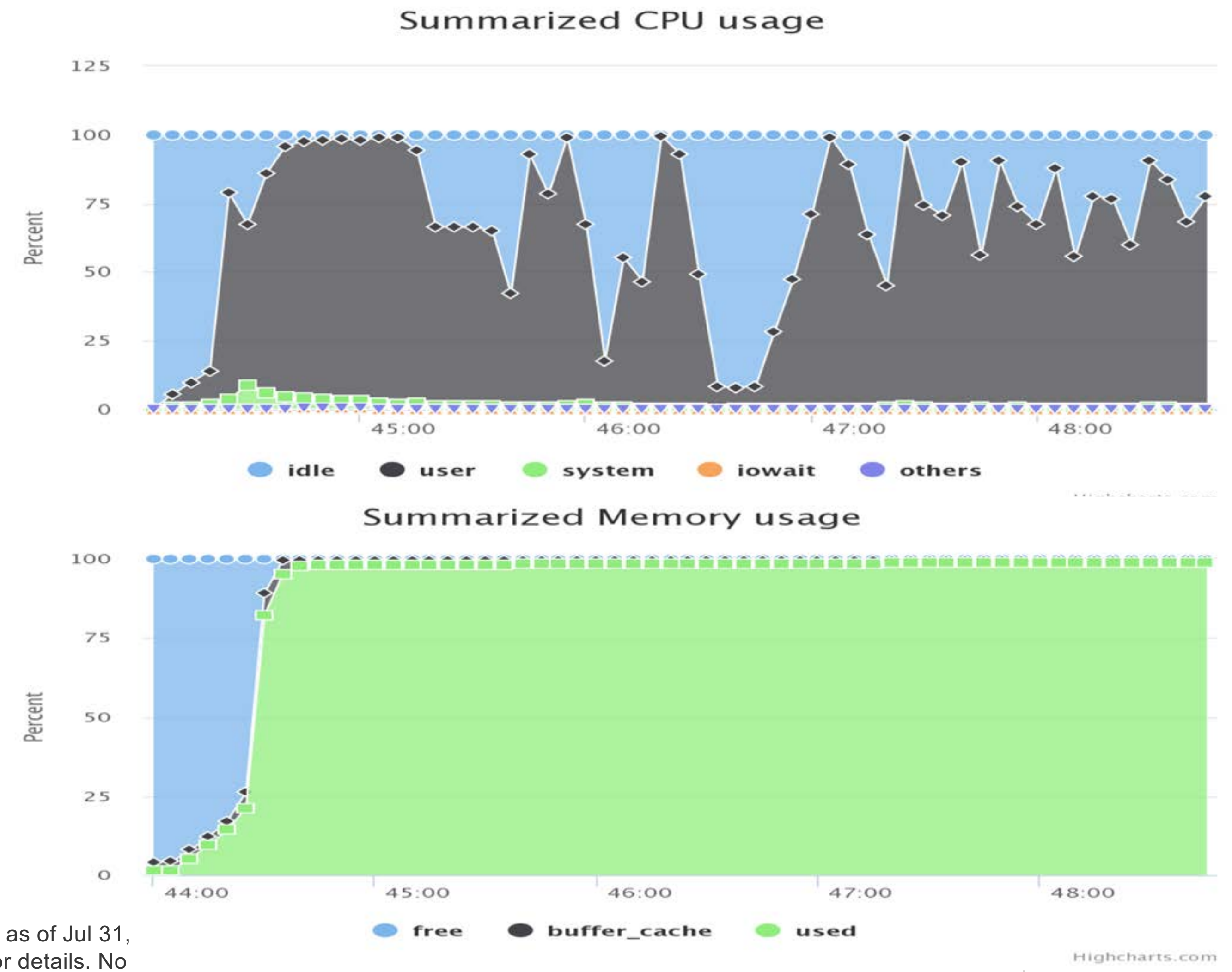
[1] For system configuration details, please refer to Slide #5. Performance results are based on testing as of Jul 31, 2018 and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure.

Workload that fits entirely into DRAM – fine tuned

Spark* Workload Configuration	
# of Executors across all Nodes	30
# of Cores per Executor	7
Memory per Executor	17 GiB
Memory Overhead per Executor	3 GiB
Driver Memory	1 GiB
Driver Memory Overhead	1 GiB
K-Means workload Scale Factor	1.2 Billion samples

Time taken to run the workload is **4.8** min[1]

- Spark* configuration is fine tuned based on Memory and CPU utilization.
- Not all workloads are alike, so each workload needs to be custom-adjusted for better resource utilization.

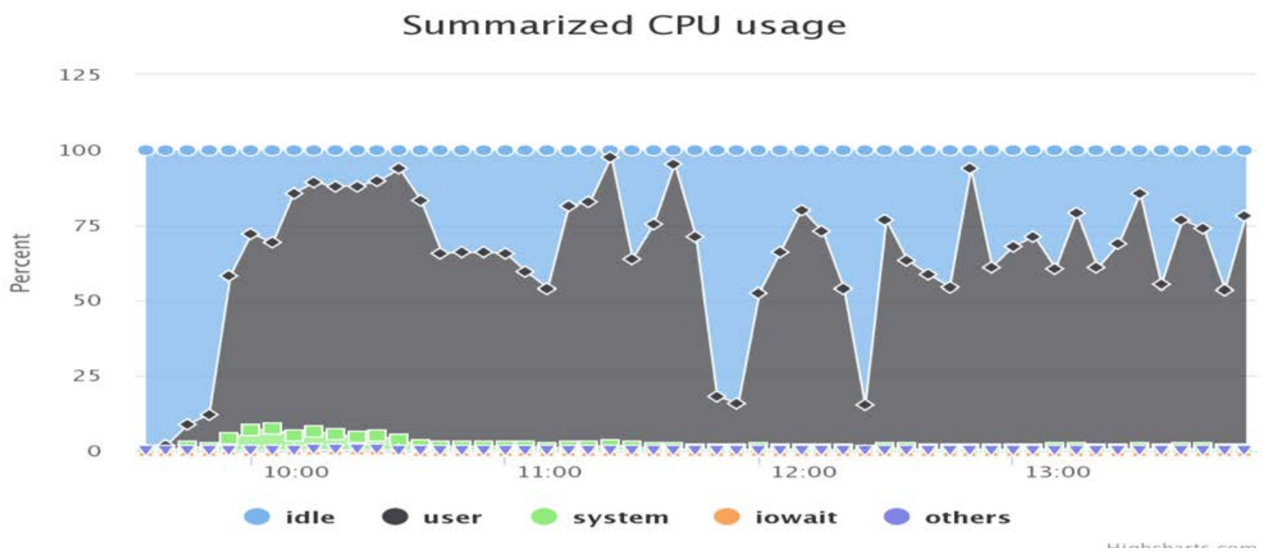


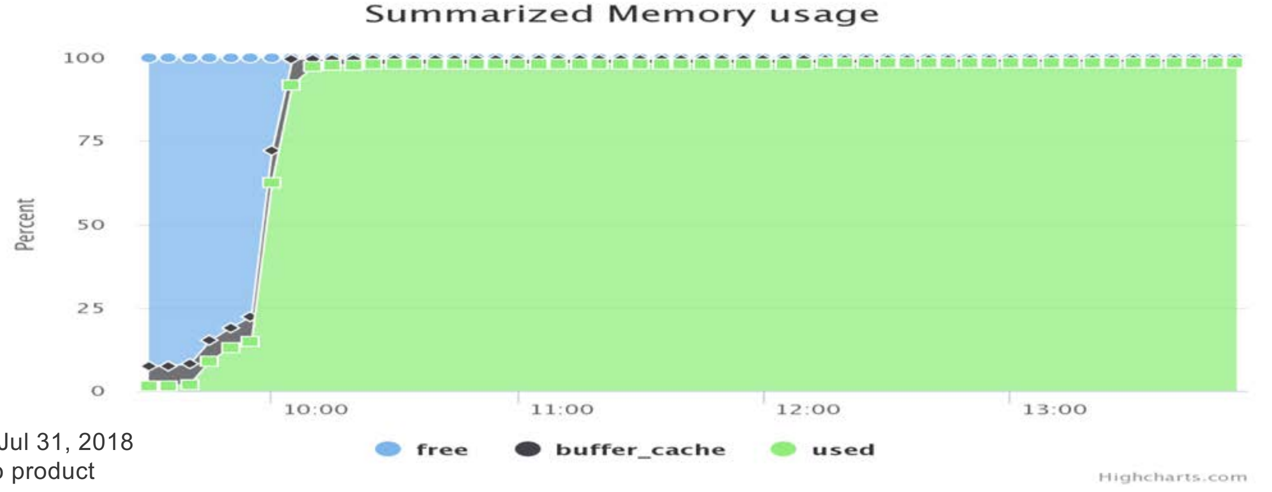
[1] For system configuration details, please refer to Slide #5. Performance results are based on testing as of Jul 31, 2018 and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure.
*Other names and brands may be claimed as the property of others

Workload that fits entirely into DRAM – fine tuned

Spark* Workload Configuration	
# of Executors across all Nodes	30
# of Cores per Executor	5
Memory per Executor	17 GiB
Memory Overhead per Executor	3 GiB
Driver Memory	1 GiB
Driver Memory Overhead	1 GiB
K-Means workload Scale Factor	1.2 Billion samples

Time taken to run the workload is **4.4** min[1]





- Utilizing max of resources available does not always yield best possible performance.
- Performance varies based on memory and other resource utilization within the application code.

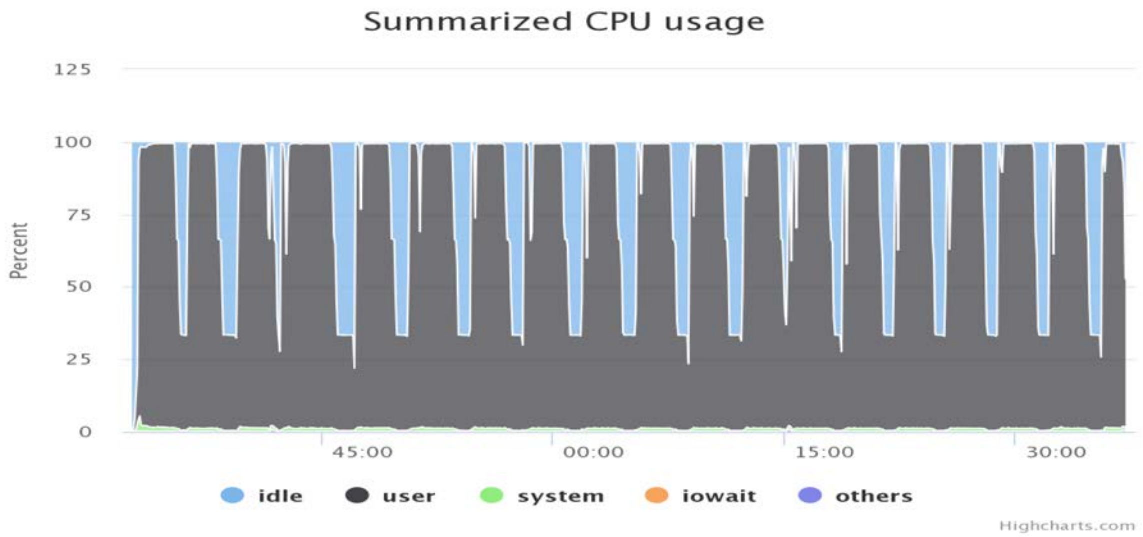
[1] For system configuration details, please refer to Slide #5. Performance results are based on testing as of Jul 31, 2018 and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure.

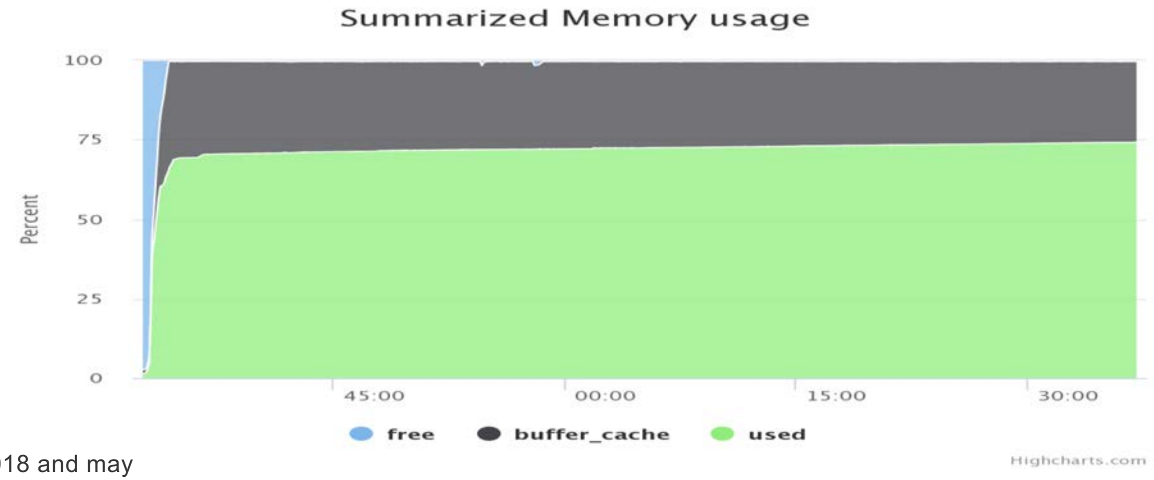
Bigger Workload using DRAM

Spark* Workload Configuration	
# of Executors across all Nodes	30
# of Cores per Executor	7
Memory per Executor	12 GiB
Memory Overhead per Executor	3 GiB
Driver Memory	1 GiB
Driver Memory Overhead	1 GiB
K-Means workload Scale Factor	2 Billion samples

Time taken to run the workload is **66** min[1]

- Spark* shuffles the data between memory and storage when dataset does not fit entirely in memory.
- If the workload is large enough that it cannot fit with fully populated memory channel, the next logical move is to scale out and add more nodes.
- Storage: 2x Intel® Optane® SSD DC P4800X (375GB)





[1] For system configuration details, please refer to Slide #5. Performance results are based on testing as of Jul 31, 2018 and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure.

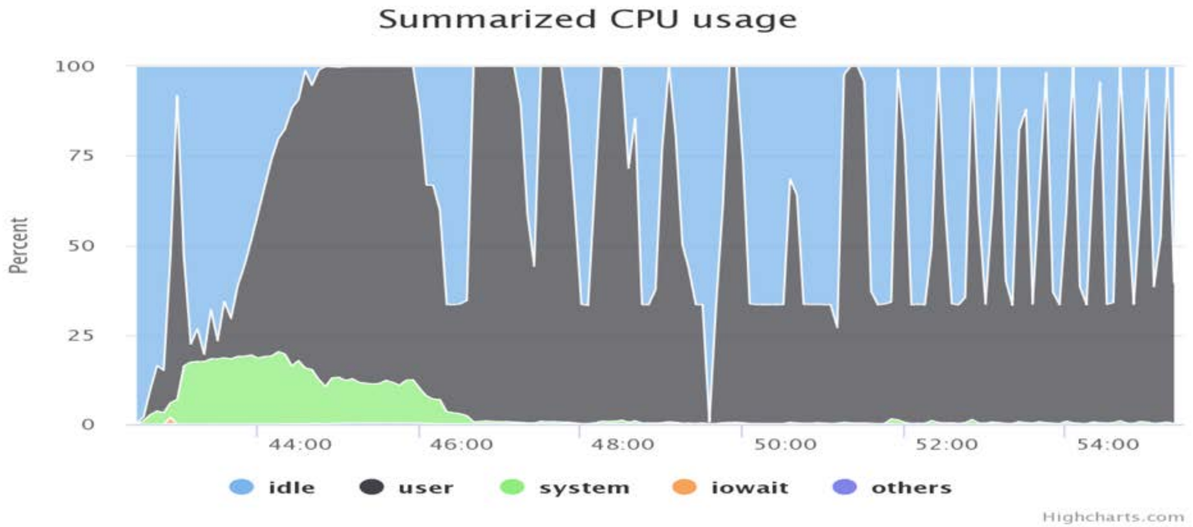
*Other names and brands may be claimed as the property of others

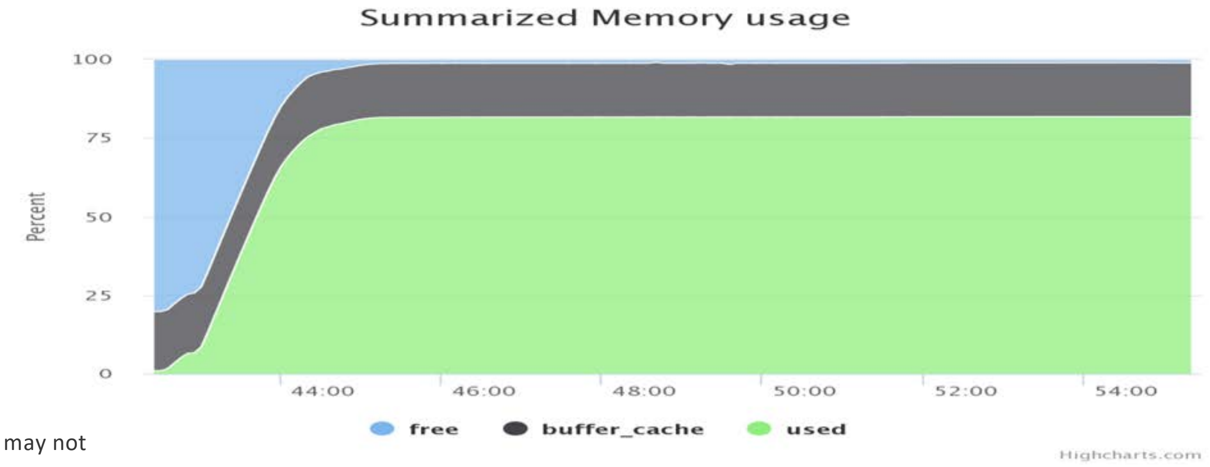
Bigger Workload using IMDT

Spark* Workload Configuration	
# of Executors across all Nodes	42
# of Cores per Executor	10
Memory per Executor	40 GiB
Memory Overhead per Executor	3 GiB
Driver Memory	1 GiB
Driver Memory Overhead	1 GiB
K-Means workload Scale Factor	2 Billion samples

Time taken to run the workload is **13** min[1]

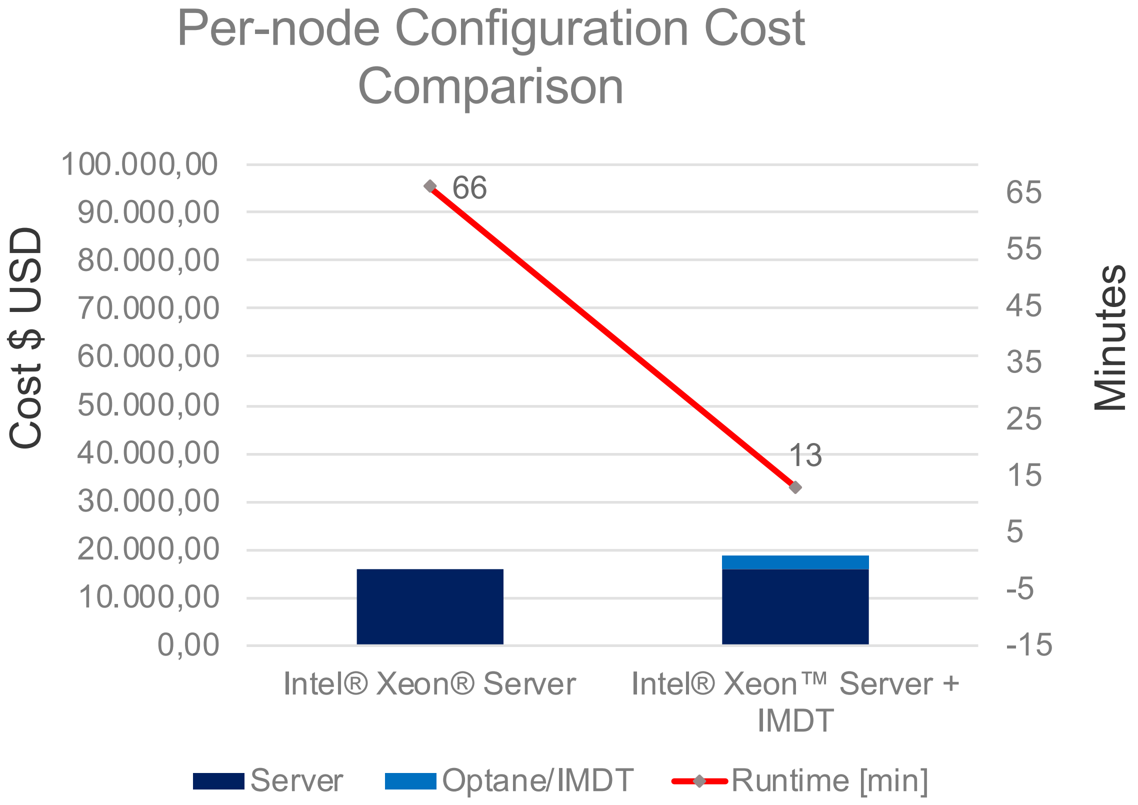
- IMDT helps to bring more memory resources without having to scale out.
- IMDT can expand memory capacity to grow x8 beyond system spec.
- That directly translates to more Spark* executors that can run in parallel.





[1] For system configuration details, please refer to Slide #5. Performance results are based on testing as of Jul 31, 2018 and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure.
*Other names and brands may be claimed as the property of others

Solution Economics



	Master Node	Data Node (x3)
CPU	Intel® Xeon® Gold 6140 CPU @ 2.30GHz	
Cores/Socket	18	
Sockets	2	
Threads per Core	2	
Total vcores	72	
Memory	192GB	
SSD	None	3.7TB Intel® SSD DC P4500 (x2)
		375GB Intel® Optane™ SSD DC P4800X (x2)
Network	10Gbps	

20% added cost[1] → reduce runtime by factor of x5.1[2]

[1] Cost estimates based on quote from Colfax International as of May 27, 2018

[2] For system configuration details, please refer to Slide #5. Performance results are based on testing as of Jul 31, 2018 and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure.

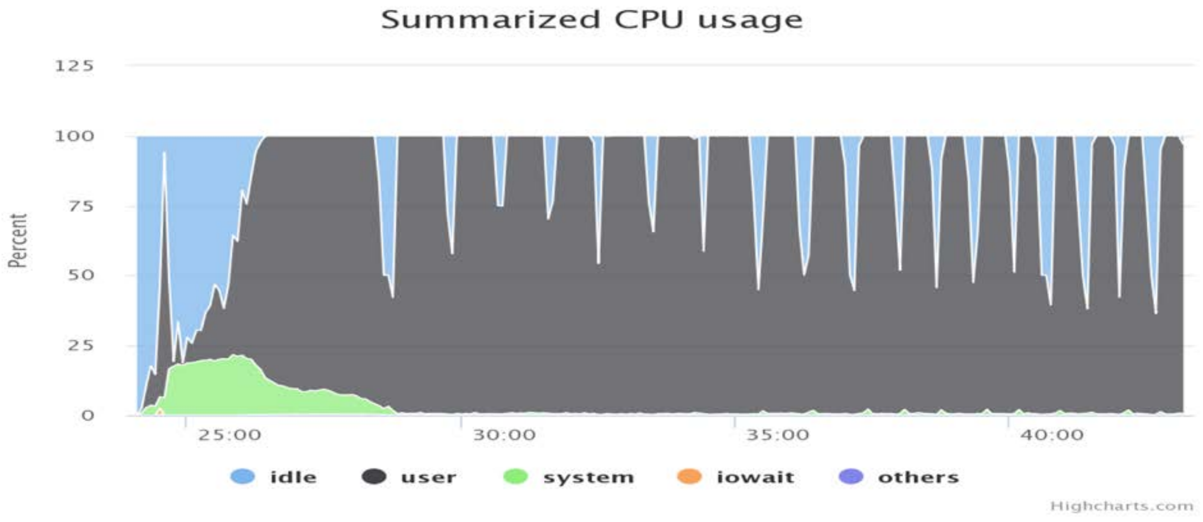
*Other names and brands may be claimed as the property of others

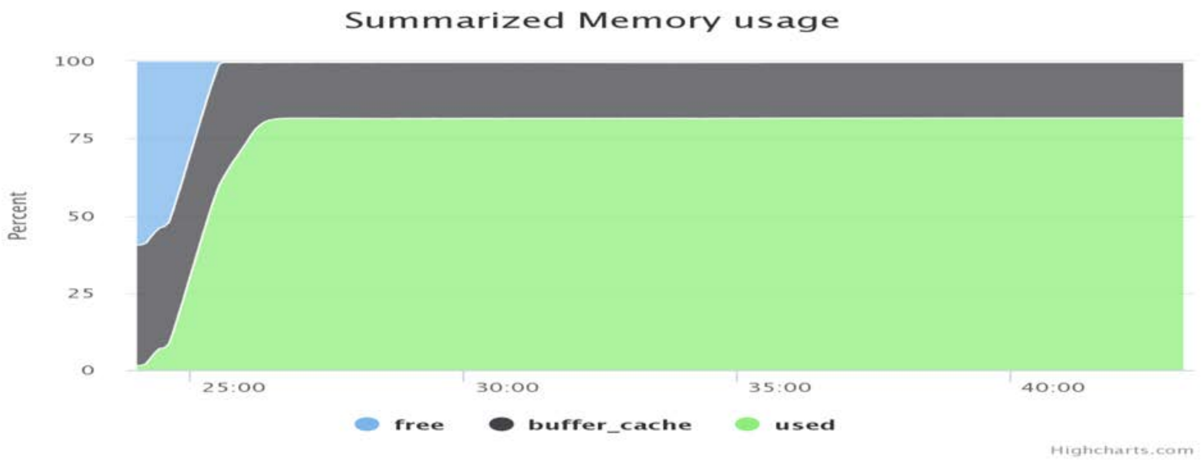
Bigger Workload using IMDT and fewer nodes

Spark* Workload Configuration (2 Data Nodes only)	
# of Executors across all Nodes	28
# of Cores per Executor	10
Memory per Executor	40 GiB
Memory Overhead per Executor	3 GiB
Driver Memory	1 GiB
Driver Memory Overhead	1 GiB
K-Means workload Scale Factor	2 Billion samples

Time taken to run the workload is **19** min[1]

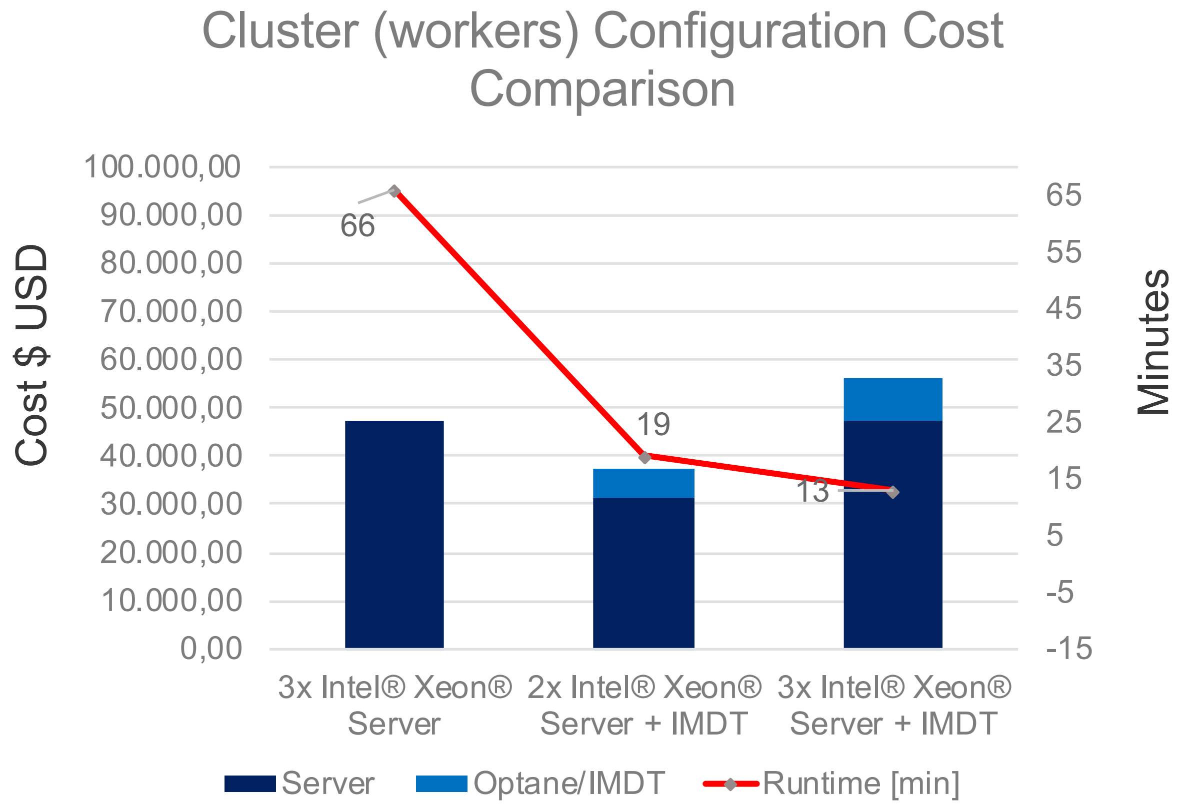
- For workloads that are not fully utilizing CPU resources in a given infrastructure, IMDT can help increase CPU utilization.
- Increasing CPU utilization allows for savings on data center footprint by reducing node-count, with larger memory per node.
- Savings can be put back into improved networks, higher-core-count CPUs, etc.





[1] For system configuration details, please refer to Slide #5. Performance results are based on testing as of Jul 31, 2018 and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure.
*Other names and brands may be claimed as the property of others

Solution Economics

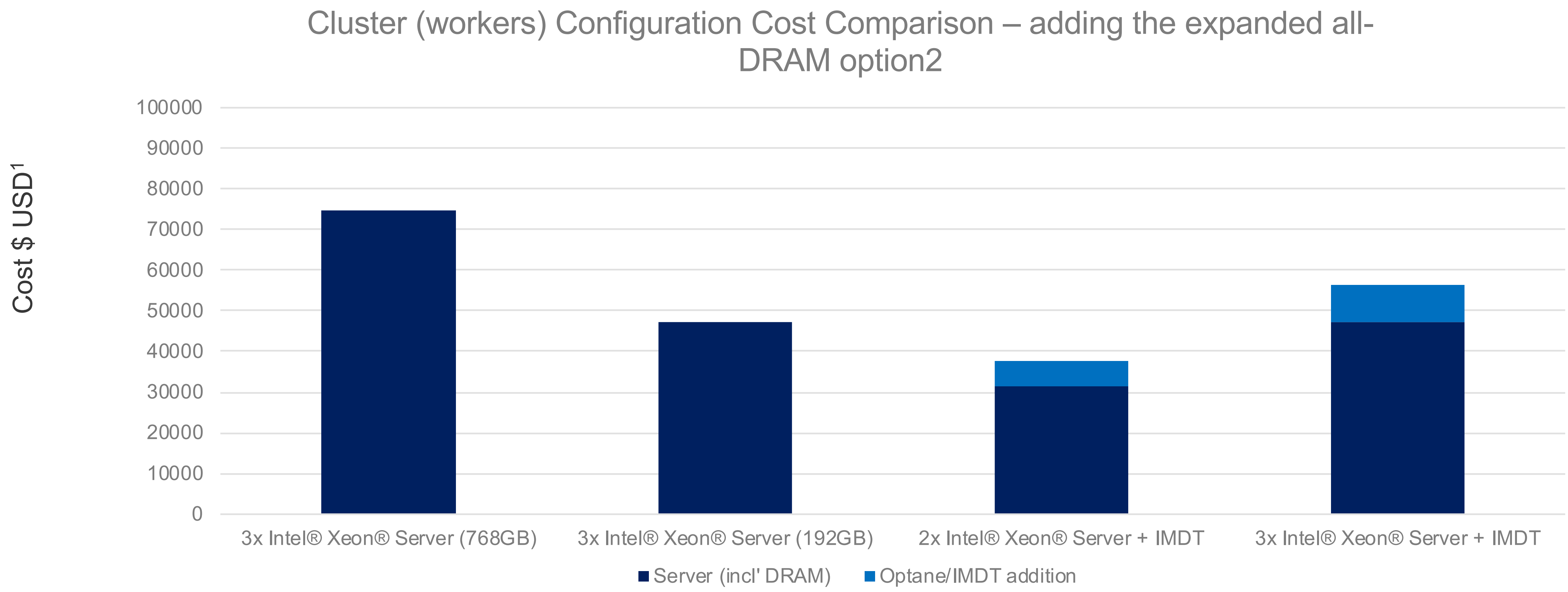


	Master Node	Data Node (x2)
CPU	Intel® Xeon® Gold 6140 CPU @ 2.30GHz	
Cores/Socket	18	
Sockets	2	
Threads per Core	2	
Total vcores	72	
Memory	192GB	
SSD	None	3.7TB Intel® SSD DC P4500 (x2)
		375GB Intel® Optane™ SSD DC P4800X (x2)
Network	10Gbps	

20% cost reduction[1] → reduce runtime by 71%[2]

[1] Cost estimates based on quote from Colfax International as of May 27, 2018
[2] For system configuration details, please refer to Slide #5. Performance results are based on testing as of Jul 31, 2018 and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure.
*Other names and brands may be claimed as the property of others

Solution Alternatives



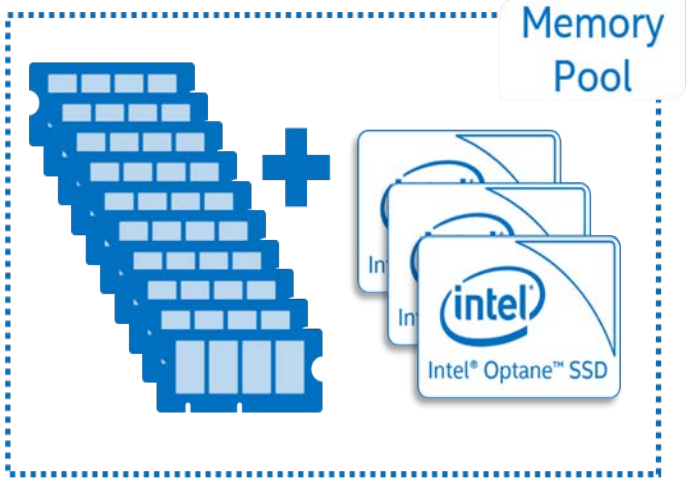
[1] Cost estimates based on quote from Colfax International as of May 27, 2018

[2] For system configuration details, please refer to Slide #5. Performance results are based on testing as of Jul 31, 2018 and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure.

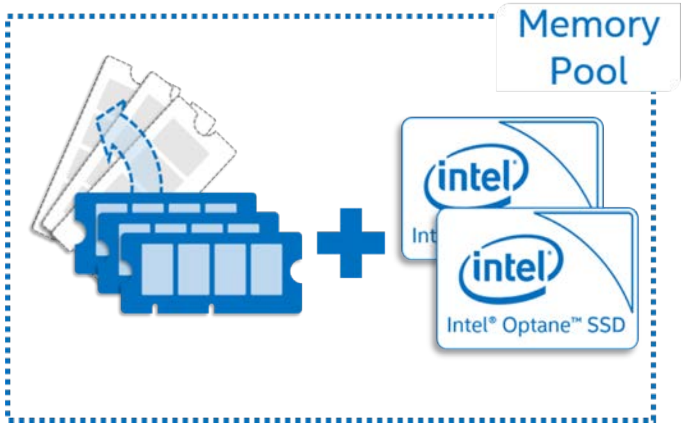
*Other names and brands may be claimed as the property of others

Summary - Optane/IMDT Benefits for Spark*

use case 1 **EXPAND beyond limited DRAM CAPACITY**



use case 2 **Displace DRAM with** Affordable SSDs

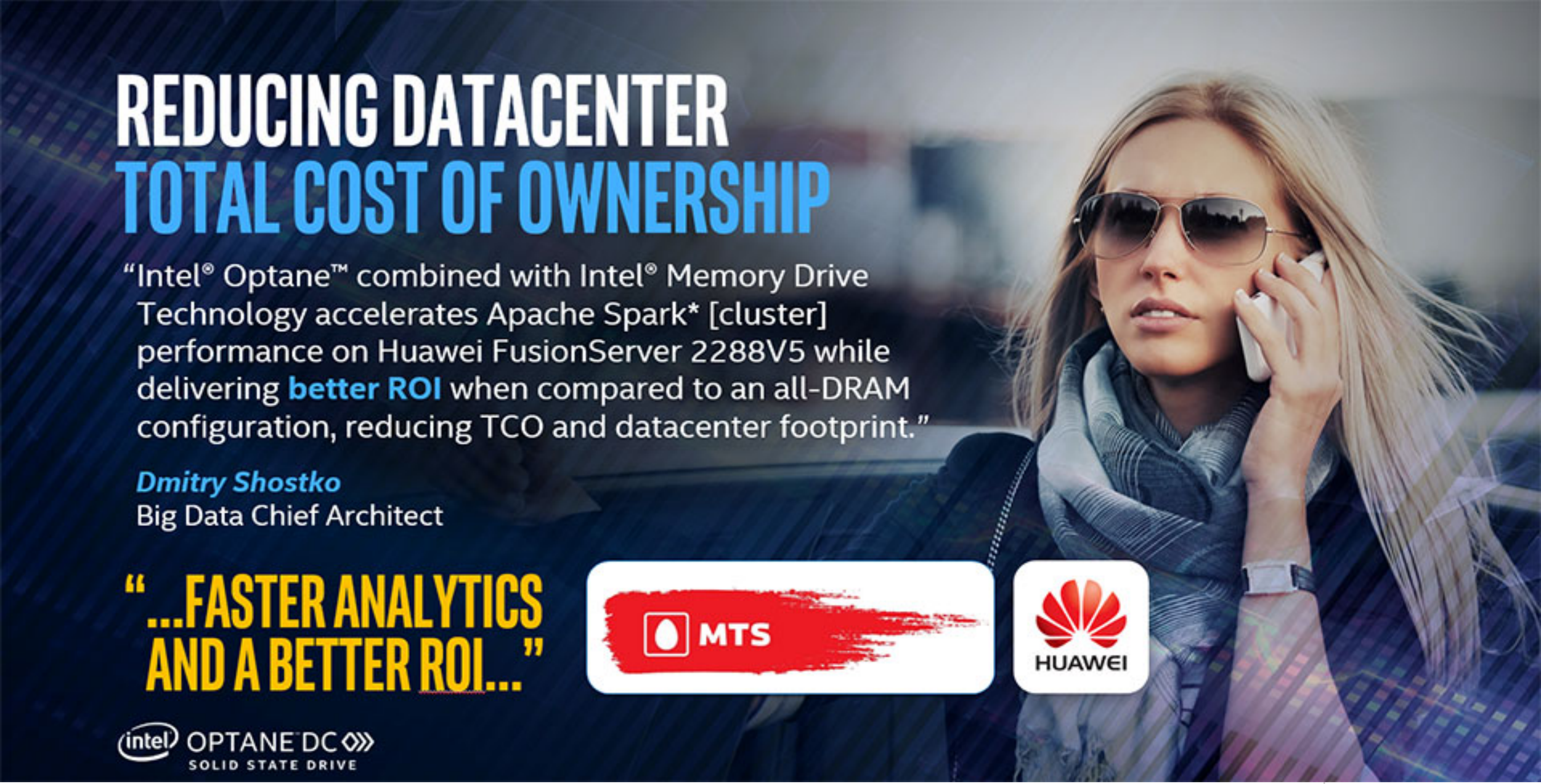


Reduce manual optimization work by having more memory available

For workloads with underutilized CPUs:

- Significantly reduce runtime
- Increase CPU utilization
- Reduce cluster node-count. Reinvest free budget in higher-core-count processors

* Other names and brands may be claimed as the property of others


REDUCING DATACENTER
TOTAL COST OF OWNERSHIP
"Intel® Optane™ combined with Intel® Memory Drive Technology accelerates Apache Spark* [cluster] performance on Huawei FusionServer 2288V5 while delivering better ROI when compared to an all-DRAM configuration, reducing TCO and datacenter footprint."
Dmitry Shostko
Big Data Chief Architect
"...FASTER ANALYTICS AND A BETTER ROI..."
MTS
HUAWEI
intel OPTANE DC
SOLID STATE DRIVE

Resources

www.intel.com/optane

www.intel.com/imdt

https://www.intel.com/content/www/us/en/software/apache-spark-optimization-technology-brief.html

QUESTIONS?