ORACLE® intel®

EXPLORING NEW SSD USAGE MODELS TO ACCELERATE CLOUD PERFORMANCE

Scott Oaks, Oracle
Sunil Raghavan, Intel
Daniel Verkamp, Intel
03-Oct-2017
3:45 p.m. - 4:30 p.m. | Moscone West - Room 3020

Big Data Talk

Exploring New SSD Usage Models to Accelerate Cloud Performance –

03-Oct-2017, 3:45 - 4:30PM,

Moscone West – Room 3020

1. 10 min – Scott

   Oracle Big Data solution
2. 15 min – Daniel

   NVMe and NVMEoF, SPDK
3. 15 min – Sunil, Case Study

   Apache Spark and TeraSort
4. 5 min - QA

Best Practices for Big Data in the Cloud - 03-Oct-2017, 4:45 - 5:30PM,

Moscone West - Room 3020

1. 10 min - Sandeep

   Oracle Big Data solution
2. 15 min – Siva

   FPGA enables new storage use cases
3. 15 min – Colin, Case Study

   Apache Spark, Big Data Analytics
4. 5 min - QA

Oracle Safe Harbor Statement

The following is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, and timing of any features or functionality described for Oracle's products remains at the sole discretion of Oracle.

Intel Notices & Disclaimers

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. Check with your system manufacturer or retailer or learn more at intel.com.

No computer system can be absolutely secure.

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. For more complete information about performance and benchmark results, visit http://www.intel.com/performance.

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 http://www.intel.com/performance.

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.

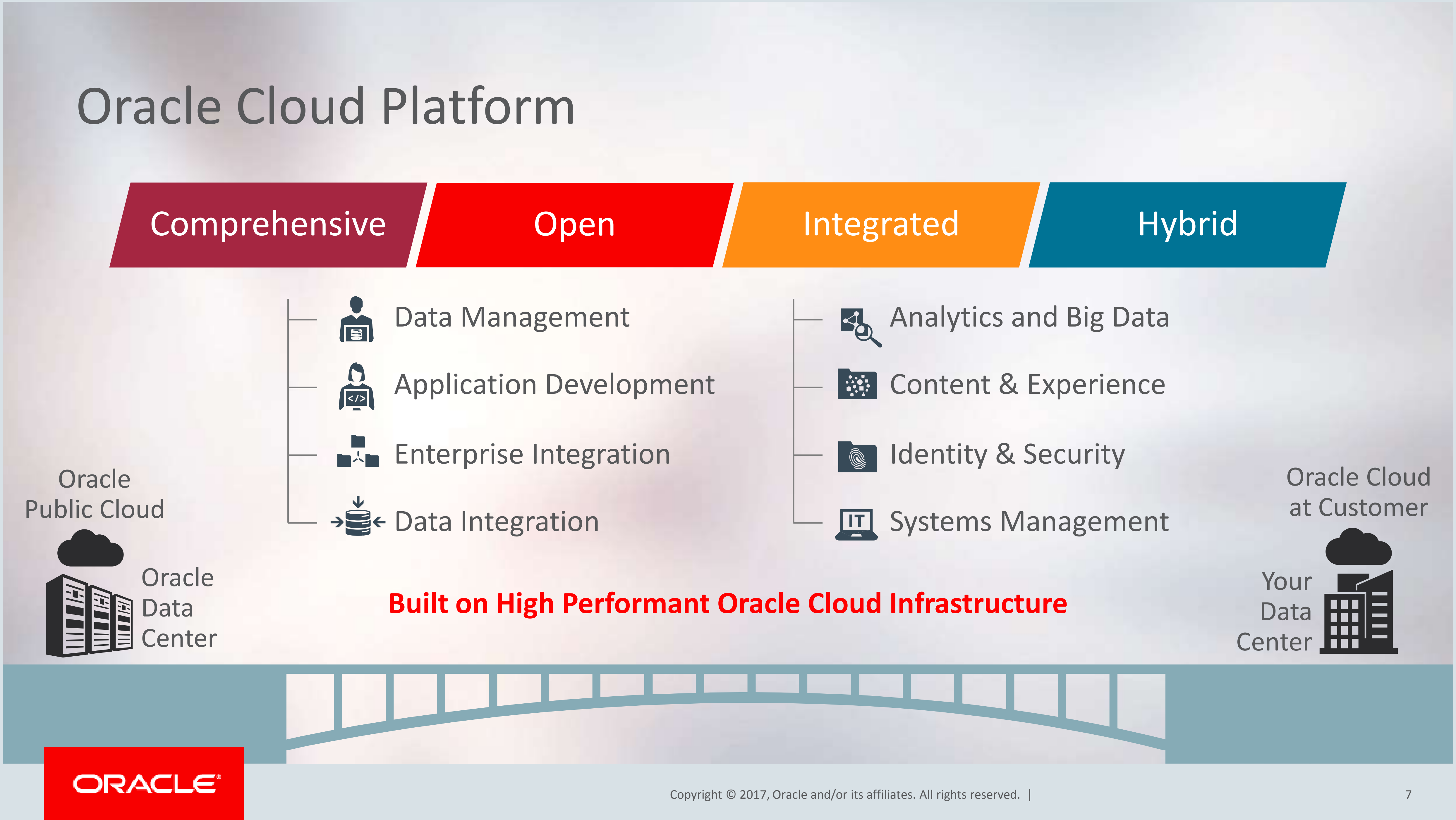
Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.

ORACLE CLOUD PLATFORM

Oracle Cloud Platform

- Develop & Deploy
- Integrate & Extend
- Publish & Engage
- Analyze & Predict
- Secure & Manage

Innovate with a Comprehensive, Open, Integrated and Hybrid Cloud Platform that is Highly Scalable, Secure and Globally Available


Oracle Cloud Platform
Comprehensive
Open
Integrated
Hybrid
Data Management
Application Development
Enterprise Integration
Data Integration
Analytics and Big Data
Content & Experience
Identity & Security
Systems Management
Oracle Public Cloud
Oracle Data Center
Oracle Cloud at Customer
Your Data Center
Built on High Performant Oracle Cloud Infrastructure
ORACLE

Oracle Cloud Platform Momentum

14,000+
Oracle
Cloud Platform
Customers

3,000+
Apps in the
Oracle Cloud
Marketplace

$1.4 Billion
FY17 Oracle Cloud
Platform
Revenue
(60% YoY Growth)

10 PaaS
Categories where
Oracle is a Leader
According to
Industry
Analysts

ORACLE

Oracle Big Data as a Service

- Work with a stack you are familiar with
- Maximum portability
- Maximum performance
  - I/O tuned for Oracle Cloud
  - OS tuned for Oracle Cloud
  - Network tuned for Oracle Cloud

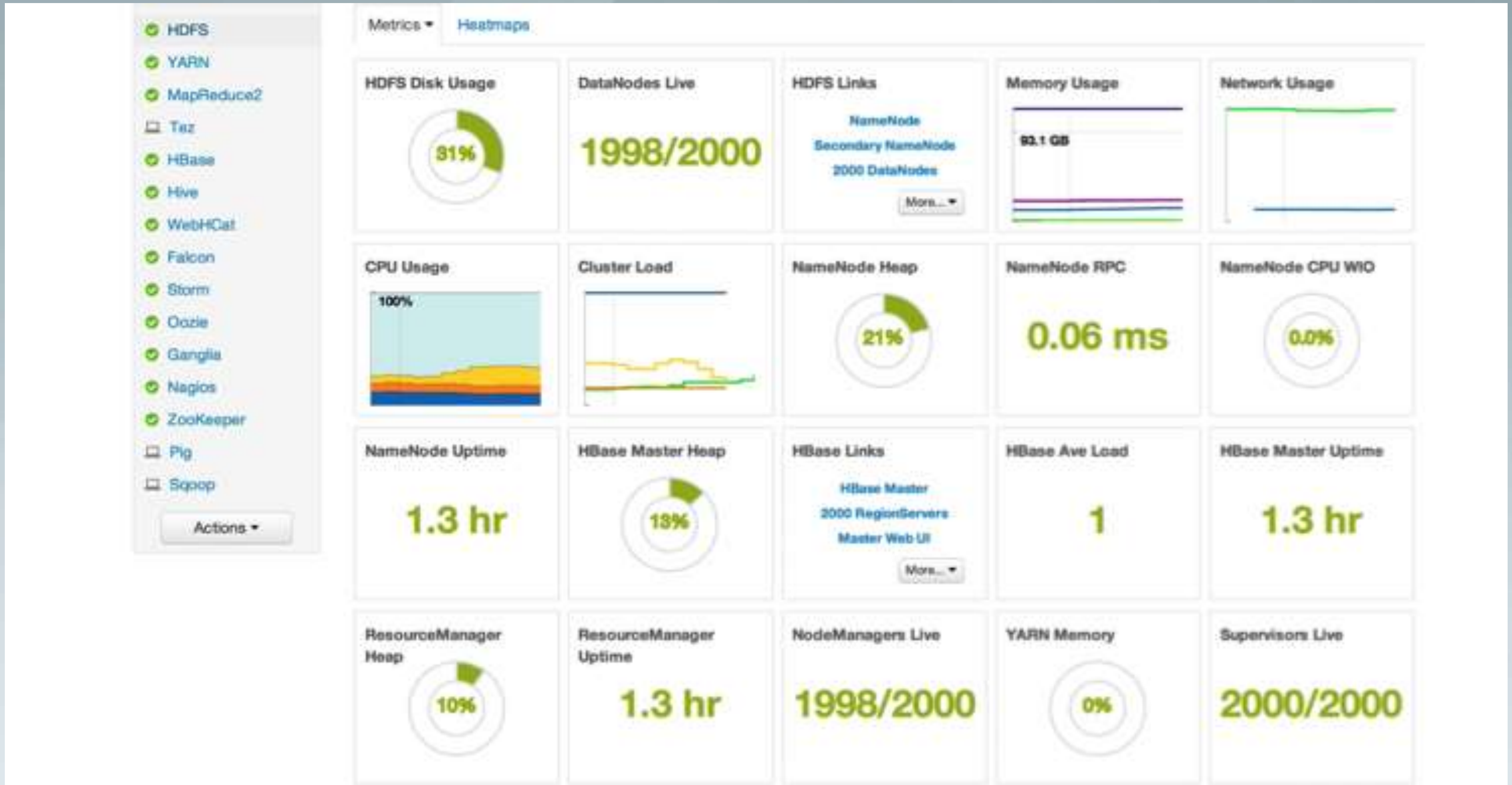
ORACLE

Oracle Cloud Platform does the grunt work

- Elastic and Scalable
  - Big Data clusters are elastic on demand
  - Storage is scaled independently
  - Choose appropriate compute shapes for your workload
- Managed
  - Automated lifecycle management
  - Service monitoring via dashboards or REST APIs

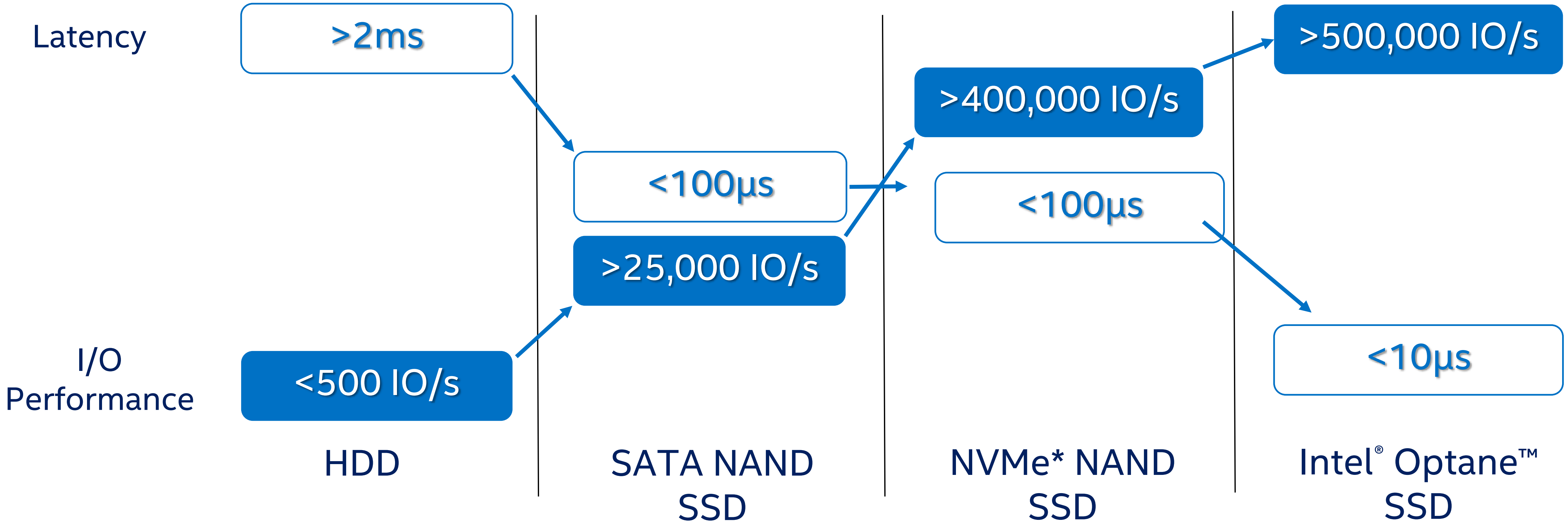
ORACLE

Big Data Cloud Service Management



ACCELERATING CLOUD STORAGE

The Problem: Software is becoming the bottleneck



The Opportunity: Intel software unlocks new media's potential

NVM Express – Key Takeaways

- Support for many independent queues with support for large queue depths
  - I/O scalability without locks for highly-parallel systems
- PCIe-attached devices with several form factors (add-in card, M.2, U.2)
  - Industry-standard bus with wide support and continued innovation
- Standard programming interface for all compliant devices
  - No vendor-specific drivers necessary
- Enterprise storage features
  - T10 DIF, dual-port controllers, reservations

NVMe over Fabrics - Key Takeaways

- Extends benefits of multi-queue NVMe protocol over the network
  - Queues are mapped to network connections
- Allows larger-scale storage systems
  - Routable network protocol allows more flexible system architecture
- Supports multiple fabrics
  - Mappings for RDMA (RoCE/iWARP) and Fibre Channel are defined today

Storage Performance Development Kit

Intel® Platform Storage Reference Software

- Optimized for Intel platform characteristics
- Open source building blocks (BSD licensed)
- Minimize average and tail latencies

Scalable and Efficient Software Ingredients

- User space, lockless, polled-mode components
- Up to millions of IOPS per core
- Designed for faster media like Intel Optane™ technology latencies

Benefits of using **SPDK**

SPDK
more performance from Intel CPUs, non-volatile media, and networking

Up to **10X MORE** IOPS/core for NVMe-oF* vs. Linux kernel

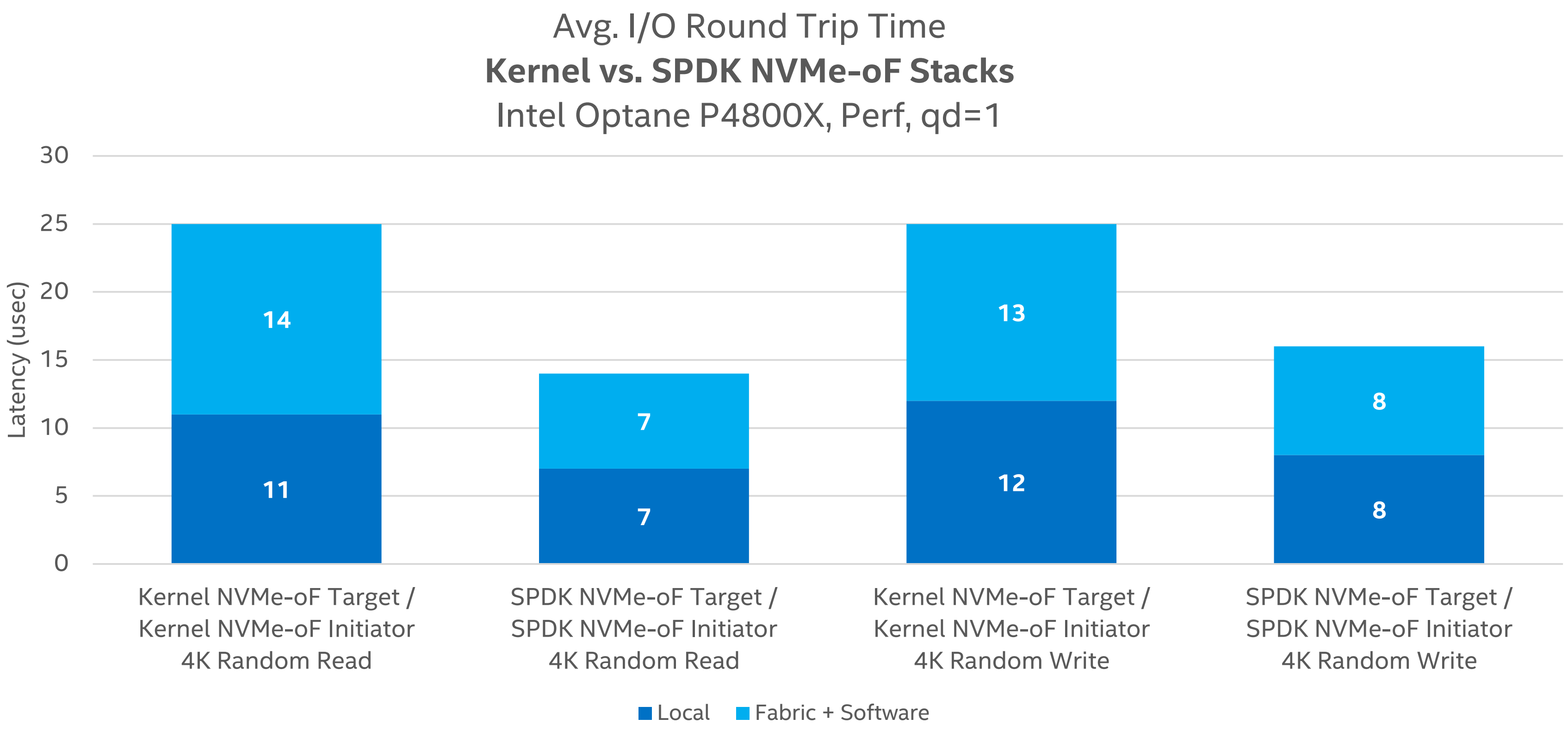
Up to **8X MORE** IOPS/core for NVMe vs. Linux kernel

Up to **50% BETTER** Tail Latency for RocksDB workloads

FASTER TTM/ **FEWER** RESOURCES than developing components from scratch

Provides Future Proofing as NVM technologies increase in performance

SPDK Host + Target vs. Kernel Host + Target



SPDK reduces Optane NVMe-oF latency by 44%, write latency by 32%!

System Configuration: 2x Intel® Xeon® E5-2695v4 (HT on, Intel® Speed Step enabled, Intel® Turbo Boost Technology enabled, 64GB DDR4 Memory, 8x 8GB DDR4 2400 MT/s, Ubuntu 16.04.1, Linux kernel 4.10.1, 1x 25GbE Mellanox 2P CX-4, CX-4 FW= 14.16.1020, mlx5_core= 3.0-1 driver, 1 ColdStream, connected to socket 0, 4KB Random Read I/0 1 initiators, each initiator connected to bx NVMe-oF subsystems using 2P 25GbE Mellanox. Performance measured by Intel using SPDK perf tool, 4KB Random Read I/O, Queue Depth: 1/NVMe-oF subsystem. numjobs 1, 300 sec runtime, direct=1, norandommap=1, FIO-2.12, SPDK commit # 42eade49

Database Use Case - SPDK

NVMe devices offer low latency and high throughput.

Databases can fully leverage these devices using SPDK.

Completely user space model for issuing database I/Os.

Asynchronous, lockless and zero copy.

Avoids kernel context switches and interrupt handling overhead.

Supports Quality of Service natively.

Ideal for OLTP databases that have extreme low latency requirements.

Database Use Case – NVMe over Fabrics

Provides shared storage capability required for Oracle RAC deployments.

Ability to scale up from one node for high performance and availability.

Ability to create clusters where each node contains local NVMe devices.

NVMe over Fabrics lets nodes access remote devices efficiently.

High availability can be achieved by replicating writes across nodes.

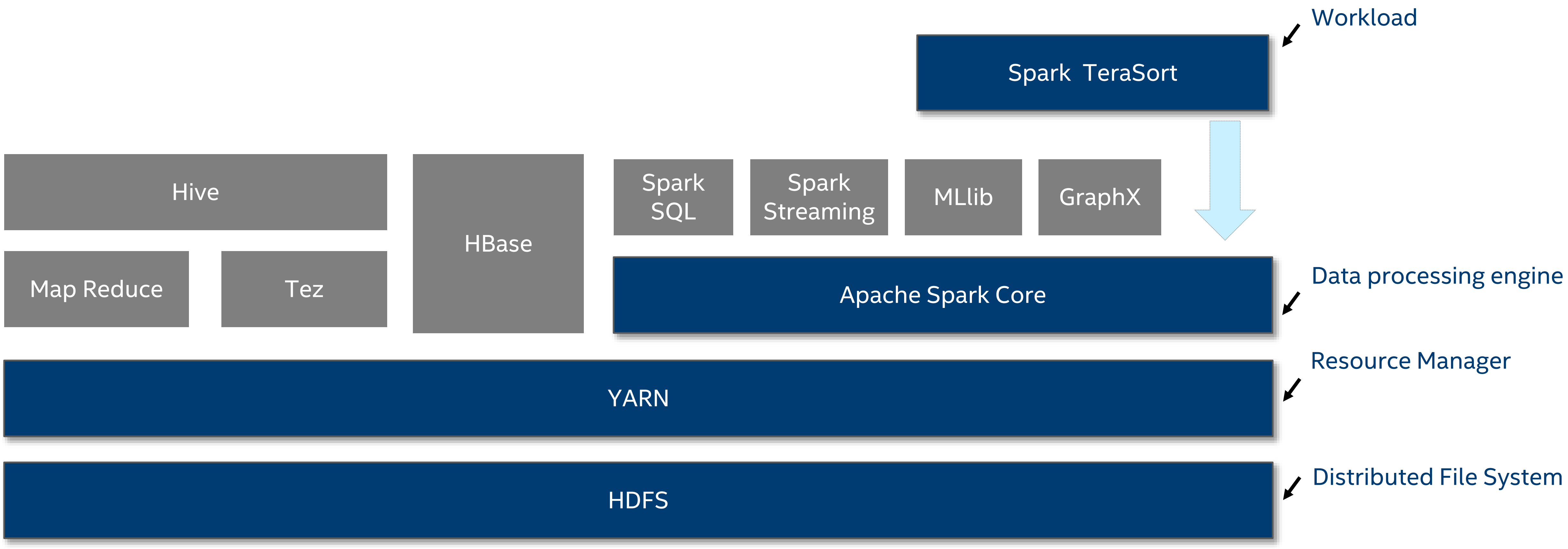
Reads can be satisfied very quickly by directly accessing local devices.

Data can also be read from a remote NVMe device using zero copy RDMA.

CASE STUDY

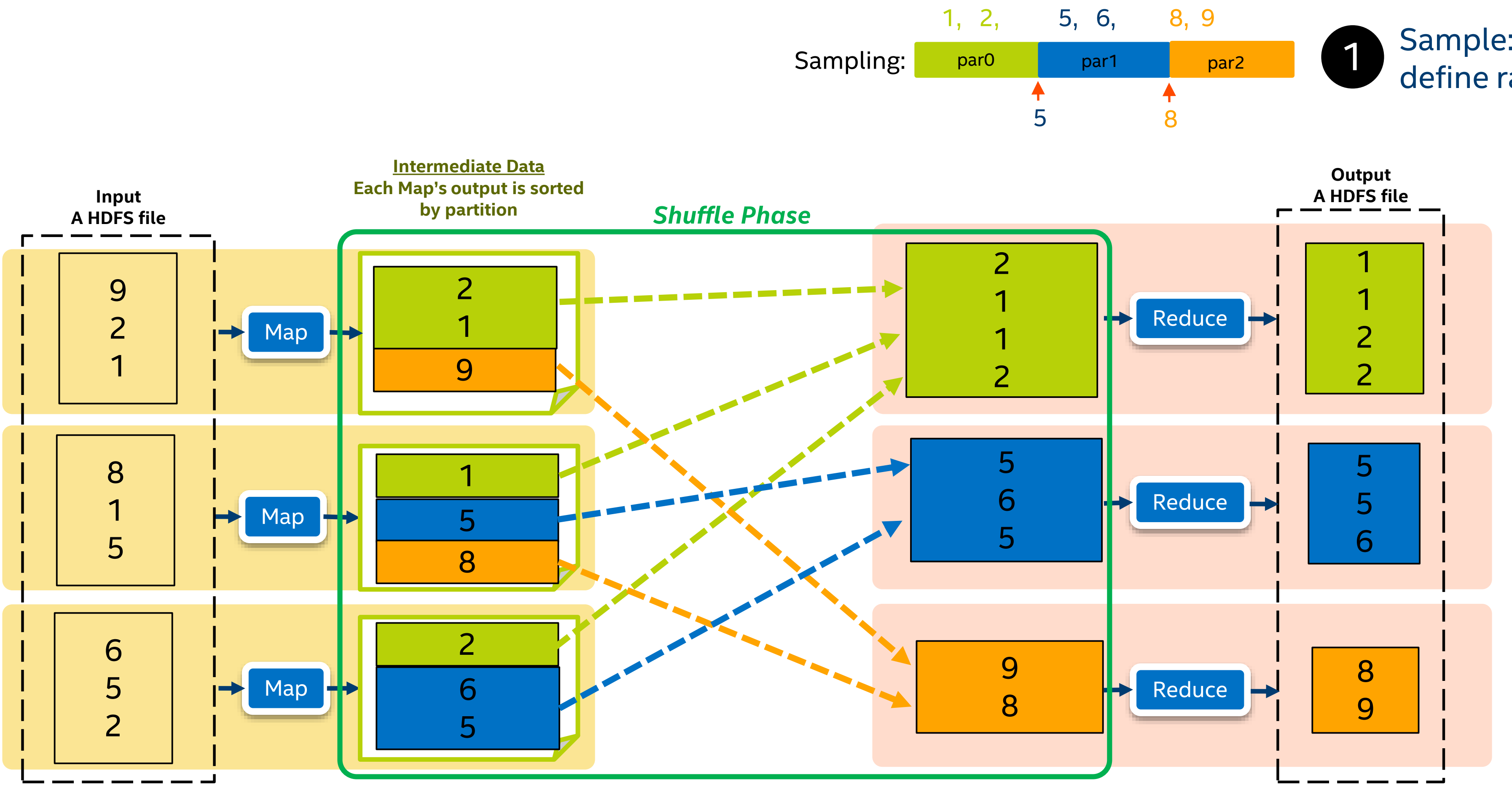
IMPROVING PERFORMANCE OF BIG DATA APPLICATIONS

Spark TeraSort



Spark TeraSort measures time to sort one terabyte of randomly distributed data using Apache Spark

How TeraSort works in Spark

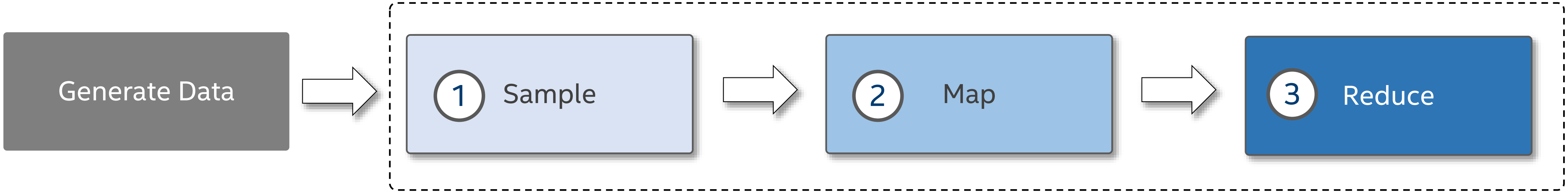


1 Sample: the data and define range for partitions

2 Map: Shuffle the data into partitions

3 Reduce: Merge-sort partitions

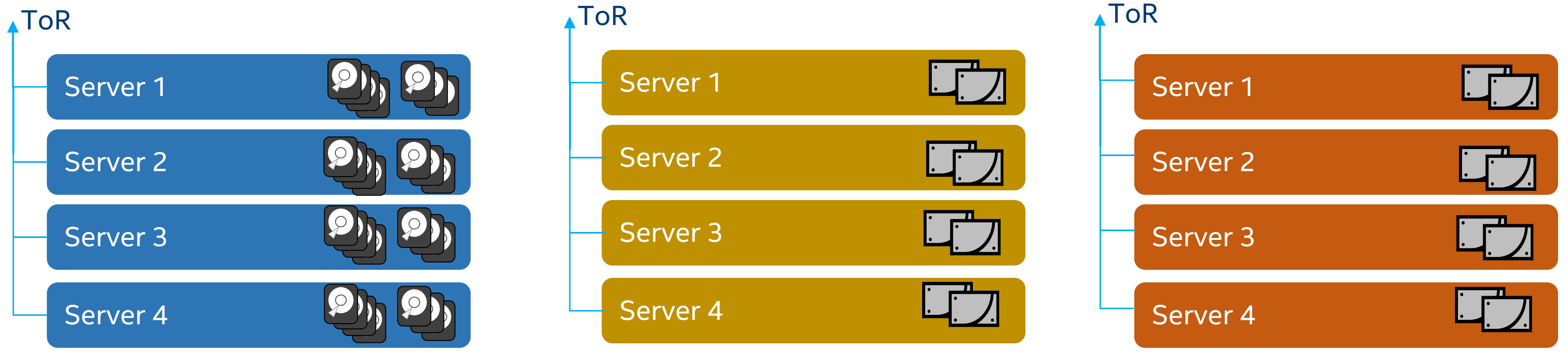
Key activities at each stage



	Stage 1	Stage 2	Stage 3
Disk	*Read HDFS (uncompressed)*	*Read HDFS (uncompressed)* *Write Spark tmp (compressed)*	*Read Spark tmp (compressed)* *Write HDFS (compressed, replicated)*
CPU	*Sample*	*Partition*	*Sort*
Network			*Merge*

TeraSort tends to be a disk intensive workload across all stages using HDFS / Spark tmp heavily.

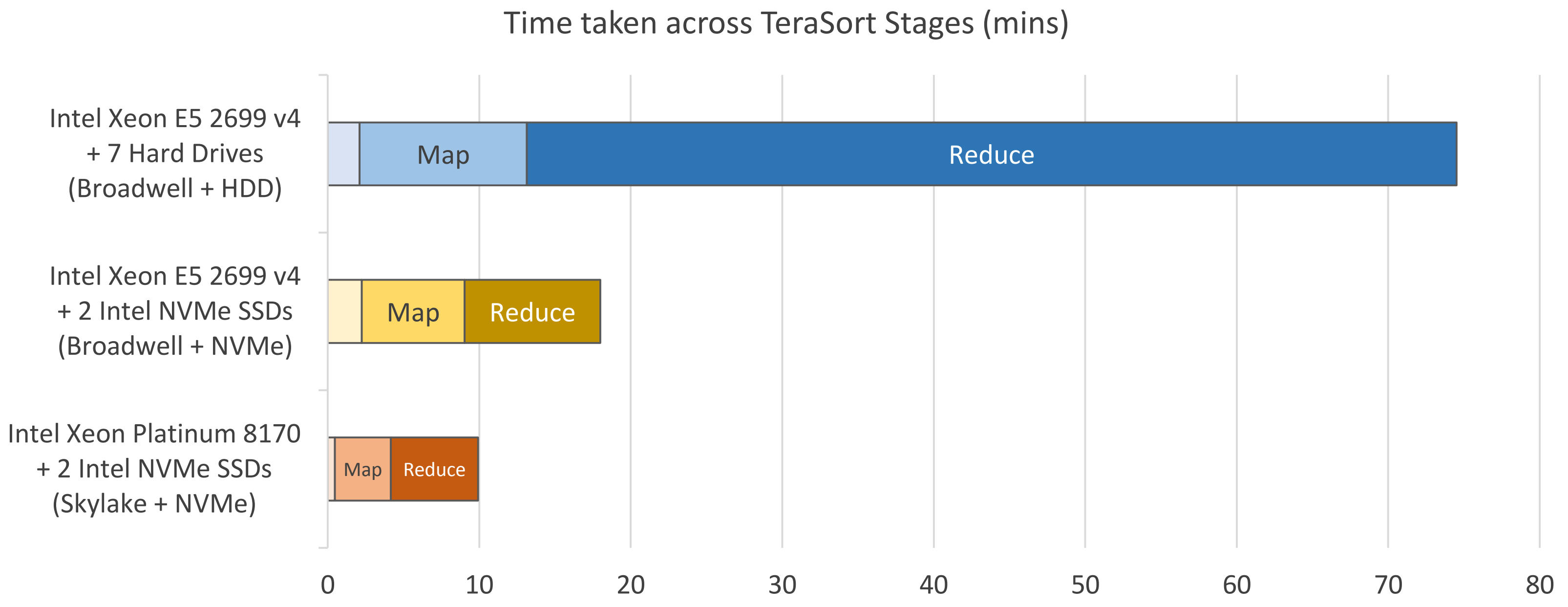
Testing TeraSort



Broadwell + HDD | **Broadwell + NVMe** | **Skylake + NVMe**

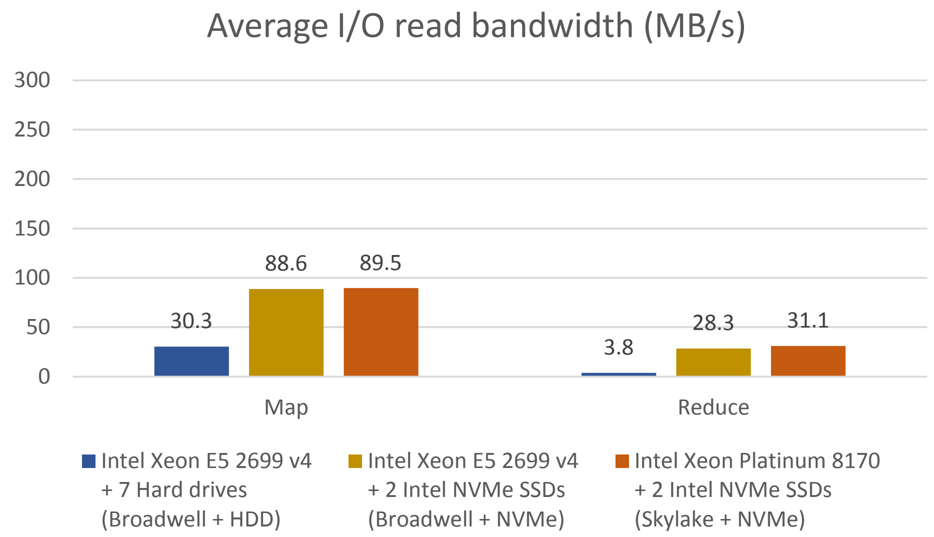
Platform	***2S Intel Xeon E5 2699 v4*** *768 GB DRAM. 10Gbps Ethernet.*	***2S Intel Xeon E5 2699 v4*** *768 GB DRAM. 10Gbps Ethernet.*	***2S Intel Xeon Platinum 8170*** *768 GB DRAM. 10Gbps Ethernet.*
Storage	***7 x Hard Drives (2TB, 7200 RPM)***	***2 x Intel NAND NVMes***	***2 x Intel 3D NAND NVMes***
OS/ Hypervisor	*Centos 7.2 / KVM*	*Centos 7.2 / KVM*	*Centos 7.2 / KVM*
Big Data SW	*Hortonworks Data Platform 2.4*	*Hortonworks Data Platform 2.4*	*Hortonworks Data Platform 2.4*
Big Data Cluster	*18 Datanode VMs (16 VCPUs 120 GB memory)* *2 Namenode VMs (4 VCPUs 30 GB memory)*	*18 Datanode VMs (16 VCPUs, 120 GB memory)* *2 Namenode VMs (4 VCPUs 30GB memory)*	*22 Datanode VMs (16 VCPUs, 120 GB)* *2 Namenode VMs (4VCPUs 30GB memory)*
Big Data Config	*HDFS replication factor=2*	*HDFS replication factor=2*	*HDFS replication factor=2*

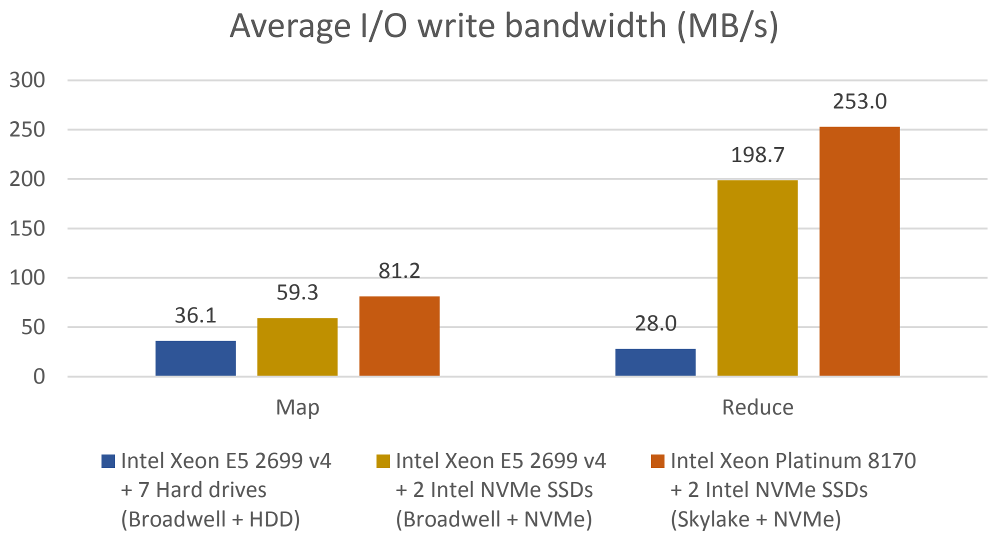
Results



- Overall, time taken TeraSort dropped ~8x from BDW+HDD to SKX+NVMe cluster
- Reduce phase runs 10x faster in 'Skylake + NVMe' compared to 'Broadwell + HDD'
  - TeraSort spends most of its time (>50%) in the Reduce phase.
  - About 80% of time is spent in reduce phase when using Hard Drives
- Performance of Map phase improves by ~3x with Skylake + NVMe cluster

Closer look at I/O performance

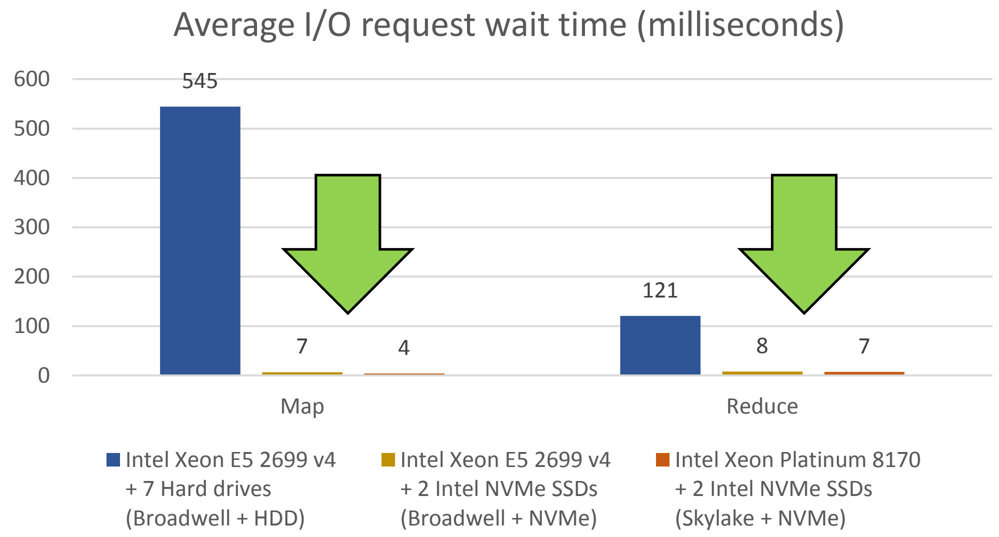


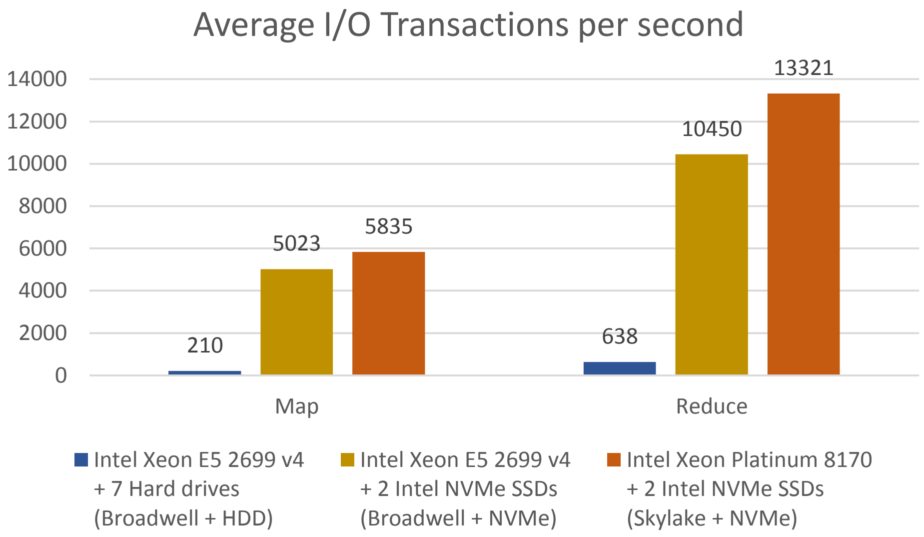


- Average I/O write bandwidth in the ‘Reduce’ phase increases 9x from Broadwell + HDD to Skylake + NVMe
  - Average I/O write bandwidth in the ‘Reduce’ phase increases 7x from Broadwell + HDD to BDW+ NVMe
- Average I/O read bandwidth in ‘Map’ phase increases ~3x from Broadwell + HDD to NVMes

Note: The I/O measurements is as measured on the disks used for both HDFS storage and Spark temp storage.

Closer look at I/O performance ..continued





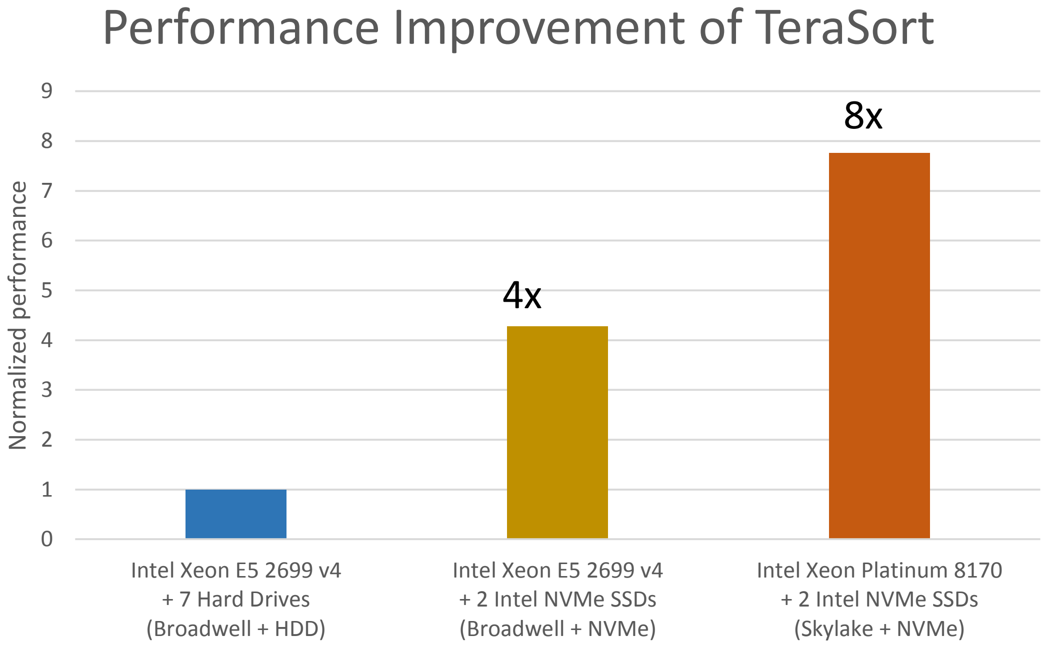
- With Intel NVMe devices the average I/O wait time for disk requests almost dropped to few milliseconds.
- Intel NVMe devices support significantly higher I/O transactions per second compared to traditional hard drives.
  - This has especially helped the performance in the Reduce phase.

Note: The I/O measurements is as measured on the disks used for both HDFS storage and Spark temp storage.

Summary

8x improvement in performance of Big Data workload

- As measure by performance of TeraSort workload
- Using 4 Intel Xeon Platinum 8170 based servers each with 2 Intel 3D NAND SSD (Intel® SSD DC P4600 - 1.6TB)
- Compared to 4 Intel Xeon E5 2699 v4 based servers each with 7 hard drives (2TB, 7200 RPM)



4x faster using Intel NAND SSDs

- Compared to cluster of same configuration but using hard drives

More Hadoop data nodes

- Bigger cluster on same number of Xeon servers, enabled by additional cores on Intel Xeon Platinum 8170 processor

BRING CLOUD PERFORMANCE TO THE NEXT LEVEL

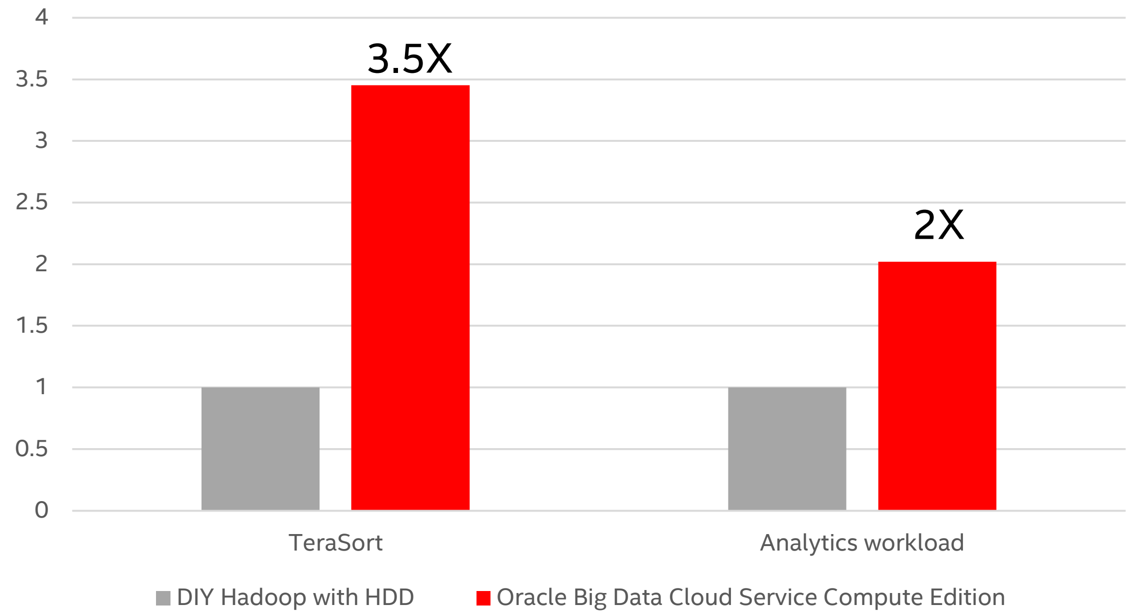
Oracle and Intel

Customers adopting both Hadoop and cloud services at a rapid rate

Oracle and Intel cloud partnership has extended to optimize Big Data

You can do it yourself but Oracle has made Big Data simple

Performance on Oracle Big Data Cloud Service

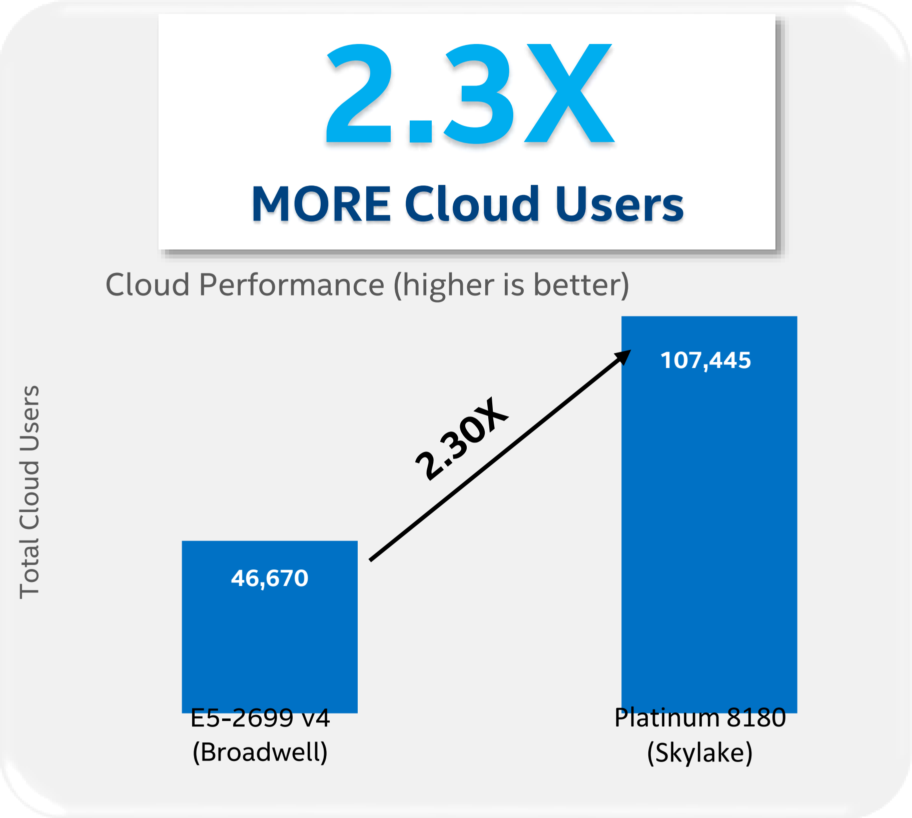


Up to 3.5X faster performance on Oracle BDCSCE over DIY Hadoop

Both clusters utilized 18 data node VMs (16VCPU and 120GB RAM) and 2 name node VMs (4VCPU and 30GB RAM), on 4 x 2S Intel Xeon E5-2699 v4 running either Oracle Big Data Cloud Services Compute Edition or DIY (Hortonworks Data Platform 2.4 on CentOS 7.2 with KVM, with 7x2.0TB 7200RPM drives per system)

Intel® Xeon® Scalable Processor optimized cache architecture

Increase cloud applications capacity, Higher VM density, Better isolation



- Intel® Xeon® Scalable processor features a new/optimized cache hierarchy
- Increased cloud performance for memory bandwidth-intensive applications
  - Larger non-inclusive Mid-Level Cache (MLC/L2) – **4X** increase (vs prior gen), enables applications to have more dedicated resources, and reduces impact of other applications on shared Last-Level Cache (LLC/L3
  - Larger total combined cache (MLC+LLC) available on Xeon® Platinum 8180 vs. Xeon® E5-2699 v4 (prior gen)

Optimized Cache Hierarchy delivers Improved Application Performance in the Cloud

ORACLE®

intel®