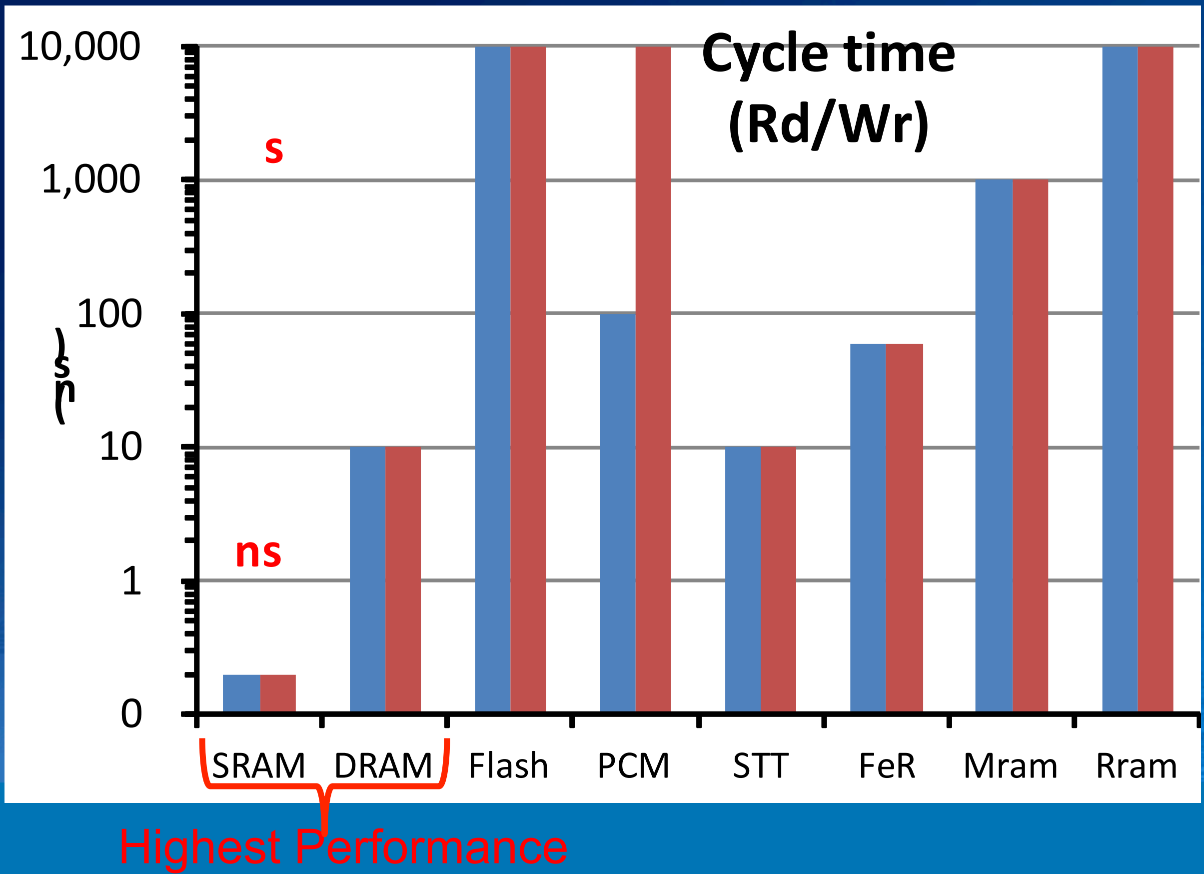
Next Generation Storage Architectures for Exascale

Mark Seager
CTO for the HPC Ecosystem
Intel Technical Computing Group

OpenSFS™
Open Scalable File Systems, Inc.

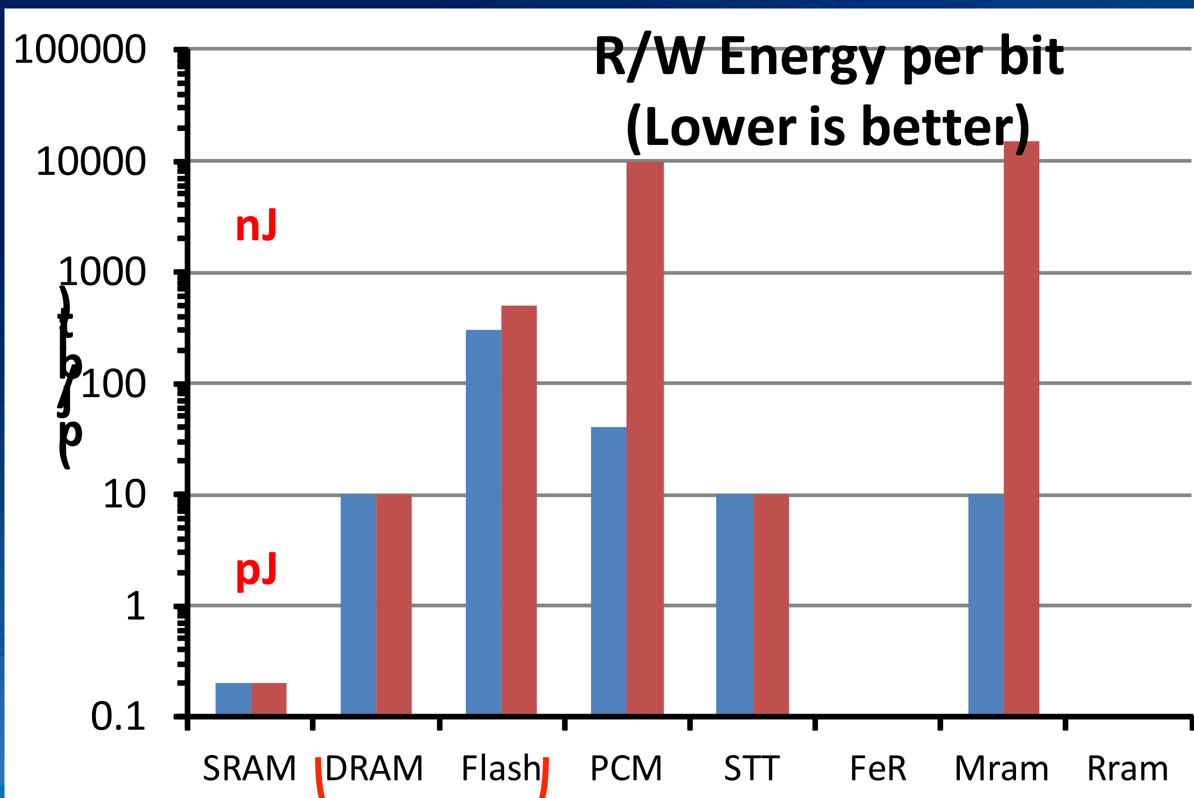
EOFS
European Open File System

Memory Technologies Latency



Memory Decoder ring
SRAM = Static RAM
DRAM = Dynamic RAM
Flash = NAND Flash
PCM = Phase Change Memory
STT = Spin Torque Transfer
FeR = Fero-electric RAM
MRAM = Magnetic RAM
RRAM = Resistance Shift RAM

Read/Write Energy



Memory Decoder ring
SRAM = Static RAM
DRAM = Dynamic RAM
Flash = NAND Flash
PCM = Phase Change Memory
STT = Spin Torque Transfer
FeR = Fero-electric RAM
MRAM = Magnetic RAM
RRAM = Resistance Shift RAM

Lowest energy given higher capacity

Memory Technology Score Card

	SRAM	DRAM	Flash	PCM	STT	FeR	Mram	Rram
Capacity	Low	High	V High	V High	Low	Low	Low	High
Performance	High	Med	Low	Low	Med	Low	Low	Low
Energy	Low	Med	High	High	Med	?	High	?
Endurance	High	High	Low	Low	High	Med	High	Med
NV	No	No	Yes	Yes	Yes	Yes	Yes	Yes
Scalable	Yes	Yes	Yes	Yes	Yes	Limited	Limited	Limited
Maturity	High	High	High	Med	Low	High	Med	Low

Small and fast
Balanced (capacity, speed, energy)

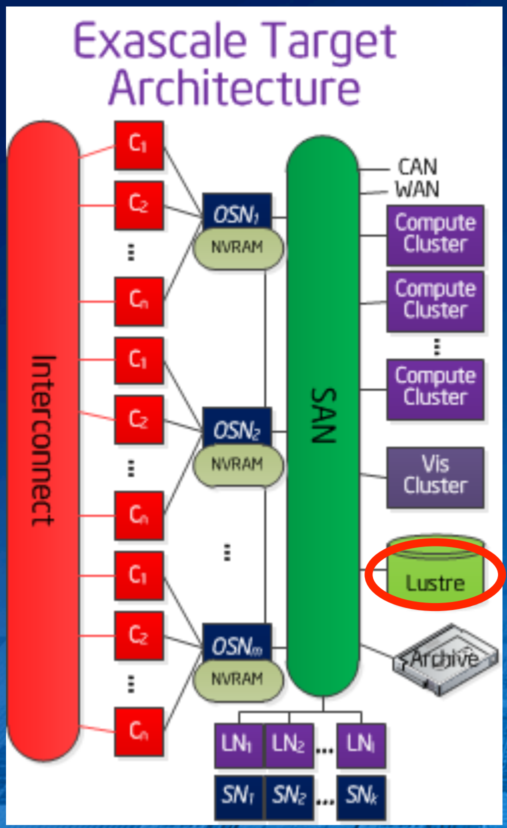
High capacity with less activity

???

New Memory Technologies will Drive a Rethink of Hierarchal Storage Management

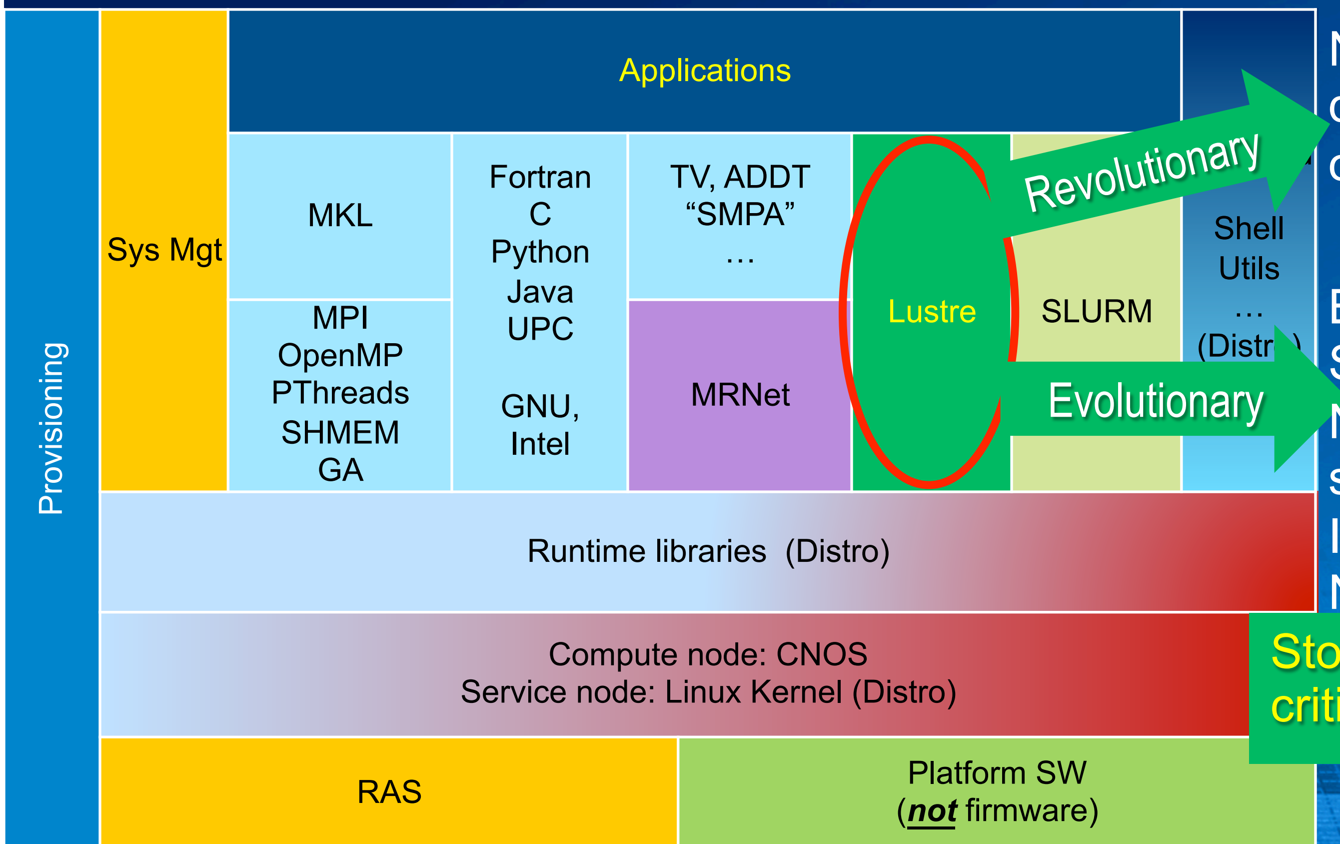
Node	Cluster	Data Center		
Processor/ Cache	Processor/ Cache	Processor/ Cache		
Mem Bus	Mem Bus	Mem Bus	PCM	10^3x BW, 10^6x IOPS
Memory	Memory	Memory		
I/O Bus	I/O Bus	I/O Bus	FLASH	10x BW, 100x IOPS
Disk	Fabric	Fabric		
	Virt BD	GW	FLASH	10x BW, 10x IOPS
		SAN		
		Virt BD	FLASH	

A new leading edge storage mechanism is required for Exascale



- Design with system focus that enables end-user applications
- Scalable hardware
  - Simple, Hierarchal
  - New storage hierarchy with NVRAM
- Scalable Software
  - Factor and solve
  - Hierarchal with function shipping
- Scalable Apps
  - Asynchronous coms and IO
  - In-situ, in-transit and post processing/visualization

HPC Software that Exascales up and also scales down for transparent user experience

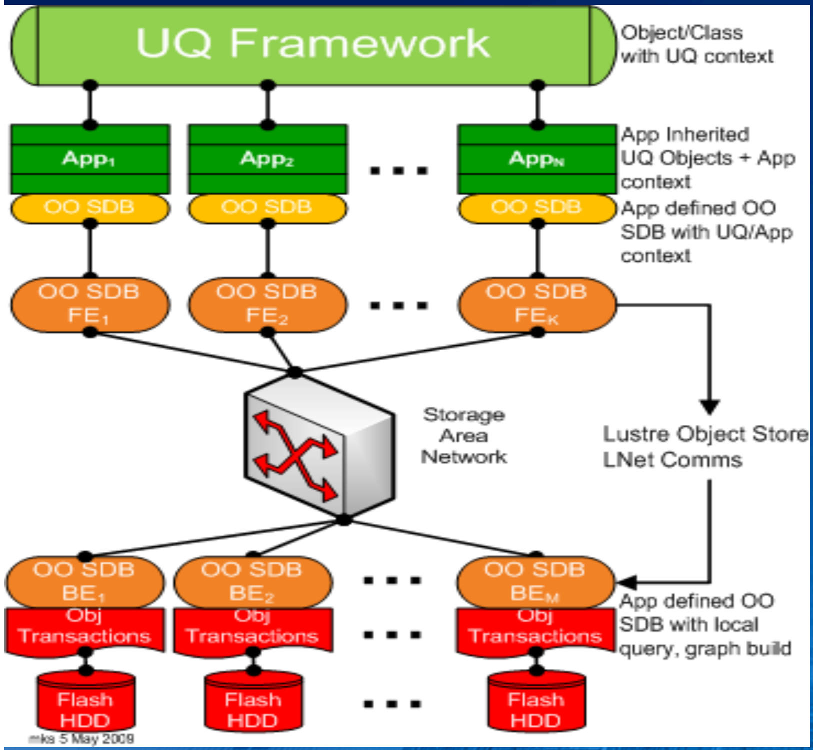


New hierarchal resilient distributed application object storage model

Burst buffer
SMP scalability
Namespace [DNE] scalability
Improved resilience
New back-end fs

Storage mechanisms are a critical system component

New approach to storage hierarchy: applications driven object oriented data storage



- UQ, Applications define objects
- Storage of objects is abstracted
- Includes remote method invocation for user computations near the data
- Access transformed from shell+ls → Python
- Metadata is accreted during object creation and IO
- Enables distributed data intensive computing model
- Enables Lustre ecosystem
- Enables analytics

The "Data Challenge"

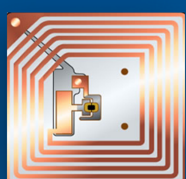
"Every two days, we create as much information as we did from the dawn of civilization up until 2003."

— Eric Schmidt, former Google CEO

1.8ZB
IN 2011
2 Days> the dawn of civilization to 2003

750M
Photos uploaded to Facebook in 2 days

966PB
Stored in U.S. manufacturing (2009)



209B
RFID tags sale in 2021; from 12 billion in 2011

1.5XB
Exabytes of traffic on Internet in 2015

200PB
Storage of a Smart City project in China

… and this is only the beginning

Big Data Graphs are Everywhere

Over 24 Petabytes
Data processed by Google every day in ***2011***

7 Exabytes
Data traffic by mobile users worldwide in ***2011***

4 billion
pieces of content shared on Facebook every day by ***July 2011***

250 Million
Tweets per day in ***Oct 2011***

5.5 million
Legitimate emails sent every second in ***2011***

158 products
ordered per second on ***Cyber Monday in 2010***

1500+ blog posts
Every minute in ***2011***

Between the birth of the world and 2003, there were 5 Exabyte of information created. We now create 5 Exabyte every 2 days
Eric Schmidt

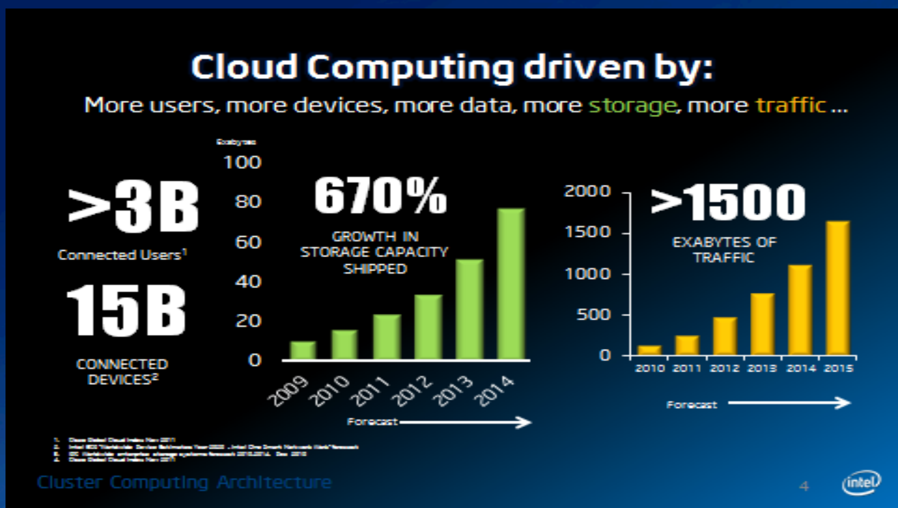
Internet devices: 1000 billion by 2013
Up from 5 billion in 2010

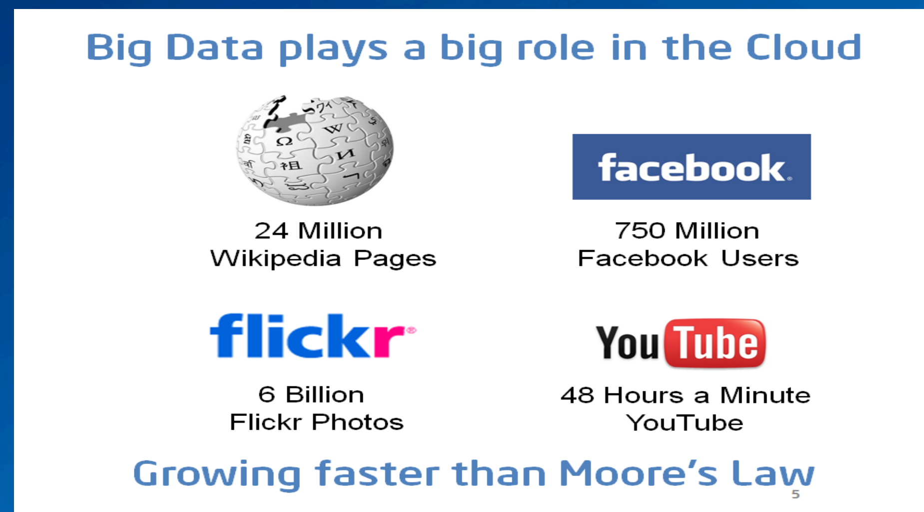
Internet traffic to increase 9x by 2013
From 5 Exabyte a month to 56 Exabyte a month in 2013

More video was uploaded to YouTube
In last 2 months, than if ABC, NBC, and CBS had been airing new content since 1948

… and graphical analysis is getting more and more sophisticated.

Grand Challenge: Knowledge Extraction





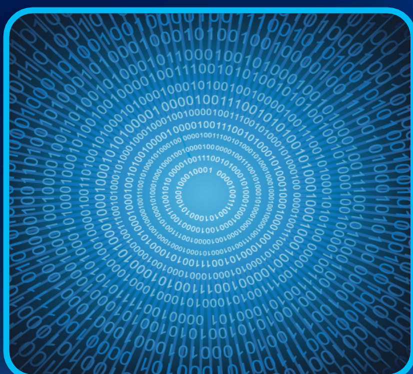
Storage and Traffic growing exponentially

… and what's vacuumed up is processed using Analytics, Machine Learning, and Data Mining methods.

Optimized Storage

ENABLES…

DATA



INFORMATION

INSIGHT

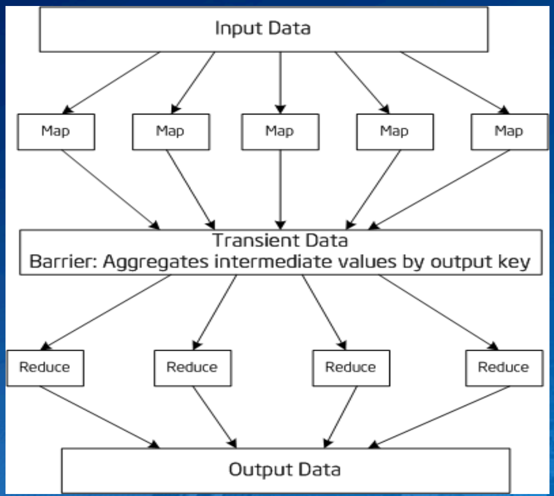
"Insight" – the Ultimate Goal

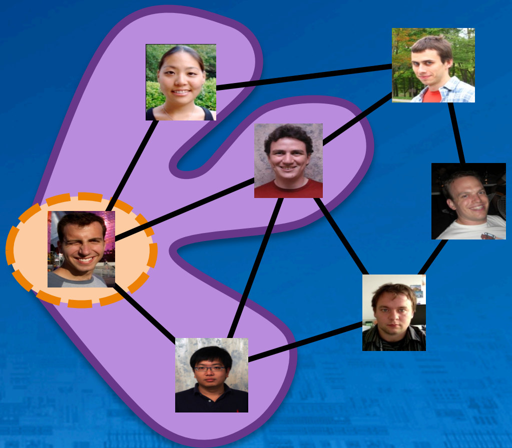
intel

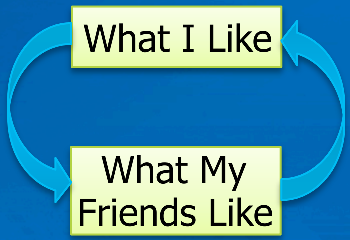
Doing this at commercial scale…

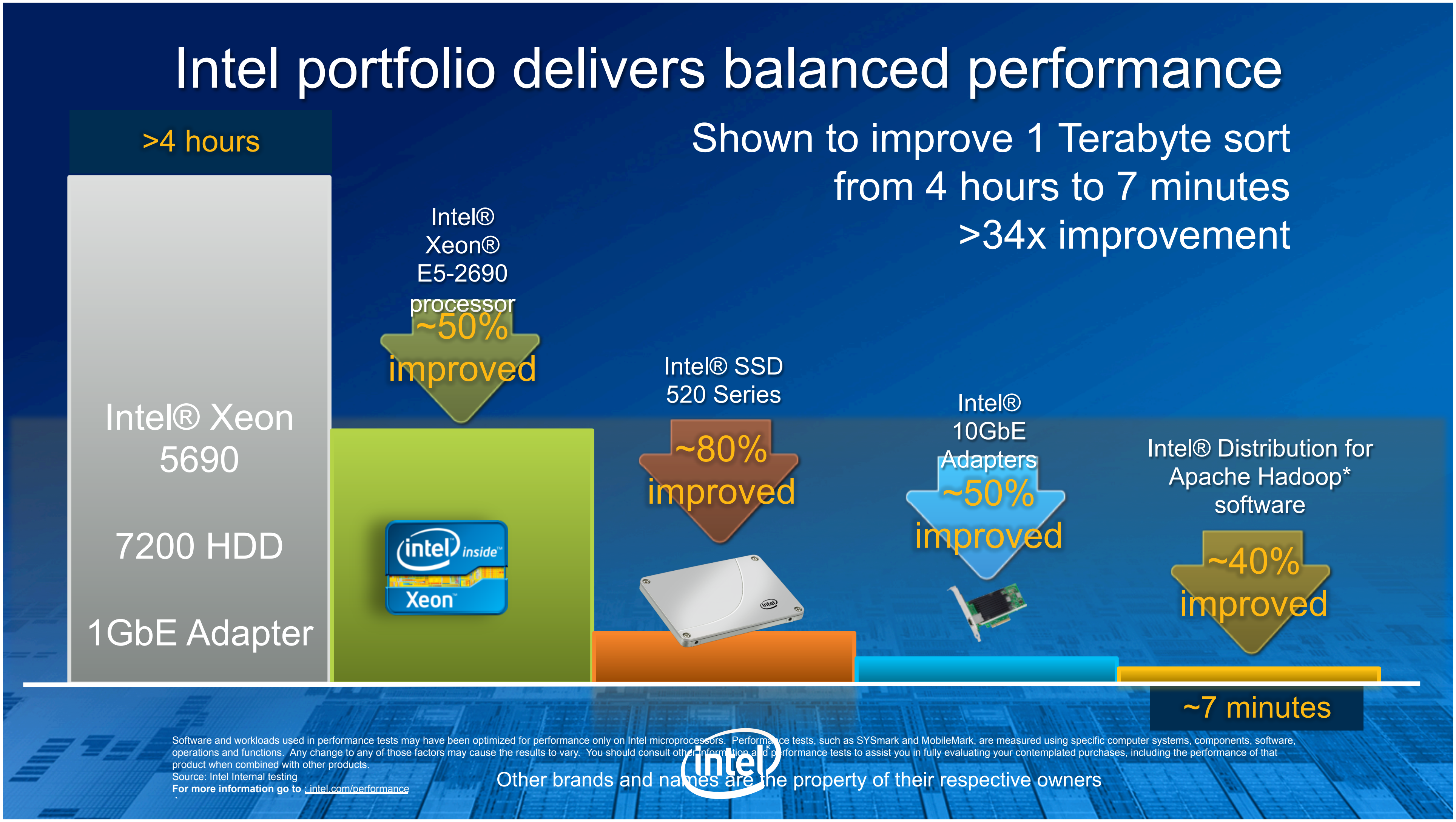
… requires some form of distributed computation.

Data-Parallel — Graph-Parallel



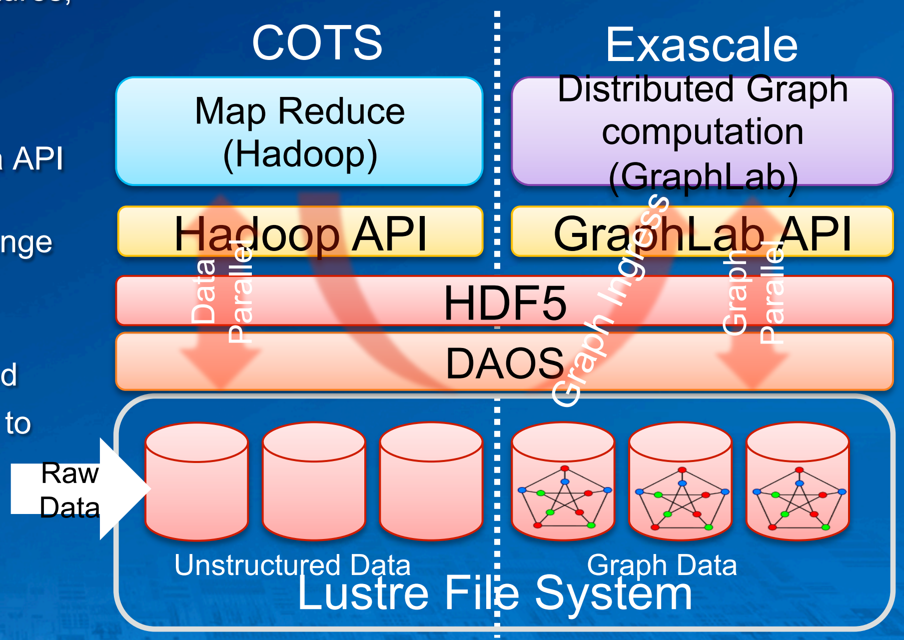





Intel portfolio delivers balanced performance
Shown to improve 1 Terabyte sort
from 4 hours to 7 minutes
>34x improvement
>4 hours
Intel® Xeon
5690
7200 HDD
1GbE Adapter
Intel®
Xeon®
E5-2690
processor
~50%
improved
intel inside
Xeon
Intel® SSD
520 Series
~80%
improved
Intel®
10GbE
Adapters
~50%
improved
Intel® Distribution for
Apache Hadoop*
software
~40%
improved
~7 minutes
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.
Source: Intel Internal testing
For more information go to : intel.com/performance
Other brands and names are the property of their respective owners

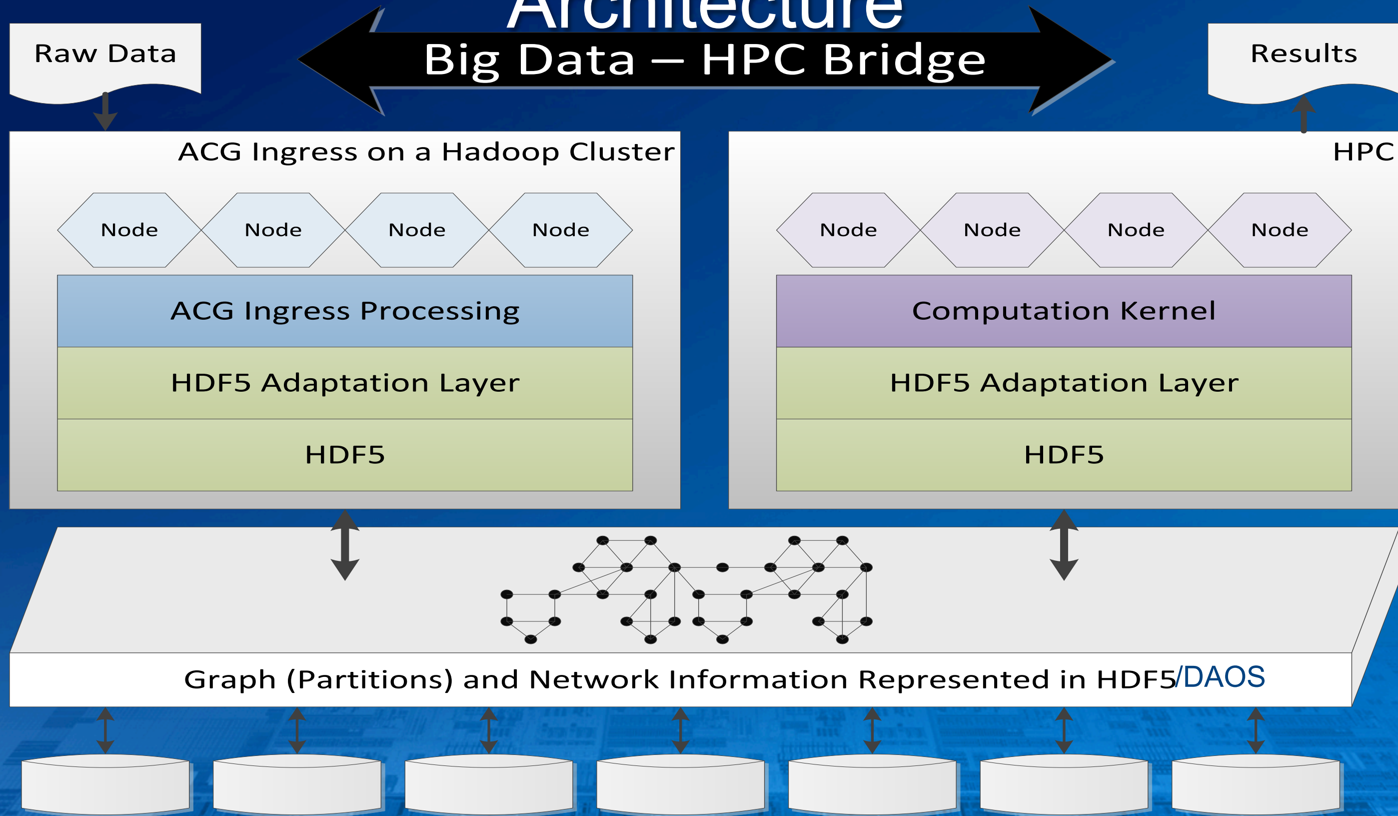
Leveraging DAOS into Big Data: Arbitrarily Connected Graph Data Analytics

- Many large-scale machine learning problems involve graph structures, and **Hadoop** is ideal for constructing graphs for Exascale computations:
  - Graph relationships built from unstructured data
  - Objects/relationships stored to DAOS via self-describing data API (HDF5) and then loaded by Exascale
- **GraphLab's** asynchronous execution model is ideal for a wide range of machine learning computations
  - Each node processes a portion of graph
  - Objects loaded from HDF5/DAOS during execution as needed
- After computation, DAOS may be used by various cloud services to query selected object values
- Intel Lab's prototyping effort:
  - Port Hadoop and GraphLab to the new DAOS interface
  - Evaluate functionality on COTS systems
  - Evaluate ingress and execution performance on Exascale prototype using large-scale machine learning benchmarks



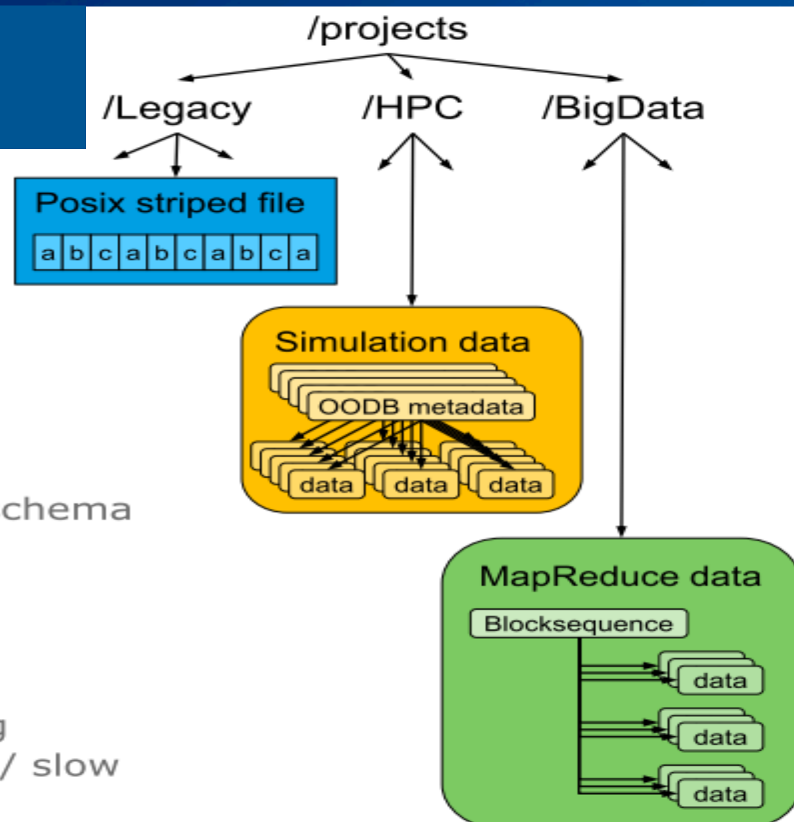
DAOS will serve as the bridge between multiple big data paradigms and also HPC

FastForward funded Big Data – HPC Bridge Architecture



The next generation storage paradigm spans HPC and BigData

- Conventional namespace
  - Works at human scale
  - Administration, security, accounting
  - Supports legacy data and applications
- DAOS Containers
  - Work at exascale
  - Separate scalable object namespace
    - Application data + metadata
    - High-level I/O models determine schema
    - Object reference invariance
  - Transactional
- Storage pools
  - Administer by usage
    - Small / random, Large / streaming
    - Low capacity / fast, High capacity / slow
  - Migration



SUMMARY

New memory technologies and Exsascale drive a different HPC storage paradigm

Big Data and HPC have similar requirements

Lets go fully object oriented. The time is now!

intel®

Legal Information

Today's presentations contain forward-looking statements. All statements made that are not historical facts are subject to a number of risks and uncertainties, and actual results may differ materially. Please refer to our most recent Earnings Release and our most recent Form 10-Q or 10-K filing for more information on the risk factors that could cause actual results to differ.

If we use any non-GAAP financial measures during the presentations, you will find on our website, intc.com, the required reconciliation to the most directly comparable GAAP financial measure.

Performance tests and ratings are measured using specific computer systems and/or components and reflect the approximate performance of Intel products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance. Buyers should consult other sources of information to evaluate the performance of systems or components they are considering purchasing. For more information on performance tests and on the performance of Intel products, reference www.intel.com/software/products.

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.

This slide MUST be used with any slides removed from this presentation

Legal Disclaimers

All products, computer systems, dates, and figures specified are preliminary based on current expectations, and are subject to change without notice.

Intel processor numbers are not a measure of performance. Processor numbers differentiate features within each processor family, not across different processor families. Go to: http://www.intel.com/products/processor_number

Intel, processors, chipsets, and desktop boards 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® Virtualization Technology requires a computer system with an enabled Intel® processor, BIOS, virtual machine monitor (VMM). Functionality, performance or other benefits will vary depending on hardware and software configurations. Software applications may not be compatible with all operating systems. Consult your PC manufacturer. For more information, visit http://www.intel.com/go/virtualization

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

Requires a system with Intel® Turbo Boost Technology. Intel Turbo Boost Technology and Intel Turbo Boost Technology 2.0 are only available on select Intel® processors. Consult your PC manufacturer. Performance varies depending on hardware, software, and system configuration. For more information, visit http://www.intel.com/go/turbo

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 select Intel® processors. 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/