Beowulf meets Exascale: A horizontally integrated framework

Mark Seager, CTO for Technical Computing Ecosystem
mark.seager@intel.com

Presented to Cetraro HPC 2014 Workshop

July 8, 2014 @ Grand Hotel San Michele, Cetraro, Italy

Legal Disclaimer

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.

NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER, AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF INTEL® PRODUCTS INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT. INTEL PRODUCTS ARE NOT INTENDED FOR USE IN MEDICAL, LIFE SAVING, OR LIFE SUSTAINING APPLICATIONS.

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.

Intel processor numbers are not a measure of performance. Processor numbers differentiate features within each processor family, not across different processor families. See www.intel.com/products/processor_number for details.

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.

Outline

Exascale Target Architecture

- Scalable System Hardware
- Scalable System Software

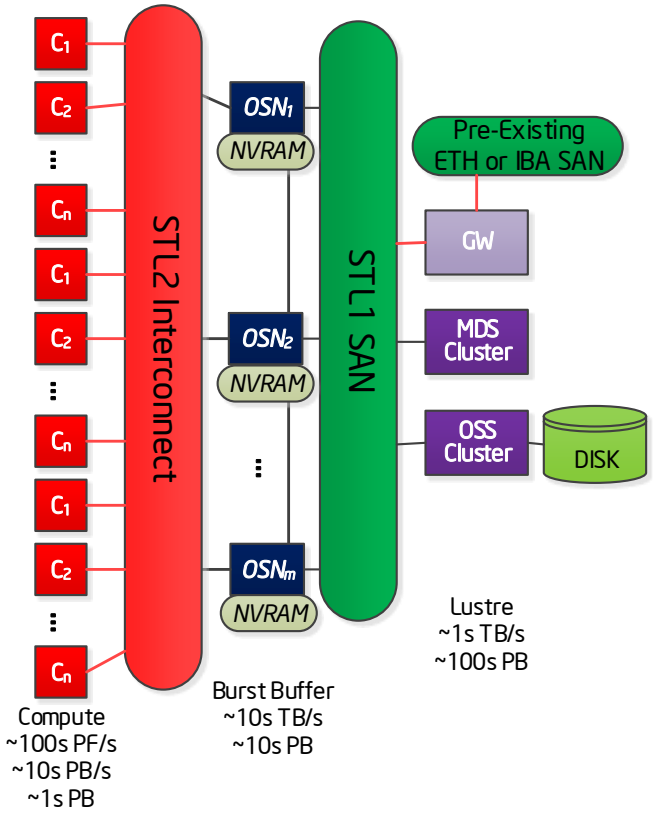
Scalable Unit Reference Design for Pre-Exascale Beowulf

Storage Scalable Unit Reference Design

Framework for a horizontally integrated scalable software stack

Pre-Exascale HPC and Big Data Scalable HW Point Designs

Pre-Exascale HPC Target Architecture



Design with system focus that enables end-user applications

Scalable hardware

- Simple, Hierarchal

Scalable Software

- Factor and solve
- Hierarchal with function shipping

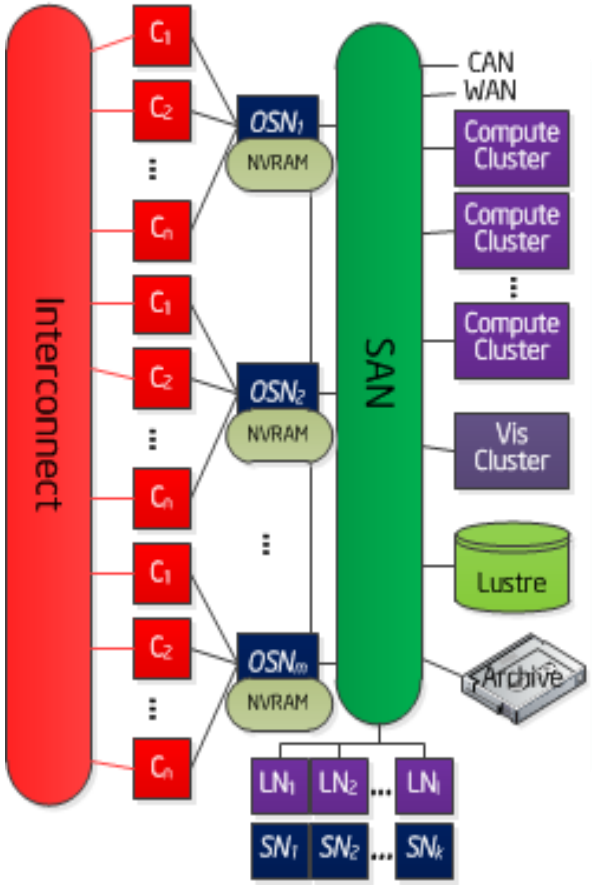
Scalable Apps

- Asynchronous coms and IO

These point designs offer evolution on HPC and new memory hierarchy for "Big Data"

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

Exascale Target Architecture



Provisioning, (logically stateless compute nodes)

System diagnostics

Network (low latency, reductionss, collectives, small msgs)

In-band system mgt

System mon

Applications

High performance parallel libraries

MPI, OpenMP PThreads SHMEM, GA, PGAS, +++

Languages Fortran, C, +++

Compilers: GNU, Intel, PGI, LLVM

Scalable debugger & performance tools

ON (tree- & mesh-based overlay networks)

Data mgt and file system unifying RAM, NVM, disk

Scalable dynamic resource manager

User-space utilities, tools, and shells, e.g., Python, perl, pipe, etc.

Unified communication stack and runtimes

Multi-OS Compute node (Linux functionality LWK performance)
OS, Login, and Service nodes: Enhanced Linux Distro

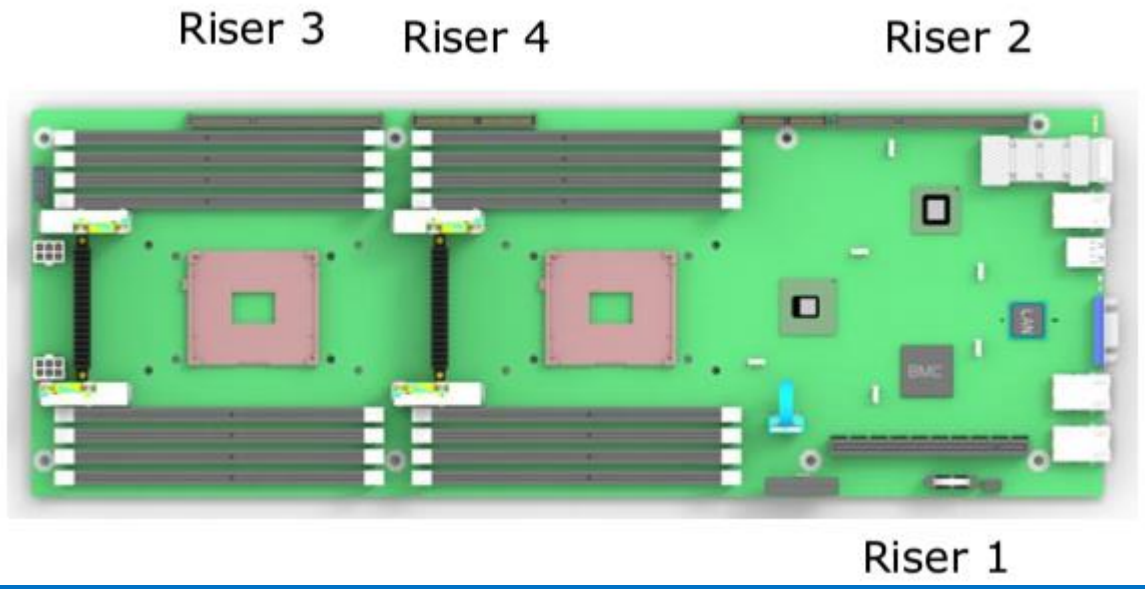
4 Pillars of RAS (gather store access process), pub-sub tree

Platform SW
(RAS interface, SMC protocol)

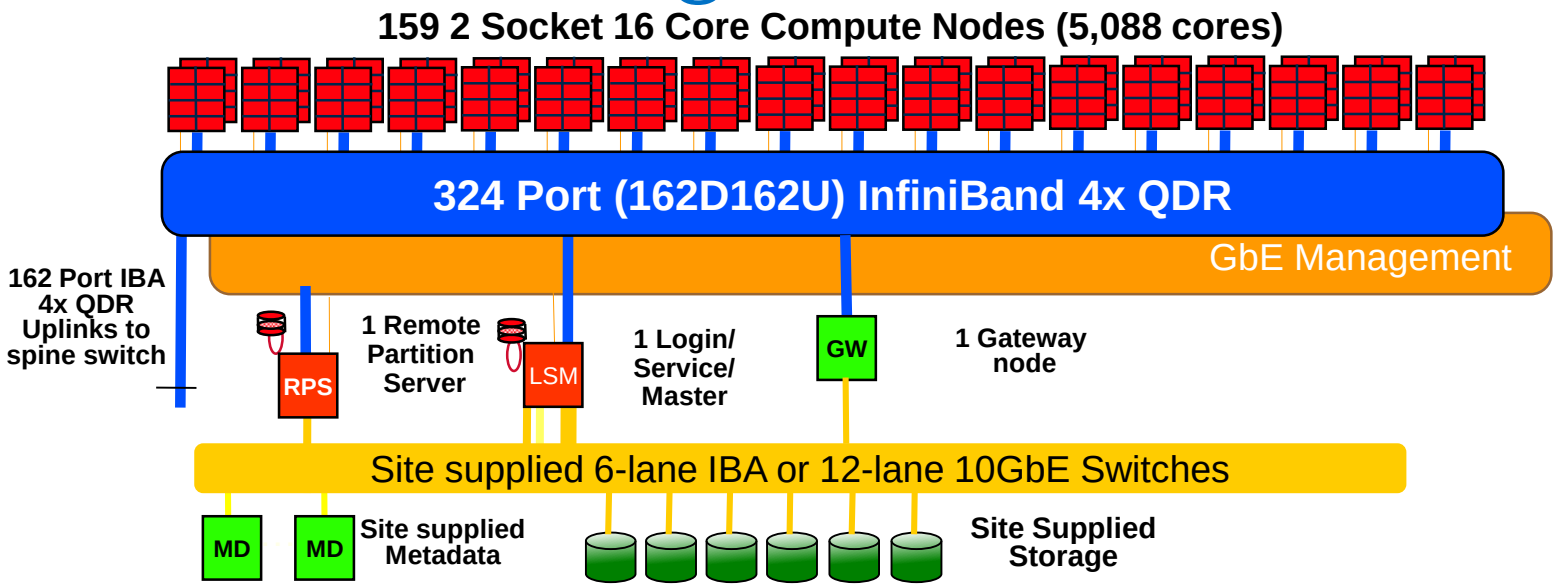
HBar nodes based on Intel Taylor Pass Motherboard

2.7 Ghz E5-2690v3 w/16 x 16GB DDR4 2133Mhz DIMMs

- 2S x2 Memory Controllers x2 channels x2 DIMM = 16 DIMMs
- 32 GB DDR4 DIMMs * 16 = 512 GB per node
  - >20 GB/core
- 1 or 2 rail TrueScale QDR IBA with Upgrade to STL1



HBar SU Configuration



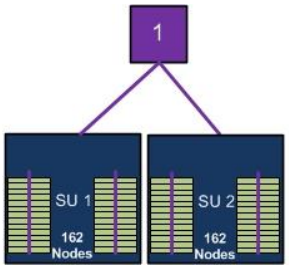
System Parameters: ~165 TF/s SU

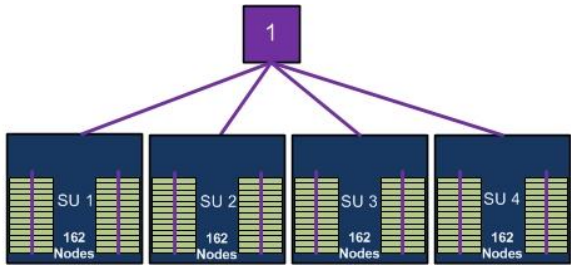
- Dual socket Intel Haswell 12 C @ 2.7 GHz nodes; 256/512 GB DDR4 (10.7/21 GB/core)
  - 16/32 GB DDR4-2133 SDRAM
- Single/Dual TrueScale InfiniBand; 4+4/8+8 GB/s Bandwidth over IBA 4x QDR
- Built from 648, 324, and 36-port IBA switches
- Compute and gateway nodes. Remote boot from RPS nodes
- IO Bandwidth ~3.3 GB/s delivered parallel I/O performance
- Software for build and acceptance Forest Peak (ICSS) V1
- May have SSD in GW nodes for accelerated checkpoint capabilities

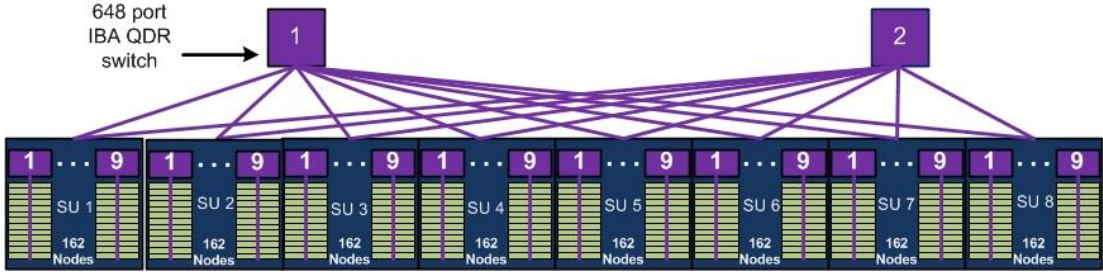
Flexibility of the Scalable Unit concept allows for a variety of cluster sizes

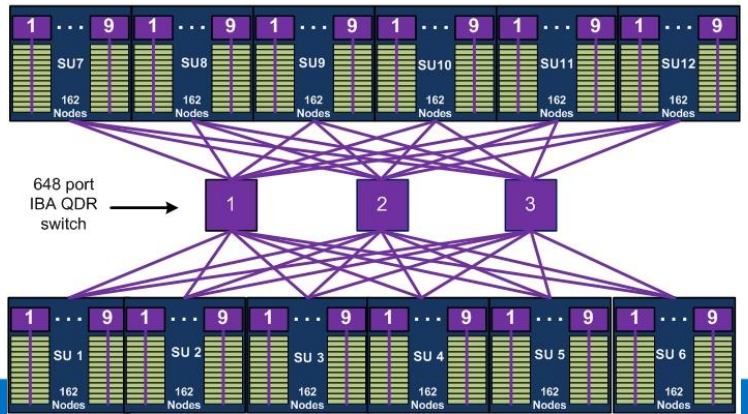
SU	2nd SW	CN	Cores@12	Peak@12c	Cores@16	Peak@16c	Mem (TiB)
1	0	159	3,816	165	5,088	187	40
2	0	318	7,632	330	10,176	374	80
4	1	636	15,264	659	20,352	749	159
8	2	1,272	30,528	1,319	40,704	1,498	318
16	4	2,544	61,056	2,638	81,408	2,996	636
32	8	5,088	122,112	5,275	162,816	5,992	1,272
48	12	7,632	183,168	7,913	244,224	8,987	1,908
96	24	15,264	366,336	15,826	488,448	17,975	3,816

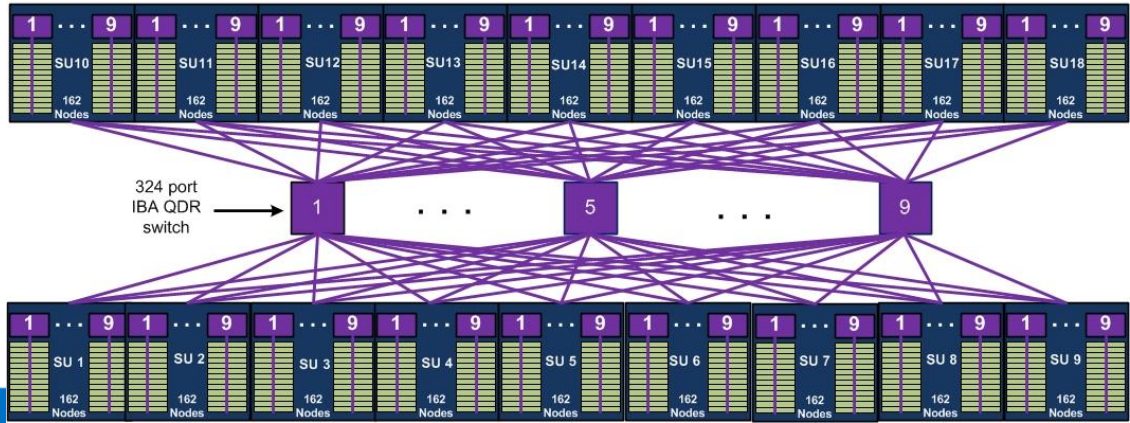
96 SU system would be #4 on June 2014 (current) TOP500 list



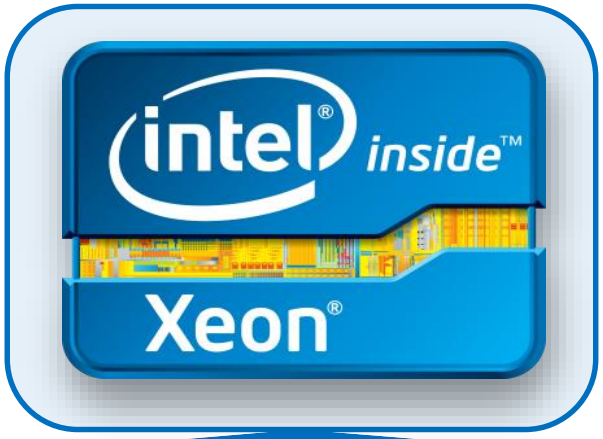








Intel's Storage Strategy: *Powering Intelligent Storage*



More Efficient Storage

- Thin provisioning
- In-line data de-duplication
- Data compression
- Scalability

Smarter Data Protection

- Distributed erasure codes
- RAID acceleration
- Integrated data integrity features

Faster Analytics & Data Placement

- Rich analytics
- Seamless data federation
- Automated data tiering with SSDs

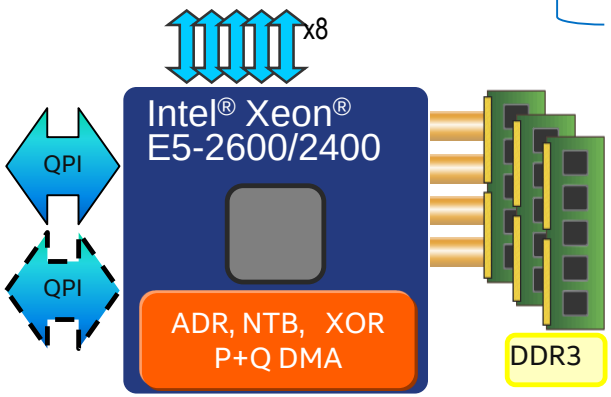
Intel® Xeon® Processor Benefits for Storage

Platform Storage Extensions

Storage Usage or Benefit	Intel Feature or Attribute
High Availability: protecting data through power disruptions	Asynchronous DRAM Refresh helps protect data in memory
High Availability: failover	PCIe Non-Transparent Bridge (NTB) connects systems over a backplane
Offload RAID calculations from the CPU with hardware RAID acceleration	• Intel® Intelligent Storage Acceleration Library (Intel® ISA-L) • Built-in high speed XOR P+Q Intel® QuickData Technology (CBDMA) & RAID-5 acceleration engine
Single write transaction to two targets to maximize memory channel bandwidth	PCIe Dual-Cast allows Host IOC to write data directly to NTB and local memory
Accelerate storage workloads, incl. tiering and thin provisioning	Intel® Advanced Vector Extensions with 256-bit vector support for integer vector operations
Data De-Duplication	• High-performance multi-Core Intel® Xeon® processors • Intel® Advanced Vector Extensions
High I/O Performance	Intel® Integrated I/O delivers reduced latency and higher bandwidth
Data Encryption	Intel® AES-New Instructions
Compression	Multi-Core Intel® Xeon® processors and optional chipsets with Intel® QuickAssist Technology
Erasure Coding	Intel® ISA-L
Reliability, availability and serviceability (RAS)	Comprehensive RAS features incl. mirroring, sparing, demand and patrol scrubbing, ECC, plus many more

Intelligent Storage:
The first converged storage-server processor starting with Haswell

Platform Storage Extensions



PCIe 3.0 Integration

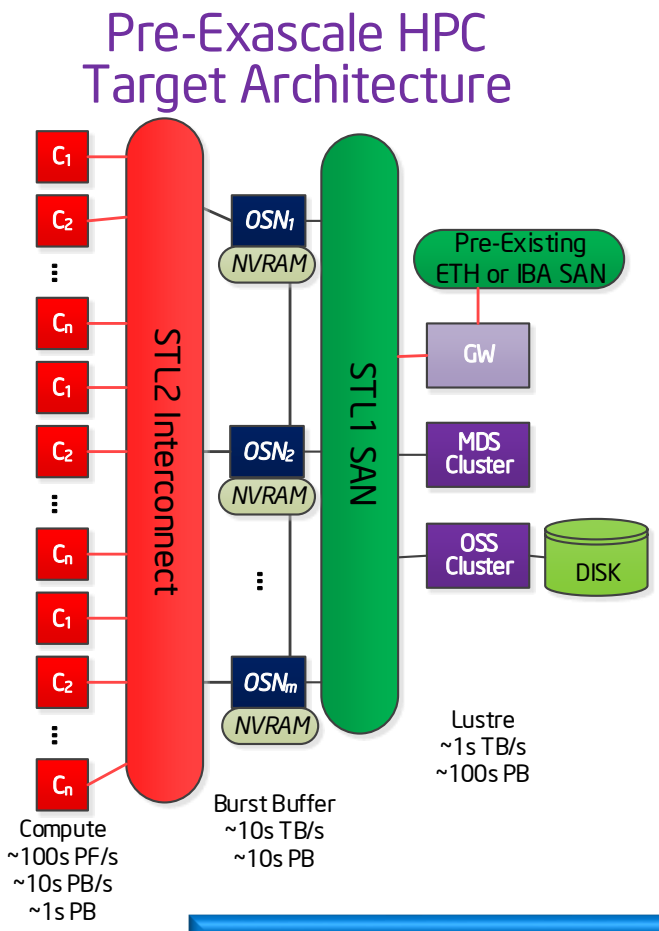
Storage Usage	Intel Feature or Attribute
High Availability: protecting data through power disruptions	Asynchronous DRAM Refresh helps protect data in memory
High Availability: failover	Non-Transparent Bridge (NTB) connects systems over a backplane
Offload RAID calculations from the CPU with hardware RAID acceleration	• Intel® Intelligent Storage Acceleration Library (Intel® ISA-L) • Built-in high speed XOR P+Q Intel® QuickData Technology (CBDMA) & RAID-5 acceleration engine
Accelerate storage workloads, incl. tiering and thin provisioning	Intel® Advanced Vector Extensions provides up to 2x throughput with 256b FP vs. Intel® Xeon® processor 5600 series
Data De-Duplication	• Multi-Core Intel® Xeon® processors with up to **80% performance boost** vs. prior gen[1] • Intel® Advanced Vector Extensions
High I/O Performance	New Intel® Integrated I/O delivers up to 3X higher I/O performance[2]
Data Encryption	Intel® AES-New Instructions
Compression	Multi-Core Intel® Xeon® processors and optional chipsets with Intel® QuickAssist Technology
Erasure Coding	Intel® ISA-L
Reliability, availability and serviceability (RAS)	Comprehensive RAS features incl. mirroring, sparing, demand and patrol scrubbing, ECC, plus many more

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.

1 Over previous generation comparable Intel® Xeon® processor 5600 series· Intel internal estimate. For more legal information on performance forecasts go to http://www.intel.com/performance

2 Intel measurements of average time for an I/O device read to local system memory under idle conditions. Improvement compares Xeon processor E5-2600 family vs. Xeon processor 5600 series

How can I build flexible Cluster File System configurations cost effectively?



Cluster File System composed of

- Multiple Scalable Storage Cluster (SCC) each with a single Lustre Mount point

SCC

- Multiple Storage Scalable Units (SSU) configured for Lustre OSS and MDS

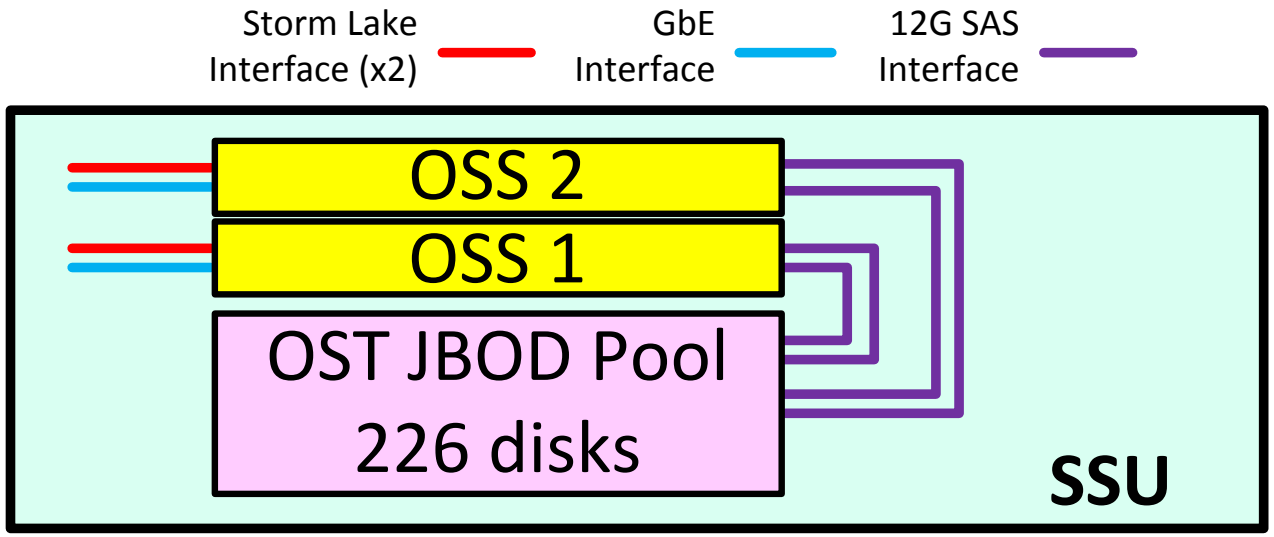
SSU

- Designed around a reasonably small storage increment
- Used to scale up capacity, bandwidth and directory operations
- Key enablers are future Lustre enhancements in ZFS software RAIDZ performance and declustered RAIDZ

Intel Enterprise Lustre can integrate SSD and multiple tiers of storage with HSM

This storage Scalable Storage Unit based file system construction offer evolution on Linux Clusters SU concept that puts cost effective config options in your hands

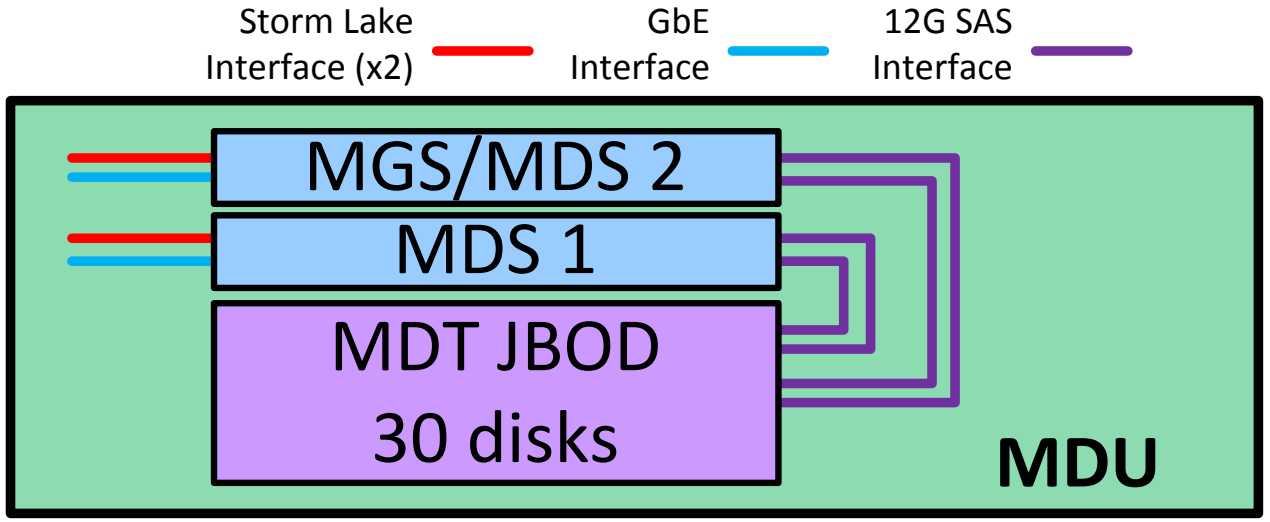
Scalable Storage Unit Architecture for Lustre Object Storage System



Scalable Storage Unit Characteristics

- OSS 1 & 2 are AA-HA failover pair
- 3x4U 84 3.5" disk arrays in 8*(8P+2P)+4 HS
- OST JBOD pool includes 240 (D+P) + 12 HS drives
  - 8TB drives for CFS1, 6TB drives for CFS2
- 8+2P RAID6 with distributed parity
- 2 vdevs – 113 disks each
- Target 14U packaging
- Intel Enterprise Lustre with ZFS based 1 OST per OSS

Metadata Unit Architecture for Lustre Meta Data Server



Metadata Unit Characteristics

- MDS 1 and MGS/MDS 2 are high availability failover pair
- MDT includes 28 data disks + 2 spare disks
- RAID10
- 2 vdevs – 14 disks each
- Target 6 rack U solution
- MDS performance can be accelerated by replacing HDDs with PCIe SSDs

Scalable Storage Cluster Architecture

Scalable Storage Cluster

- 8 Scalable Storage Units
- 1 Metadata Unit
- GbE management switches
- Three 42U racks
- Rack 1+2 with 1,512 HDD
- Rack 1+2+3 with 2,016 HDD
- Intel Lustre with ZFS enhancements

CFS1 (8 TB HDD, 67.5 MB/s/HDD)

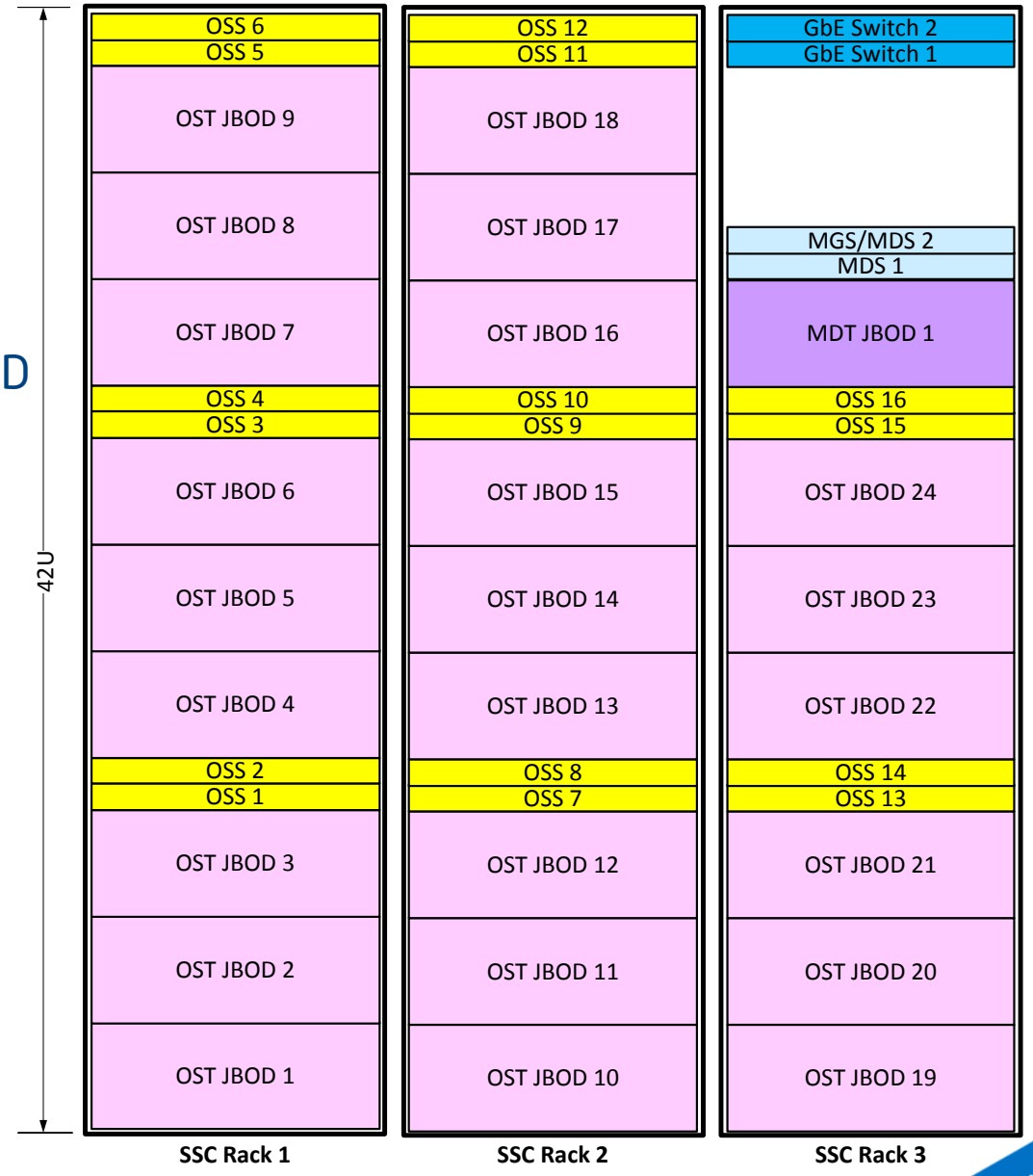
- 12.3 PB per SSC
- 103.5 GB/s per SSC

CFS2 (6 TB HDD, 67.5 MB/s/HDD)

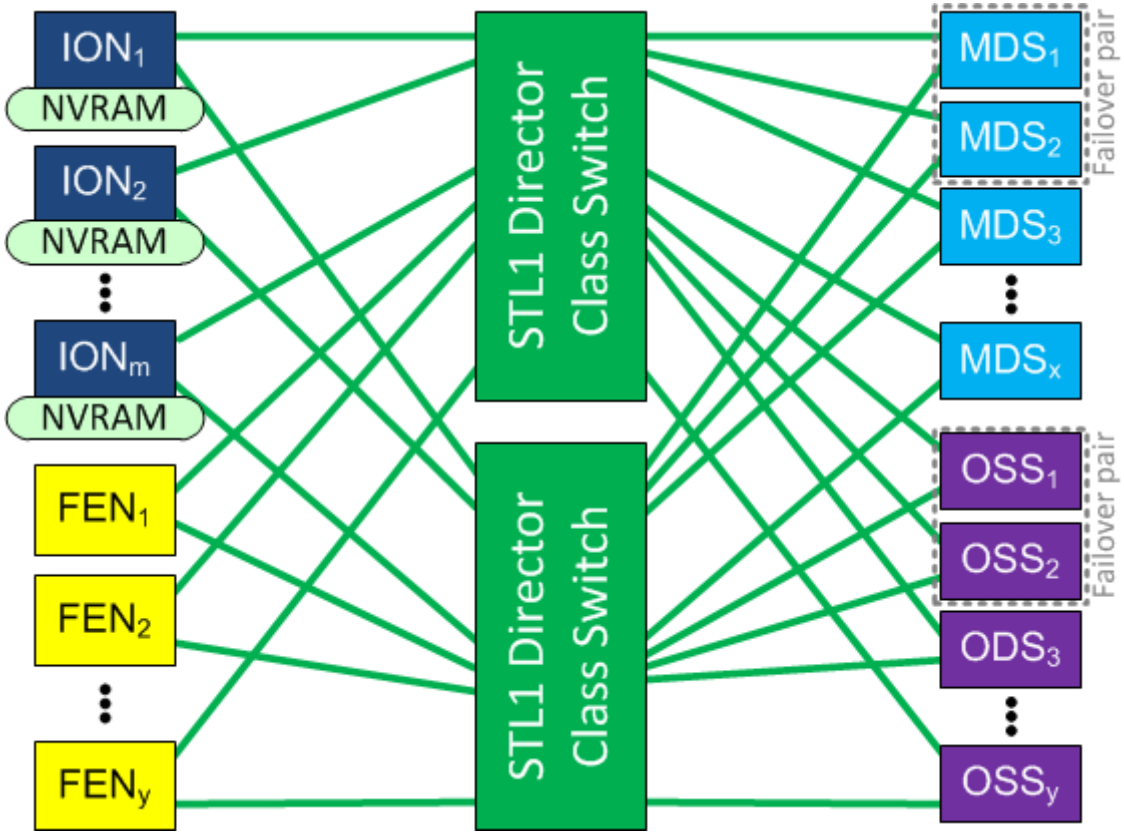
- 9.2 PB per SSC
- 103.5 GB/s per SSC

Stretch Goal (80 MB/s)

- 123 GB/s per SSC



Storm Lake Network (SAN) provides CFS design and expansion flexibility



Exascale System Software Stack Horizontal Integration Elements

Architectural Elements

- TBON (based on MRNET)
- Publish, Subscribe APIs
- Notification Transport APIs
- God network API (GOSSIP protocols)
- RAS and other system state clustered DB
- System Services

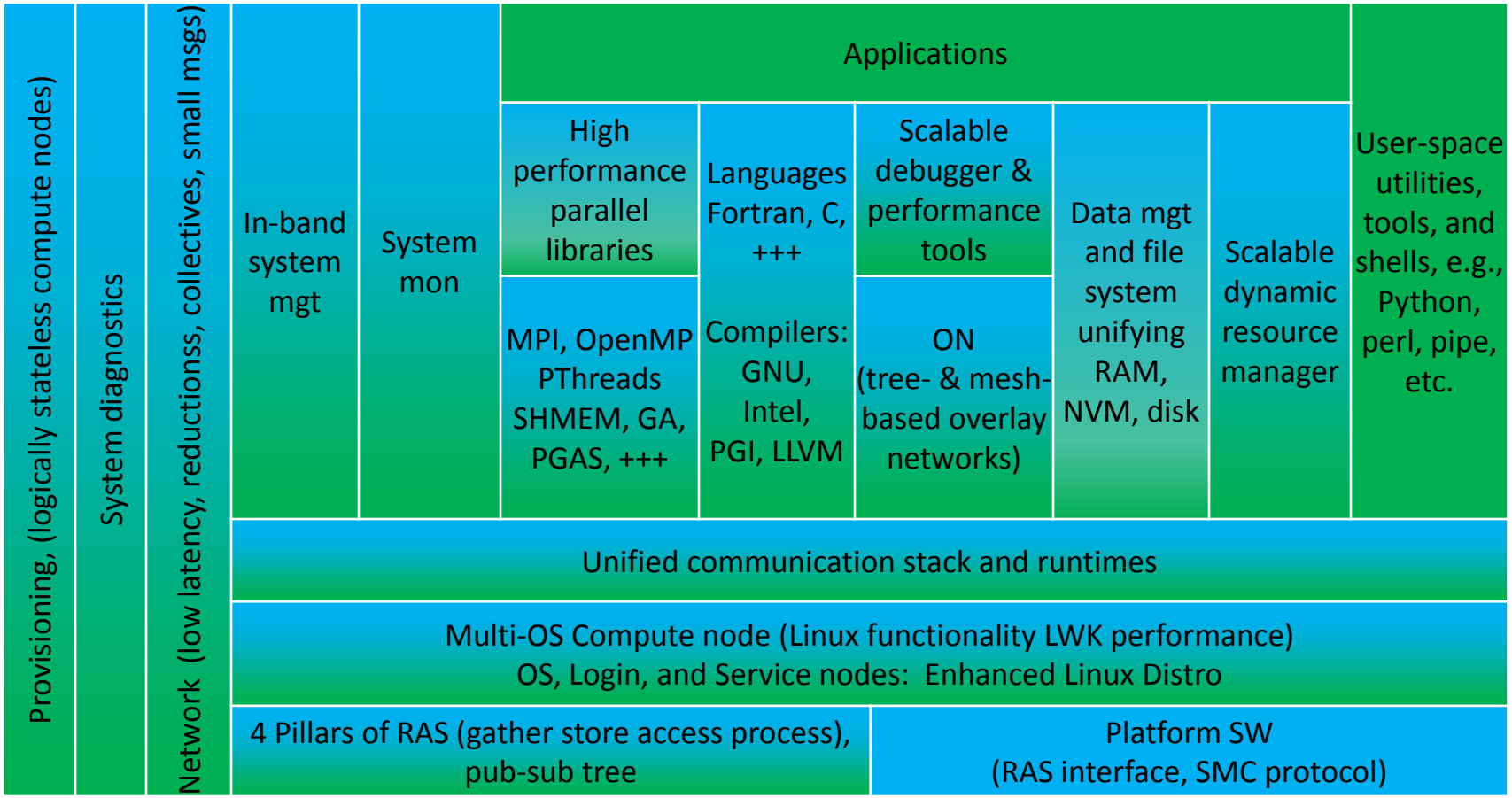
Enhancements for Hierarchal Model

- Multi-OS compute node kernel
- Function shipping: CN←→OSN+BB
- Four Pillars of RAS
- New storage paradigm for Big Data

Horizontal Integration enables the IA Ecosystem to meet the challenges Exascale systems

System Software Stack Progression

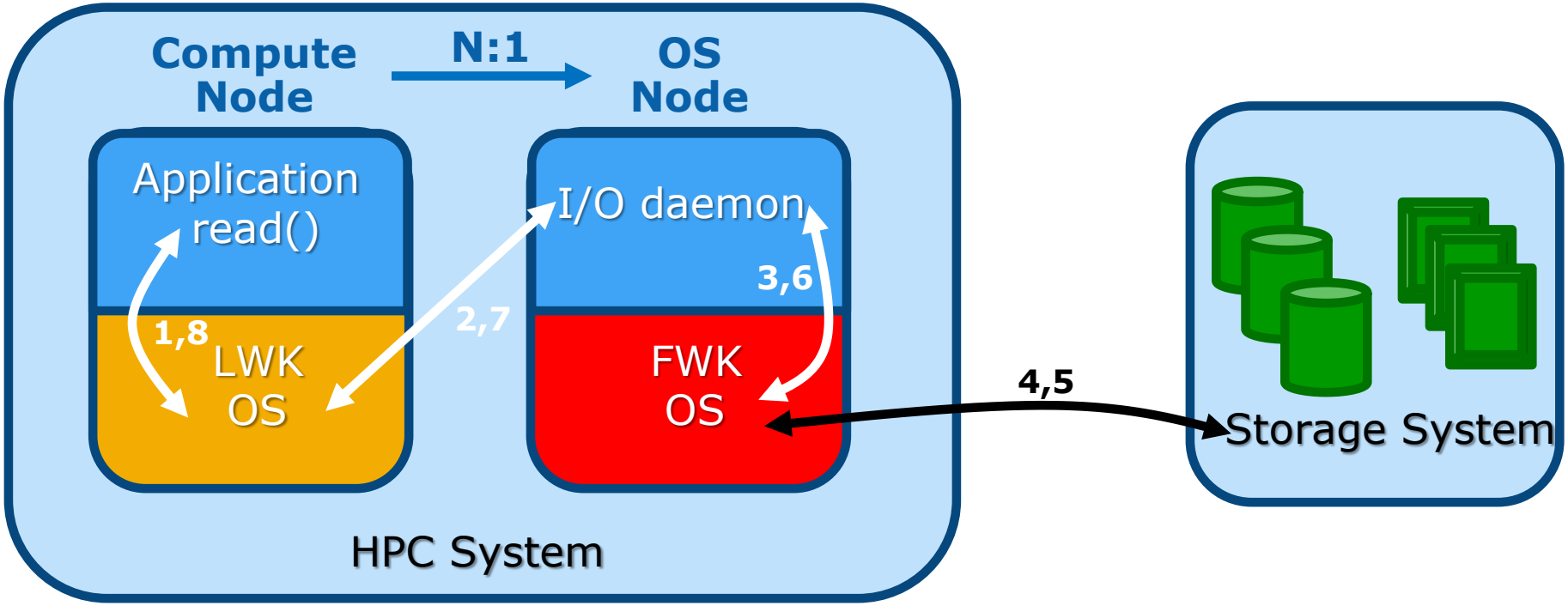
(Pre-Exascale circa 2018)



Intel products, projects, or contributions

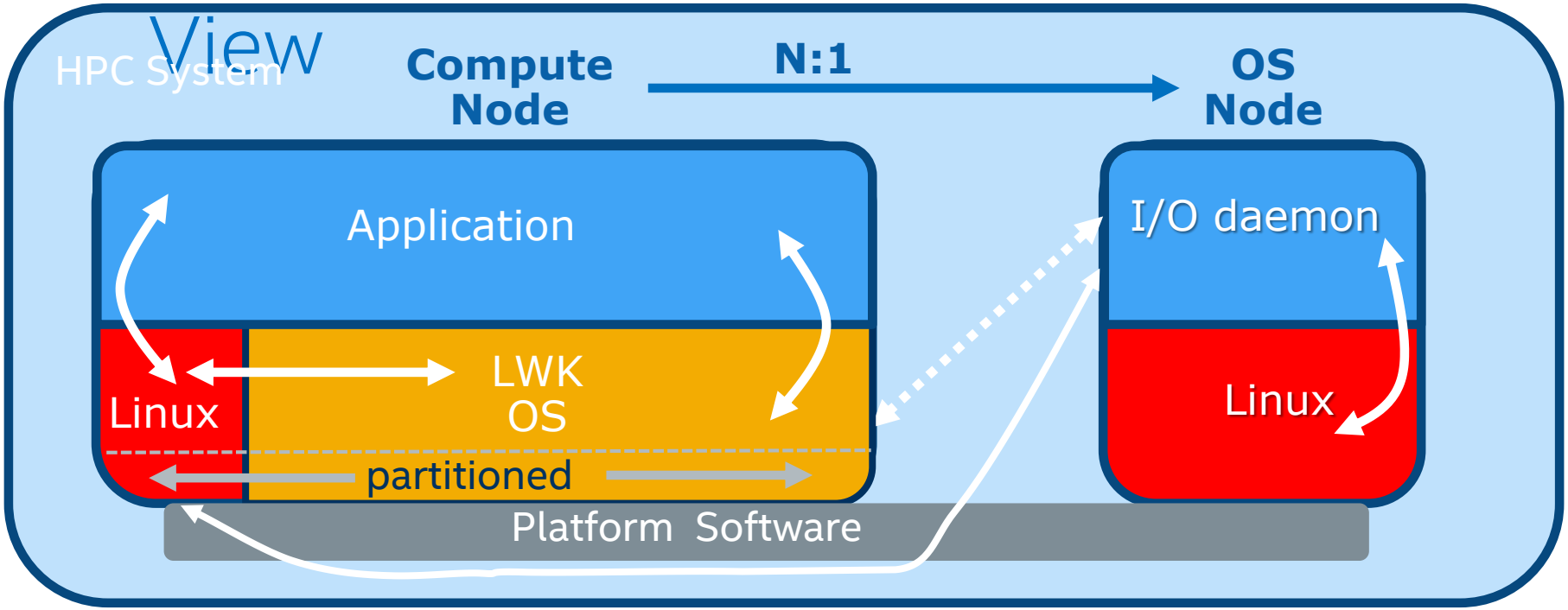
3rd party or community

OS I/O Offload



Aggregation (file systems can not handle 100sK+ clients)
Noise reduction
Reduced cache and memory pollution

OS Expanded Compute Node View



Core specialization

LWK performance for core HPC application resource requests

- Nimble to adapt to new technology, fine-grained threading

True Linux for application (and runtimes and tools) on compute node

- Previous approach was Linux compatible

Ability to leverage core variability on the compute node

- Heterogeneous not currently in plan, but manufacturing variation produces differences

System Management

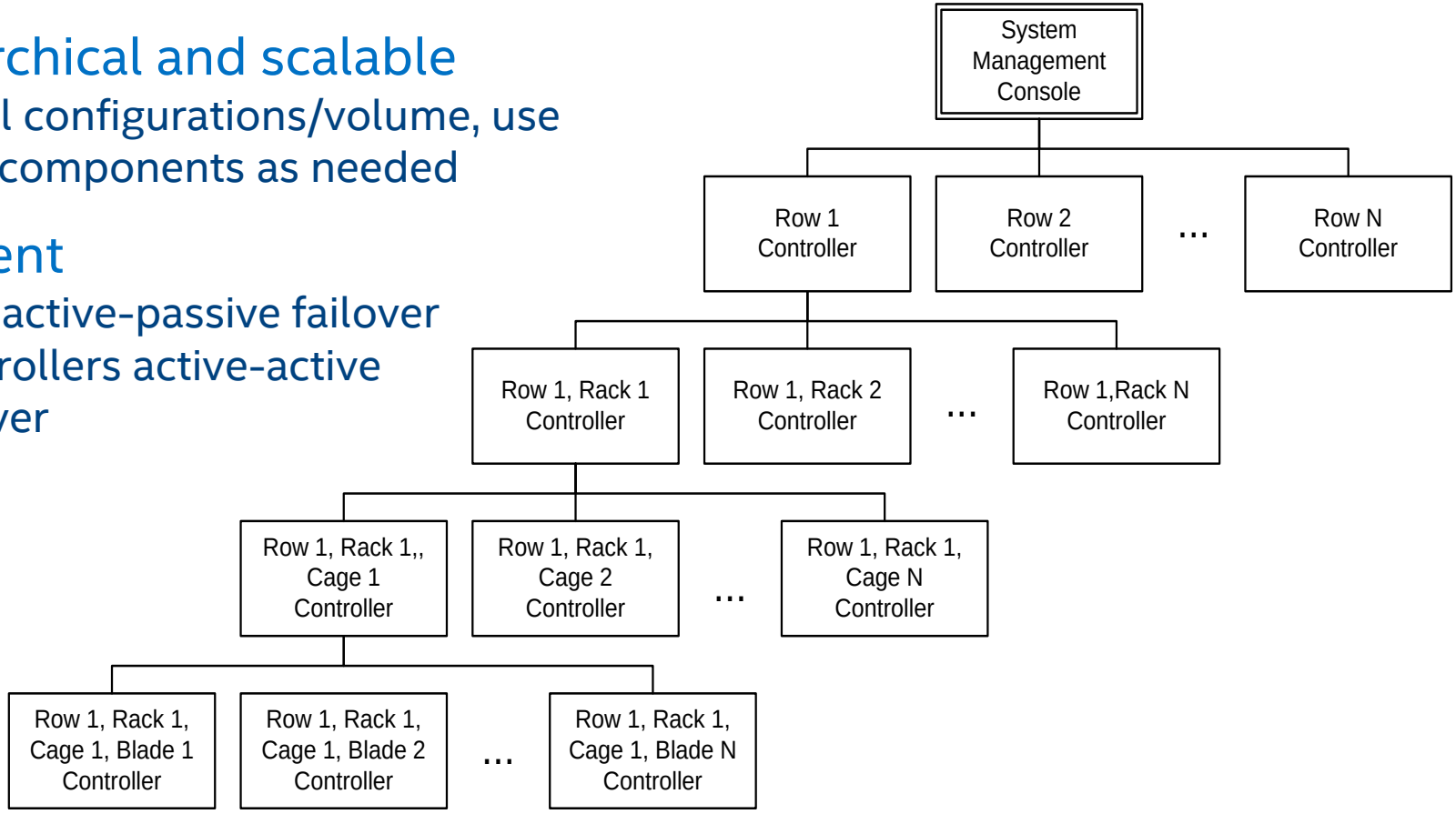
Provide single comprehensive view of system

Hierarchical and scalable

- Small configurations/volume, use only components as needed

Resilient

- SMC active-passive failover
- Controllers active-active failover



Power

HPC jobs managed as unit
- Versus transactional loads

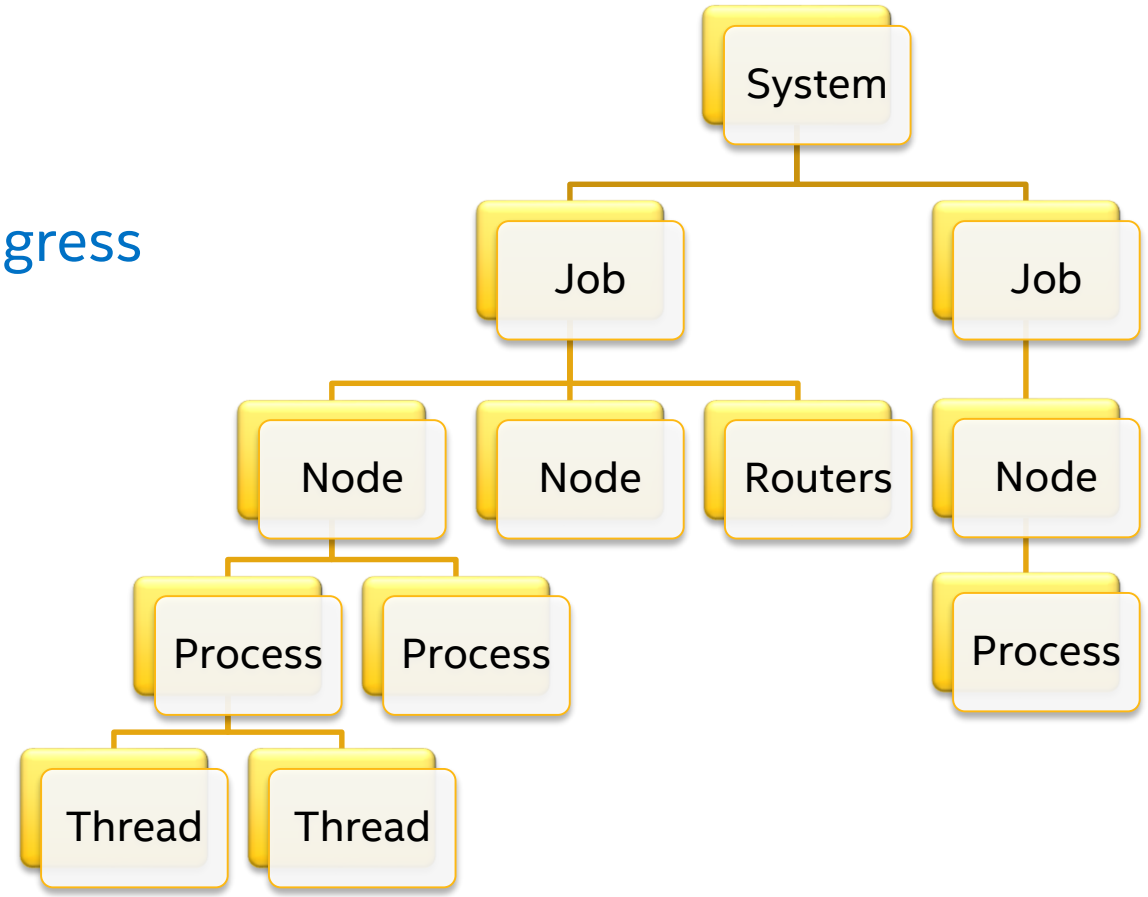
Hierarchical

Manage application progress

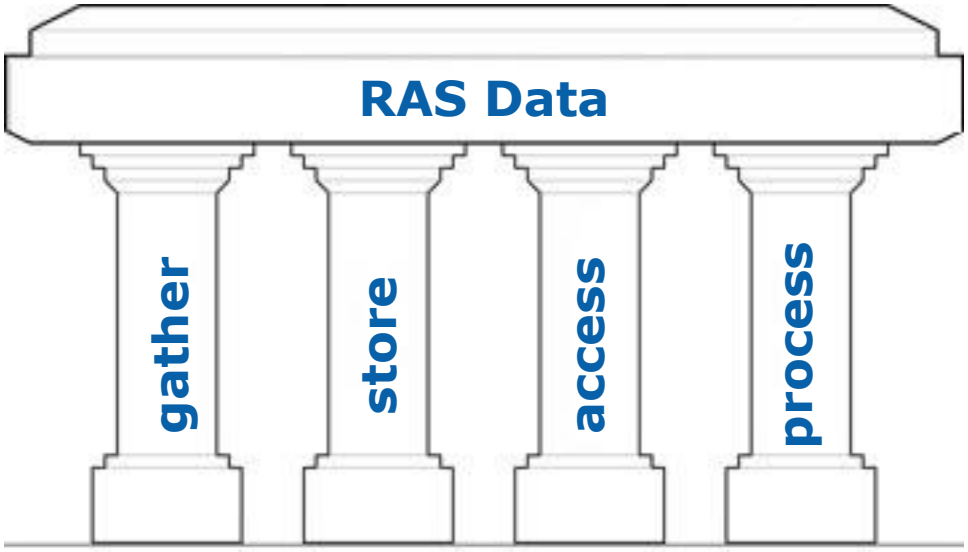
More dynamic

Schedule for KW hours
- In addition to CPU hours

Job placement
- Different topology tradeoffs



Scalable RAS Infrastructure



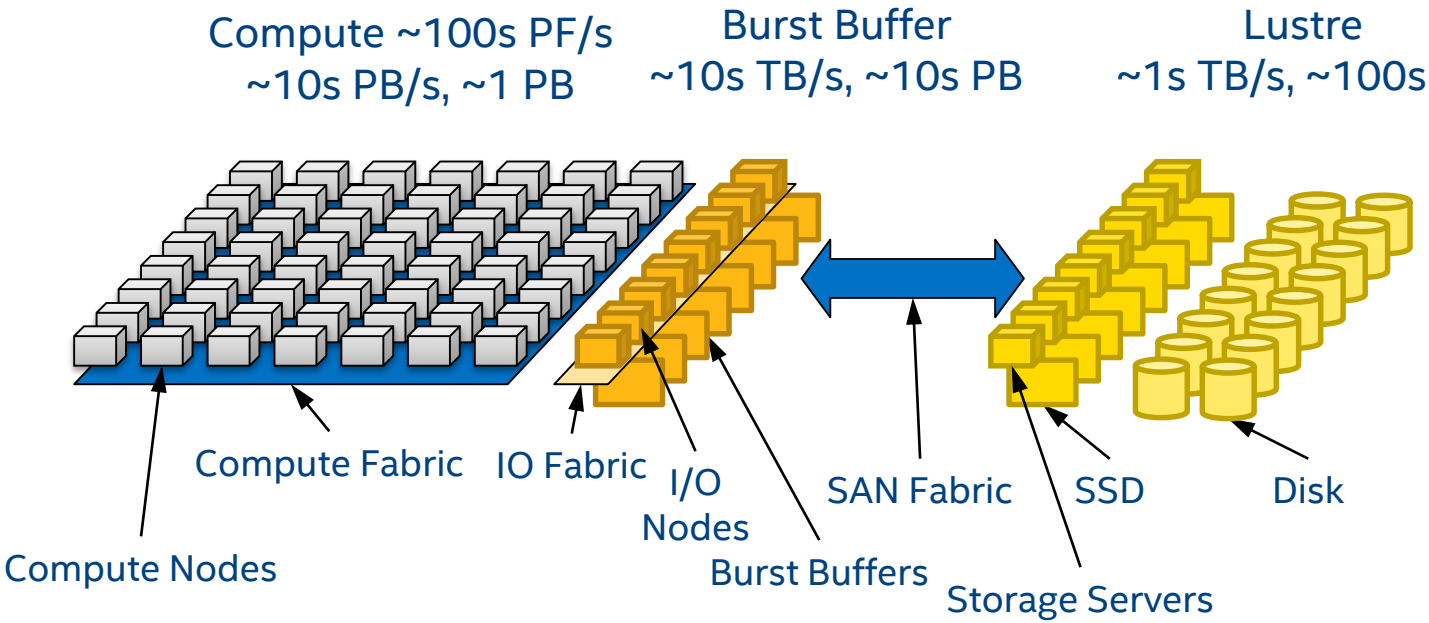
Gather: As extensive as possible, consistent format

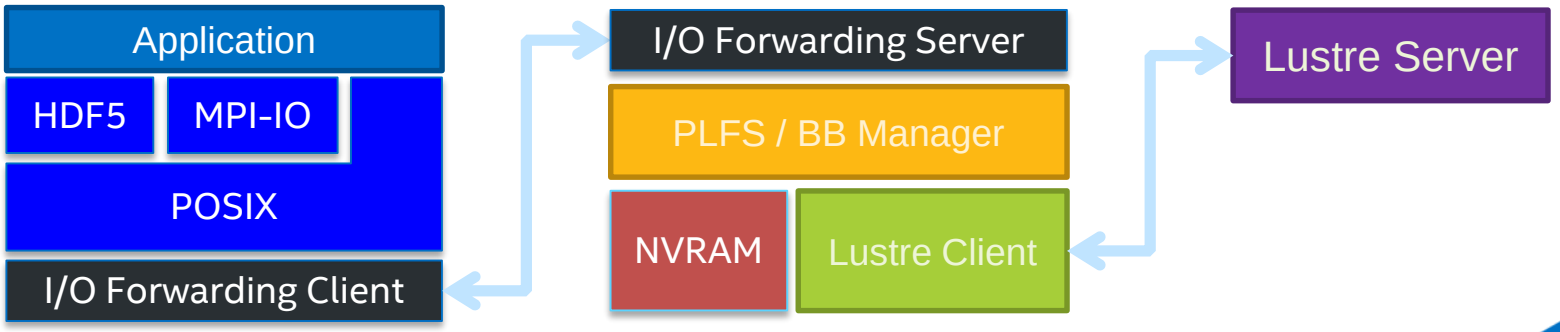
Store: Database for searching and associating

Access: Real-time pub-sub access by all components

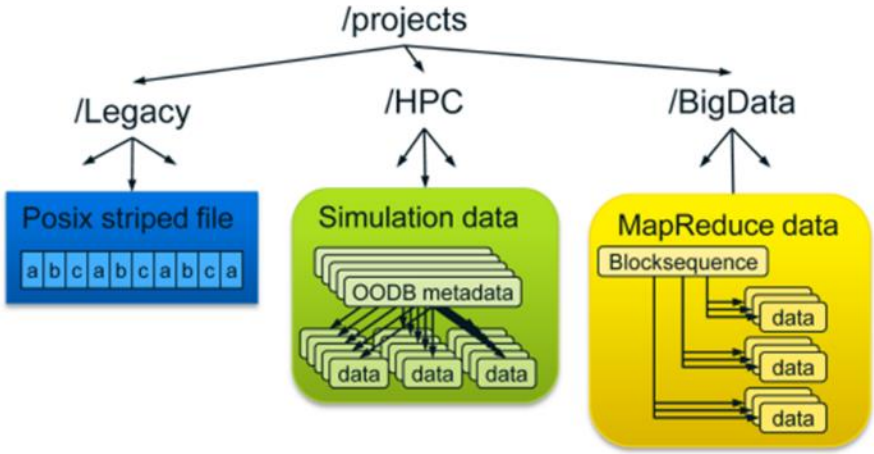
Process: Agents aggregate, trigger, notify, filter, etc.

HPC+Big Data I/O Architecture Leverages NVRAM for Bandwidth and Disk for Capacity





DAOS Containers



Virtualizes Lustre's underlying object storage

- Shared-nothing
  - 10s of billions of objects
  - Thousands of servers

Private object namespace / schema

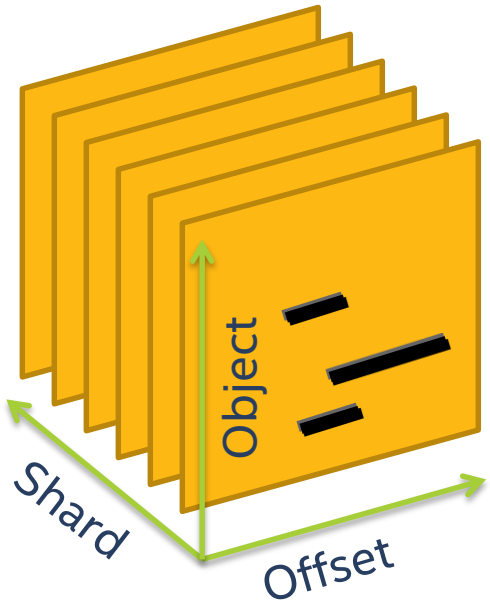
- Filesystem namespace unpolluted

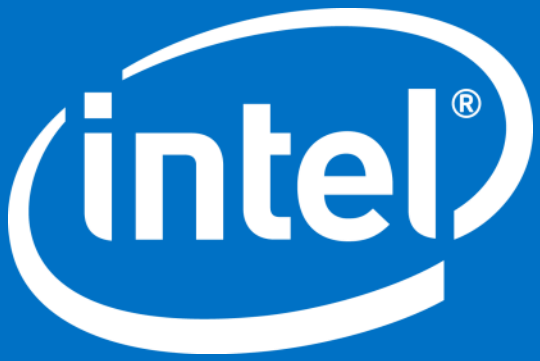
Transactional PGAS

- Baseline: addr = <shard.object.offset>
- HA: addr = <ha-schema.object.offset>

Read & Write

- No create/destroy
- Punch == store 0s efficiently




intel®