Big Data: at the crossroad of Cloud and HPC

Legal Disclaimers

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.

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

Relative performance is calculated by assigning a baseline value of 1.0 to one benchmark result, and then dividing the actual benchmark result for the baseline platform into each of the specific benchmark results of each of the other platforms, and assigning them a relative performance number that correlates with the performance improvements reported.

SPEC, SPECint, SPECfp, SPECrate. SPECpower, SPECjAppServer, SPECjbb, SPECjvm, SPECWeb, SPECompM, SPECompL, SPEC MPI, SPECjEnterprise* are trademarks of the Standard Performance Evaluation Corporation. See http://www.spec.org for more information. TPC-C, TPC-H, TPC-E are trademarks of the Transaction Processing Council. See http://www.tpc.org for more information.

Intel® Virtualization Technology requires a computer system with an enabled Intel® processor, BIOS, virtual machine monitor (VMM) and, for some uses, certain platform software enabled for it. Functionality, performance or other benefits will vary depending on hardware and software configurations and may require a BIOS update. Software applications may not be compatible with all operating systems. Please check with your application vendor.

Hyper-Threading Technology requires a computer system with a processor supporting HT Technology and an HT Technology-enabled chipset, BIOS and operating system. Performance will vary depending on the specific hardware and software you use. For more information including details on which processors support HT Technology, see here

Intel® Turbo Boost Technology requires a Platform with a processor with Intel Turbo Boost Technology capability. Intel Turbo Boost Technology performance varies depending on hardware, software and overall system configuration. Check with your platform manufacturer on whether your system delivers Intel Turbo Boost Technology. For more information, see http://www.intel.com/technology/turboboost

No computer system can provide absolute security under all conditions. Intel® Trusted Execution Technology (Intel® TXT) requires a computer system with Intel® Virtualization Technology, an Intel TXT-enabled processor, chipset, BIOS, Authenticated Code Modules and an Intel TXT-compatible measured launched environment (MLE). Intel TXT also requires the system to contain a TPM v1.s. For more information, visit http://www.intel.com/technology/security. In addition, Intel TXT requires that the original equipment manufacturer provides TPM functionality, which requires a TPM-supported BIOS. TPM functionality must be initialized and may not be available in all countries.

Intel ® AES-NI requires a computer system with an AES-NI enabled processor, as well as non-Intel software to execute the instructions in the correct sequence. AES-NI is available on Intel® Core™ i5-600 Desktop Processor Series, Intel® Core™ i7-600 Mobile Processor Series, and Intel® Core™ i5-500 Mobile Processor Series. For availability, consult your reseller or system manufacturer. For more information, see http://software.intel.com/en-us/articles/intel-advanced-encryption-standard-instructions-aes-ni/

Intel processor numbers are not a measure of performance. Processor numbers differentiate features within each processor series, not across different processor sequences. See http://www.intel.com/products/processor_number for details. Intel products are not intended for use in medical, life saving, life sustaining, critical control or safety systems, or in nuclear facility applications. All dates and products specified are for planning purposes only and are subject to change without notice

Parviz Peiravi

Principal Architect – Enterprise Computing

Enterprise Solution Sales Group

Big Data & Cloud Strategy, Architecture and Design

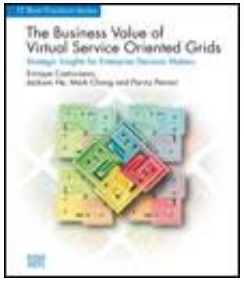
Intel Corporation

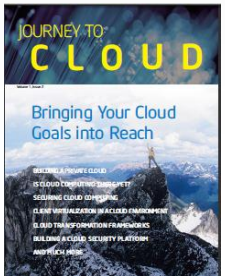
Parviz Peiravi is a Principal Architect with Intel Corporation responsible for Enterprise Infrastructure solutions and design. He is primarily responsible for designing and driving development of Big Data, Service Oriented Architecture, Cloud computing architectures in support of Intel's focus areas within Enterprise Computing. Parviz has designed large scale Clusters using Oracle, Microsoft SQL Server, and IBM DB2, Data Grid and Cloud infrastructure using Grid, SOA and virtualization technologies. He has numerous certifications in Enterprise Architecture Framework, SOA, ITIL, XML\Web Services, VMware, Xen, Hadoop and Database design. His current focus is researching the application of Big Data, Virtualization, and SOA within Cloud Computing Infrastructure Framework. He is member of Cloud Computing Group, Cloud Security Alliance (CSA), DMTF, OGF, and IEEE. Parviz has been with Intel 16+ years and holds a degree in Computer and Electrical Engineering and a recipient of Intel Achievement Award (IAA).

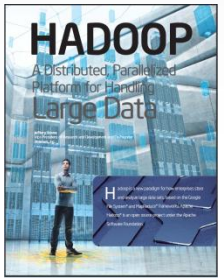
His e-mail address is parviz.peiravi@intel.com

Recent Publication: "The Business Value of Virtual Service Oriented Grids" by Enrique Castro-leon, Jackson He, Mark Chang and Parviz Peiravi, Intel Press (2008), ISBN 978-1934053102. He is also editor in chief of "Journey to Cloud" eMagazine published by Intel Corporation.

www.intel.com

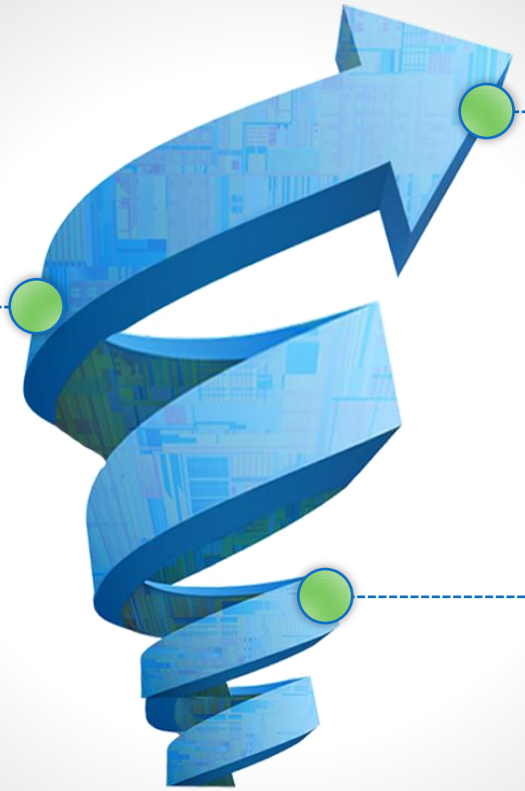






Data-Driven Innovation Fuels ‘Big Data’

40 Zettabytes of data will be generated WW in 2020[1]

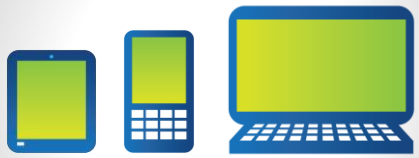


2.8 Zettabytes of data will be generated WW in 2012[1]

Cloud

Richer data to analyze

Clients



Richer user experiences

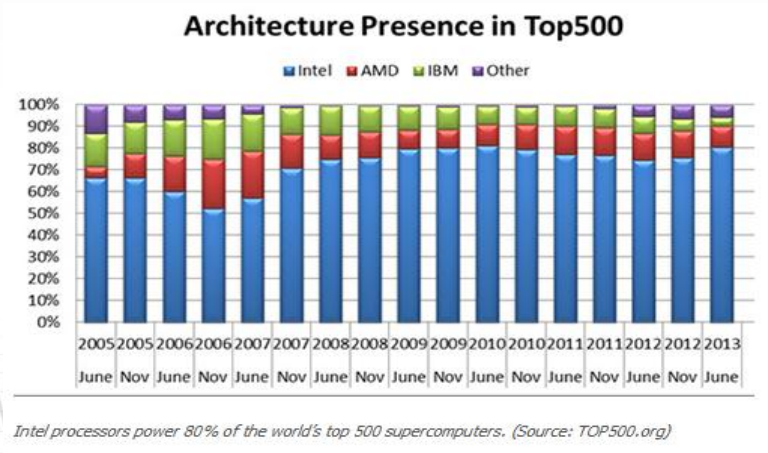
Intelligent Systems

Richer data from devices

(1) IDC Digital Universe 2020 (2) IDC

Intel Powers the World's Fastest Supercomputer

The National Supercomputing Center in Guangzhou supercomputer is called the "**Milky Way-2**" system
It is powered by a staggering 80,000 Intel processors: **32,000** 12-core future generation Intel® Xeon® E5-2600 v2 processors and **48,000** Intel Xeon® Phi™ coprocessors, for a total of **3,120,000** computing cores.



Intel processors power 80% of the world's top 500 supercomputers. (Source: TOP500.org)

Big Data

Large volume of Data, Not new to HPC

High **Velocity** and **Varity** of Data, it is a game changer!

Velocity; speed of generating and ingesting, processing, analyzing and delivery analytic services

Variety; Multi-data source,

Multi-Structured

HPC Technologies and Big Data

- Transforms Big Data from **batch to real-time** orientation
- Moves from static to **dynamic, streaming** operations requires a shift from simple to intelligent storage architecture to reap full ‘Big Data’ benefits
- Move compute to data or data to compute as appropriate

Balanced compute, storage and fabric performance is essential

Cloud Architecture Framework, Big Data and HPC services

Data Access
User
Authentication

BI & Analytic as a Service

- Streaming Analytics
- BI
- Integrated Analytics with Hadoop Support

PaaS
Application Development Environment

HPC as a Service

Data as a Services

Data ingestion, Integration and Processing Services

Enterprise Infrastructure Services

Public Cloud ⟷ Private Cloud

Data Sources

- Human Genome & Drug Discovery
- Diagnostic Images
- Surveillance and Medical Device Streaming Data
- Log Files
- Text, Video and Audio
- Social Media
- Medical Records
- GIS

Build on Secure and Open Cloud Architecture from ground up

Cloud Architecture Framework, Big Data and HPC services

Data Access
User
Authentication

Service Catalog

BI & Analytic as a Service

- Streaming Analytics
- BI
- Integrated Analytics with Hadoop Support

PaaS
Application Development Environment

HPC as a Service

NLP/Semantic Search/ Machine Learning Knowledge Management

Data as a Services

Distributed	Virtual	Persistence
Event, Message	Real-Time, Cached, Federated	EDW, Marts

Data Lifecycle Management

Master Data Management

Federated Secure Cloud Services

Data ingestion, Integration and Processing Services

- Distributed Data Processing Hadoop MapReduce
- Real-Time Streaming Data
- Data Integration Tools

Data classification and Quality

Enterprise Storage Services

- High Performance Storage
- Scale Out Storage
- Distributed File Systems (HDFS)
- NoSQL Databases
- MPP Databases DW Appliances
- DBMS Existing DW, DM

SAN, NAS, DAS | HDD, SSD, NVRAM | 10GB, Infiniband

Metadata Management

Data Tiering De-Dup & Encryption

Data Sources

- Human Genome & Drug Discovery
- Diagnostic Images
- Surveillance and Medical Device Streaming Data
- Log Files
- Text, Video and Audio
- Social Media
- Medical Records
- GIS

Security Services Privacy Compliance

Build on Secure and Open Cloud Architecture from ground up

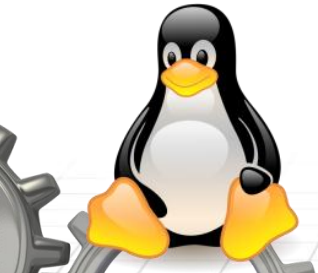
The Convergence of Advanced Technologies

HPC

Enabling exascale computing on massive data sets

Cloud

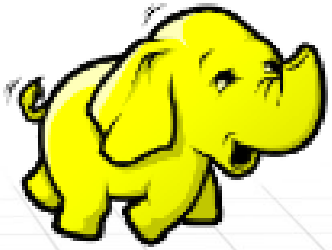
Helping enterprises build open interoperable clouds



lustre

Open Source

Contributing code and fostering ecosystem



Intel® Distribution for Apache Hadoop*

HADOOP AND HPC

Intel® Distribution for Apache Hadoop*

File-based Encryption in HDFS

Up to 20x faster decryption with AES-NI*

Role-based access control for Hadoop services

Up to 8.5X faster Hive queries using HBase co-processor

Optimized for SSD with Cache Acceleration Software

Adaptive replication in HDFS and HBase

Integrated text search with Lucene

Simplified deployment & comprehensive monitoring

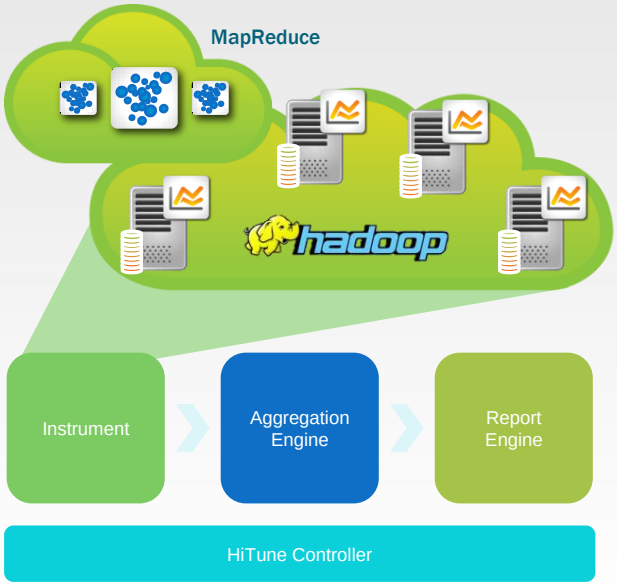
Deployment of HBase across multiple datacenters

Automated configuration with Intel® Active Tuner

Detailed profiling of Hadoop jobs

Simplified design of HBase schemas (+ in 2.4)

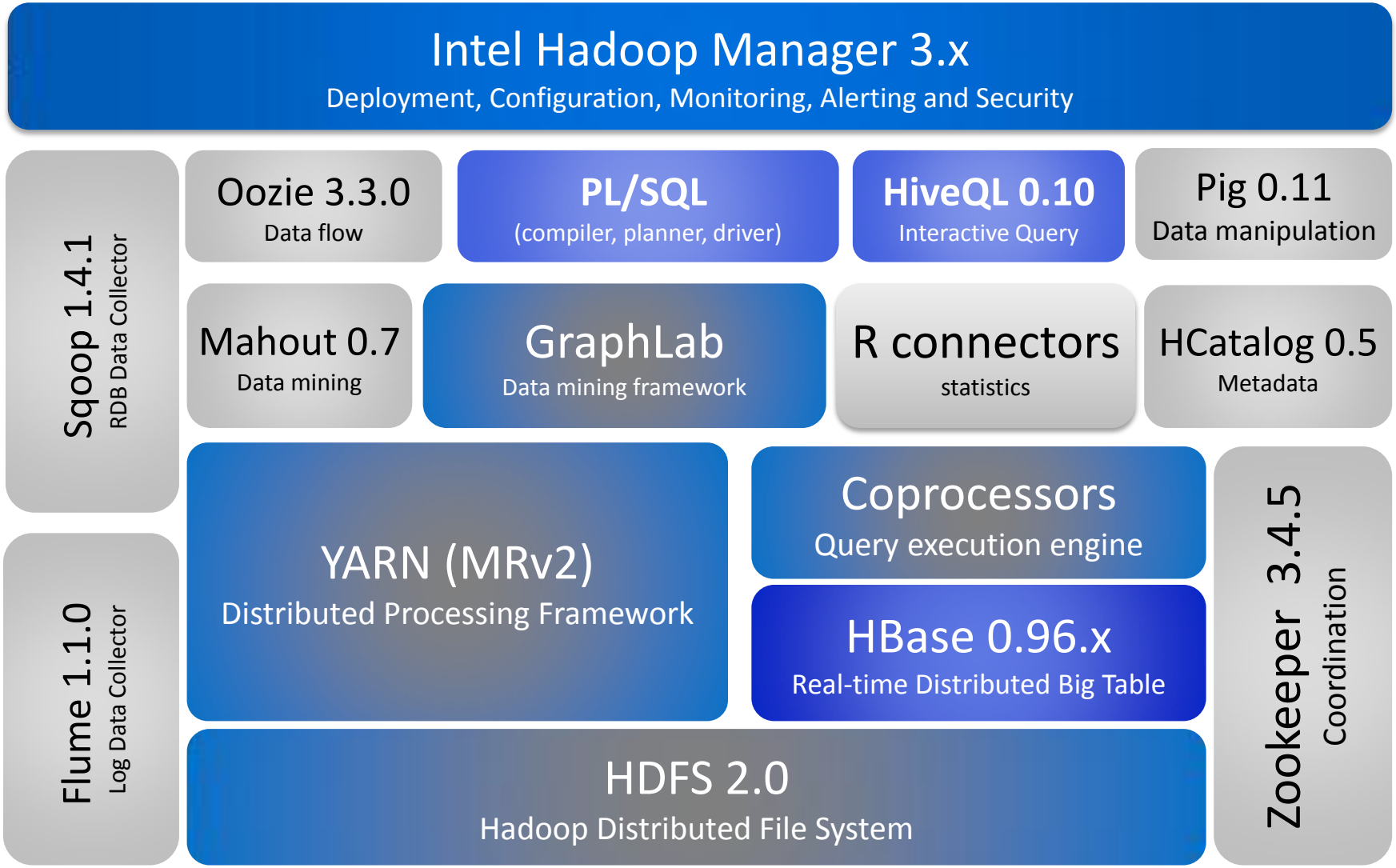
REST APIs for deployment and management (+ in 2.4)



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

Big Data

Intel's Distribution of Apache Hadoop



Lustre for Big Data & Technical Computing

Shared, parallel Lustre storage accelerates 'Big Data' workloads

Intel® Manager for Lustre*

- **Simplified, configuration monitoring and management**
- **Browser and CLI interfaces** lowers complexity and costs
- **Highly extensible** storage plug-in design and REST API

Lustre File System

- **Proven sustained performance at massive scale**
- **Flexible, scale-up and scale-out solutions**
- **Intel leads the development and delivery of new features and releases**

Intel Support Services for Lustre

- **Unrivaled Lustre support expertise**
- **Professional services offerings ensures success**

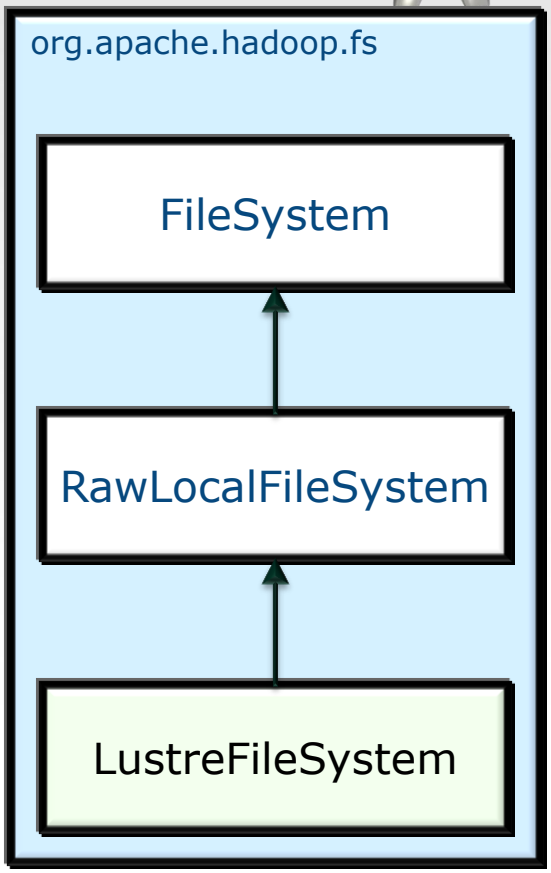
lustre

whamcloud

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

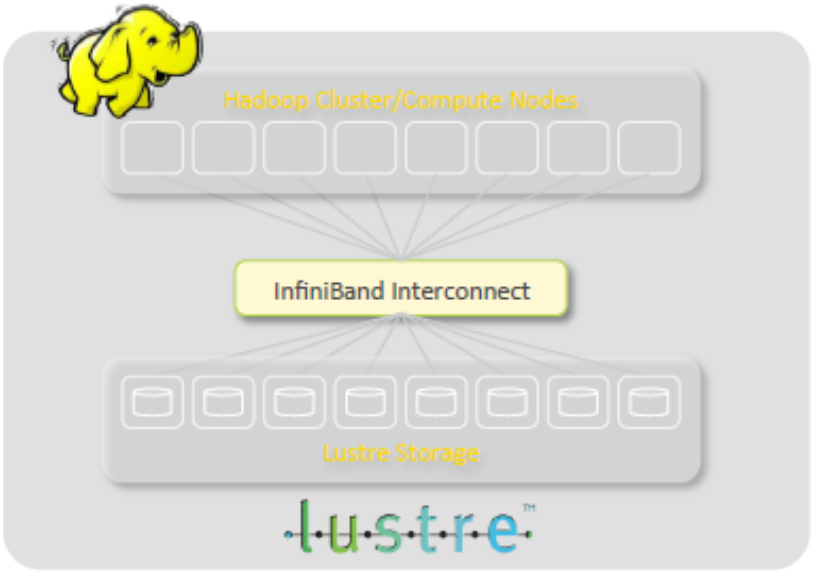
How to make them cooperate?

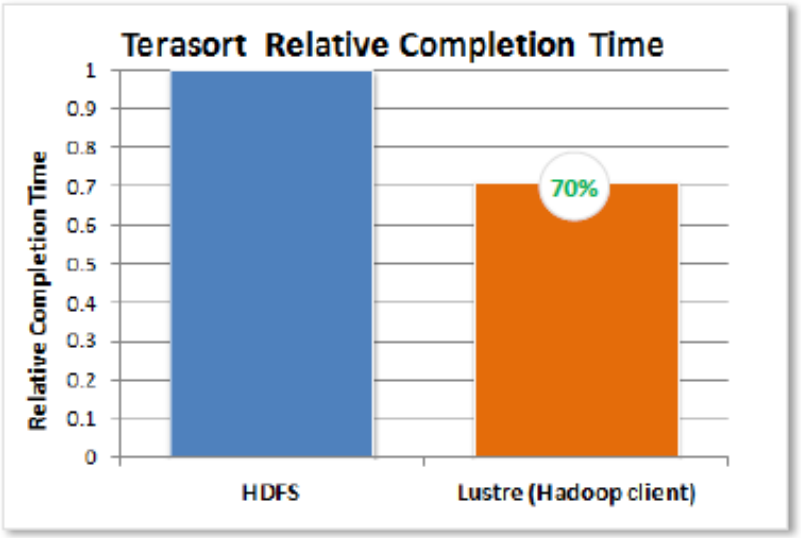
- Hadoop uses pluggable extensions to work with different file systems
- Lustre is POSIX compliant:
  - Use Hadoop's built-in LocalFileSystem class
  - Uses native file system support in Java
- Extend and override default file system behavior
- Optimizes the performance of the shuffle phase



Hadoop on Lustre

- Hadoop on Lustre: extend Hadoop analytics to HPC environments
  - *Exploit performance and scalability of shared storage*
  - *Scale storage and compute nodes separately*
- Initial work demonstrates Lustre performance advantage
- In plan to enable/optimize Hadoop analytics stack on Lustre in 2013
- Early results indicate existing HPC sites do well changing HDFS->Lustre
  - *No need to put disk back in the compute nodes*
  - *Scale compute and I/O separately to balance work*

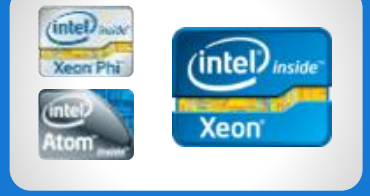
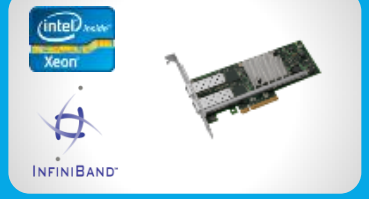
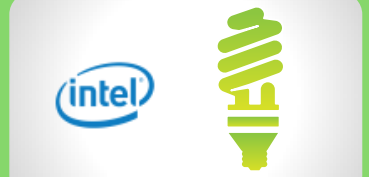




Lustre* Partner and Solution Ecosystem

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

Cloud, Big Data, HPC Building Blocks

	Compute	Network	Storage Hardware	Datacenter Software
Responsive				
Energy Efficient	Intel® Xeon® Product Family E3-E5-E7 Intel® Atom™ Intel® Xeon Phi™	Intel® Ethernet Controllers Intel® Ethernet Adapters Intel® Ethernet Switch Silicon Intel® True Scale Fabric	Intelligent Storage[1] Scale-out Storage[1] Scale-up Storage[1]	Intel® Distribution for Apache Hadoop*
Secure			Intel® SSD 710 series, DC S3700 (SATA) Intel® SSD 910 series (PCIe)	Intel and Lustre*
High Availability				
Choice	Intel's Foundational Technologies Offer Advanced Solutions for Big Data Analytics			

Xeon-based storage systems are available in a wide range of configuration options from the industry's leading storage vendors

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

Be Sure to Visit Us On-Line

Big Data

http://www.intel.com/content/www/us/en/big-data/big-data-analytics-turning-big-data-into-intelligence.html

Hadoop

https://hadoop.intel.com/

Lustre*

http://www.whamcloud.com/

Question & Respond!

Parviz.peiravi@intel.com