Intel® Scalable System Framework for Everyscale Computing

Mark Seager

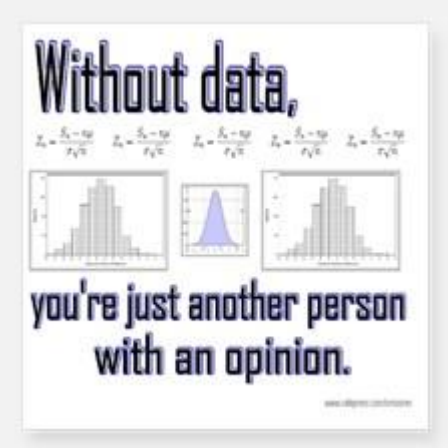
Intel Fellow, CTO for the HPC Ecosystem

High Performance Computing Group

Presented to RMACC @ CSU Fort Collins
August 10, 2016

Agenda

- Emerging HPC computing paradigms/workflows
  - Data is the new Bacon
  - Democratization of HPC requires scale down and scale up
- Scalable System Framework – What is it?
- What is the SSF value proposition?
- Compute Scalable Units for building Linux Clusters
- Storage Scalable Units for building Lustre parallel file systems

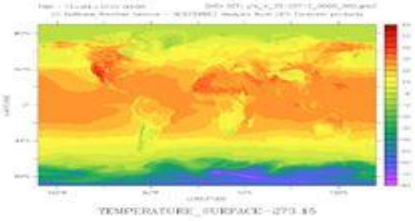


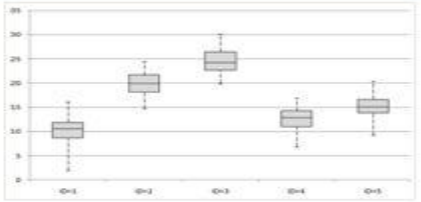
Big Data Analytics for Climate Modeling

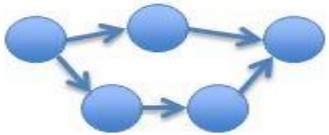
Requirements and needs focus on:

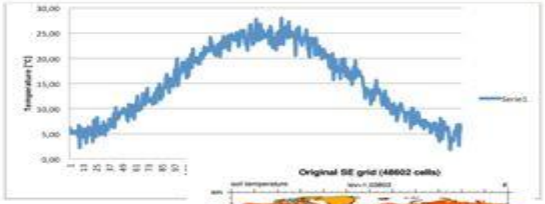
- *Time series analysis*
- *Data subsetting*
- *Model intercomparison*
- *Multimodel means*
- *Massive data reduction*
- *Data transformation (through array-based primitives)*
- *Climate change signal*
- *Maps generation*
- *Ensemble analysis*
- *Worflow support*
  - *Tens, hundreds of tasks*
- *Metadata management support*
- *….*

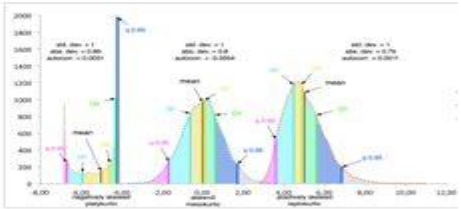
Big data analytics Framework

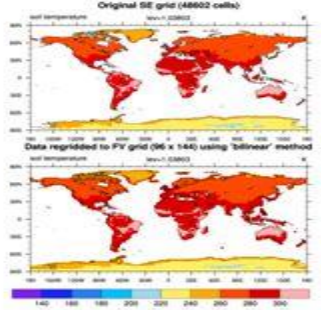












The Data Intensive Infrastructure for Autonomous Vehicles Machine Learning

In-Vehicle
(Scoring)

Autonomous Driving Functions
Trajectory Enumeration, Path Planning, Selection & Maneuvering
Driving Policy, Path Selection

Real Time Environment Modeling
Localization

Sensor Processing & Fusion
Object ID & Classification

Anomaly Detection

Real Time HD Map Updates

Compressed DL Model

OTA SW /FW Updates

5G

Captured Sensor Data

Data Formatting, Annotation Annotation Tools

Data Storage

Dataset Management & Traceability

Data Center
(Training & Scoring)

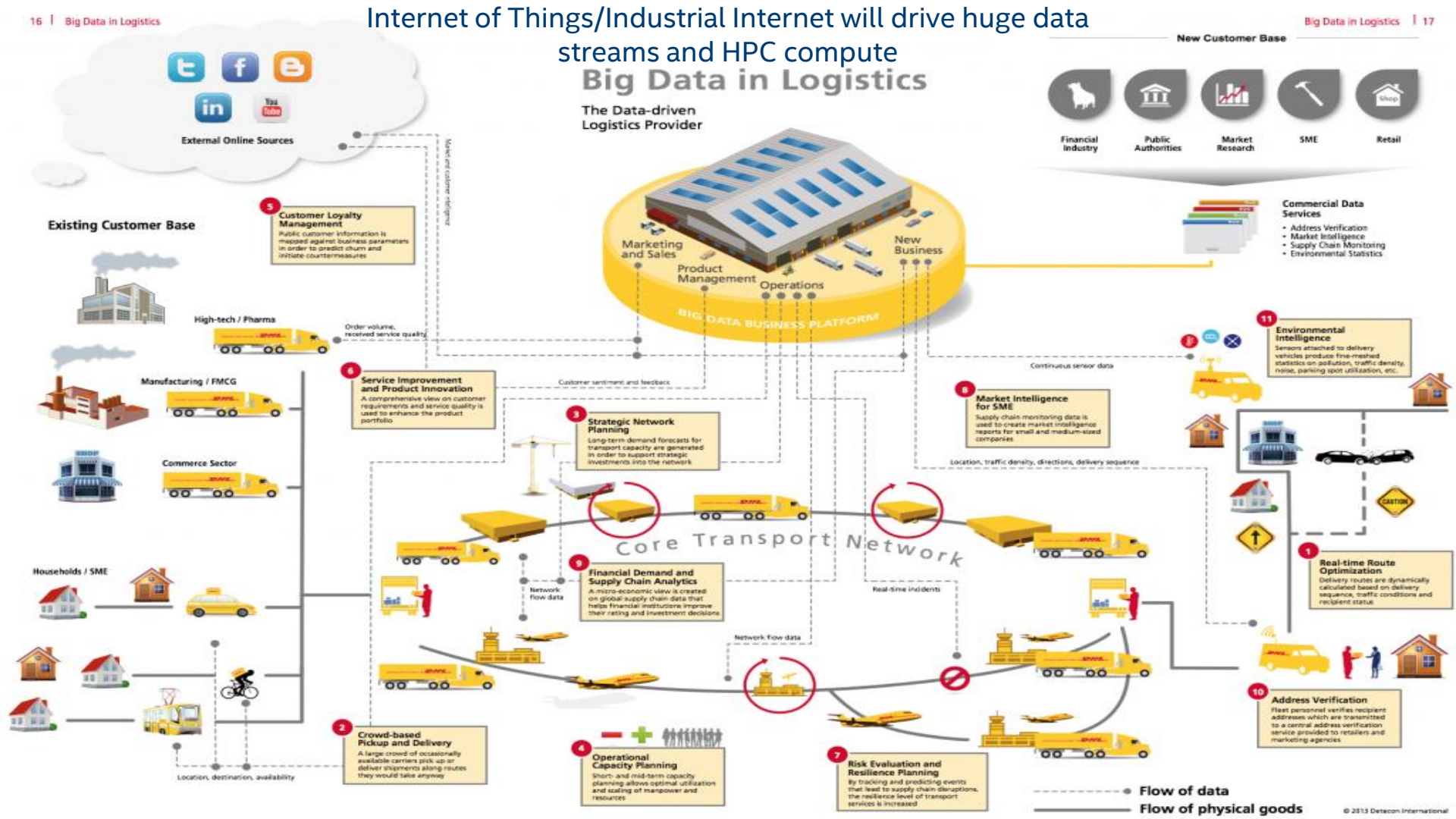
Neural Network Design for Target Hardware, & usage
(Vision, Data Driven, etc.)

Model Training
Single & Multi-node optimized Frameworks

Model Scoring
>than Real Time Model Simulation & Verification

intel

Internet of Things/Industrial Internet will drive huge data streams and HPC compute



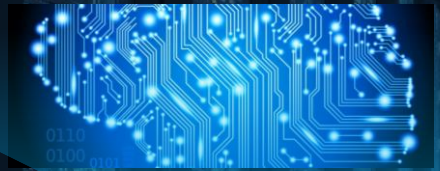
Scalable System Framework – What is it?

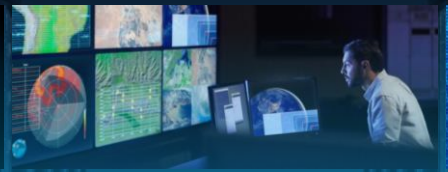
Many workloads | One architectural framework

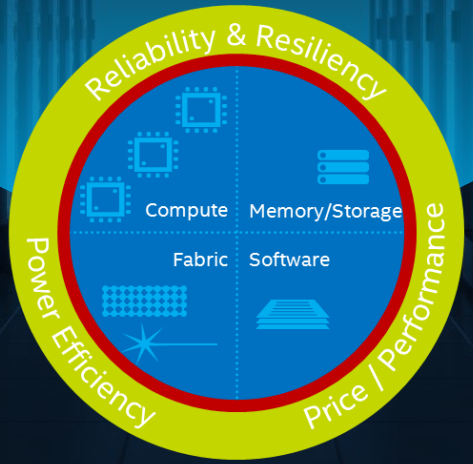
Machine Learning

Modeling & Simulation

High Performance Data Analytics







Small Clusters Through Supercomputers

Compute and Data-Centric

Standards-Based Programmability

On-Premise and Cloud-Based

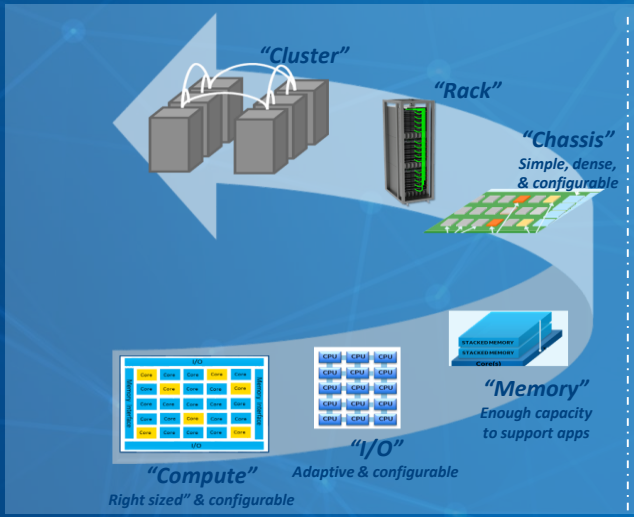
Intel® Scalable System Framework

intel

SSF: Enabling Configurability & Scalability

from components to racks to clusters ➔ Next Step is Solutions

SSF Path To Exascale



- Xeon or Xeon-Phi – based on workload needs
- Compute flexibly aggregated
- Lowest latency compute to compute interconnect

SSF for Scalable Clusters

- I/O Topologies for best performance
- Configurable I/O bandwidth director switch
- Burst buffer to decouple storage from I/O

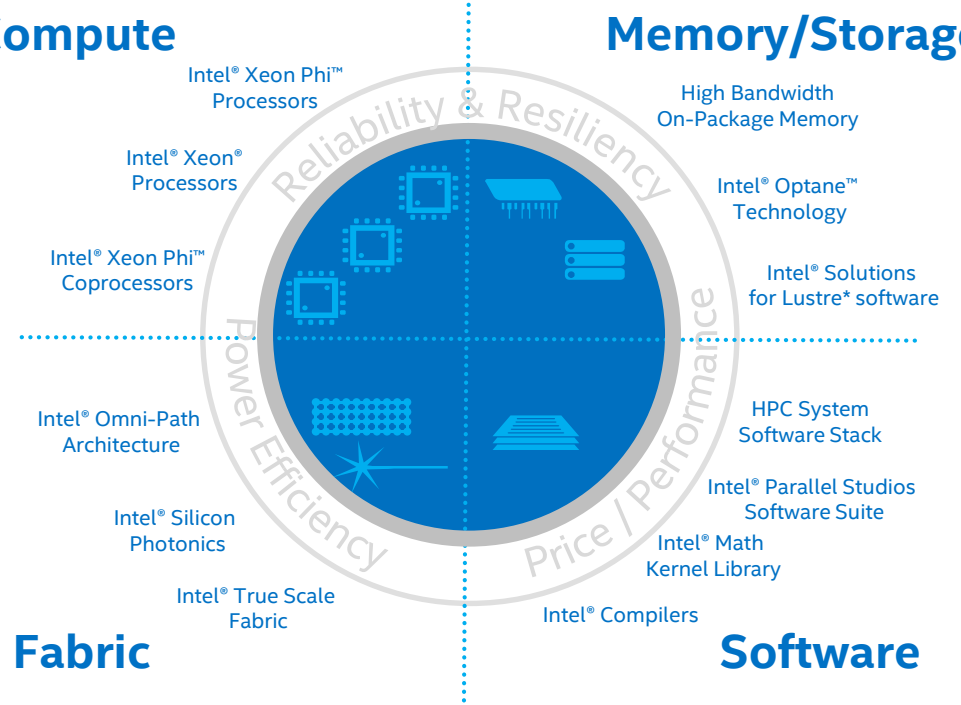
What is the SSF value proposition?

Jack Johnson - Better Together - YouTube
https://www.youtube.com/watch?v=seZMOTGCDag

Lyrics: Mmm, it's always better when we're together / Yeah, we'll look at them stars when we're together / Well, it's always better when we're together / Yeah, it's always better when we're together… Full lyrics on Google Play
Artist: Jack Johnson
Album: In Between Dreams
Released: 2005

How It Works

Innovative Technologies

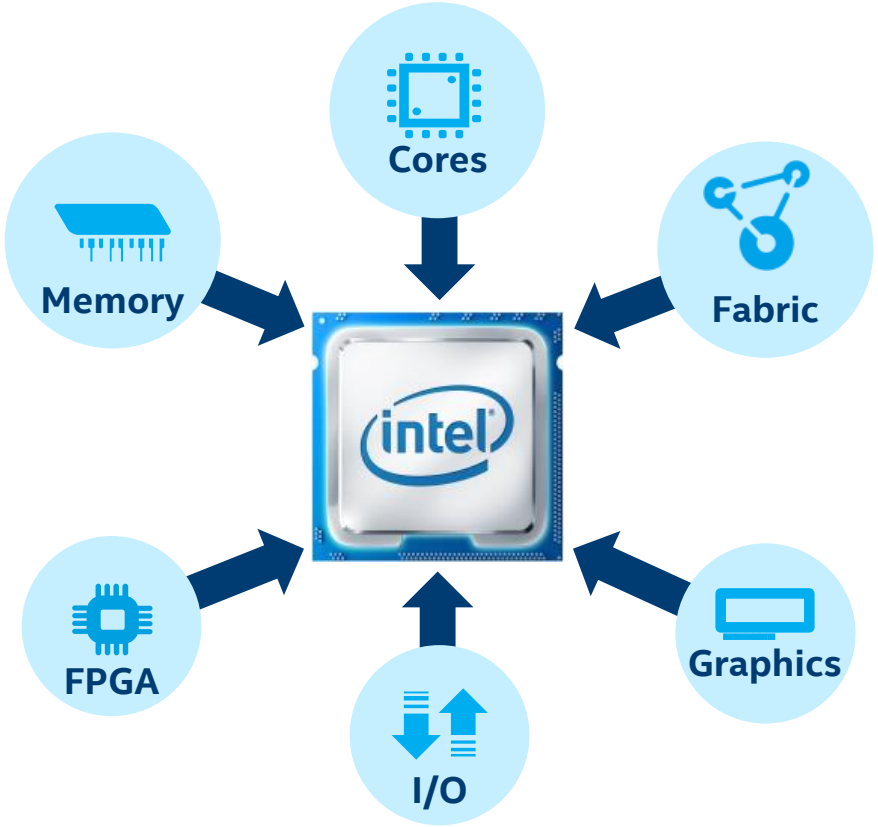


Tighter Integration
and Co-Design

Increased System Density
Reduced System Power Consumption

Intel = ***INT***egrated ***El***ectronics, Founded 1968, 48 YaG

Tighter Component Integration



Intel® Scalable System Framework

Benefits

Bandwidth

Density

Latency

Power

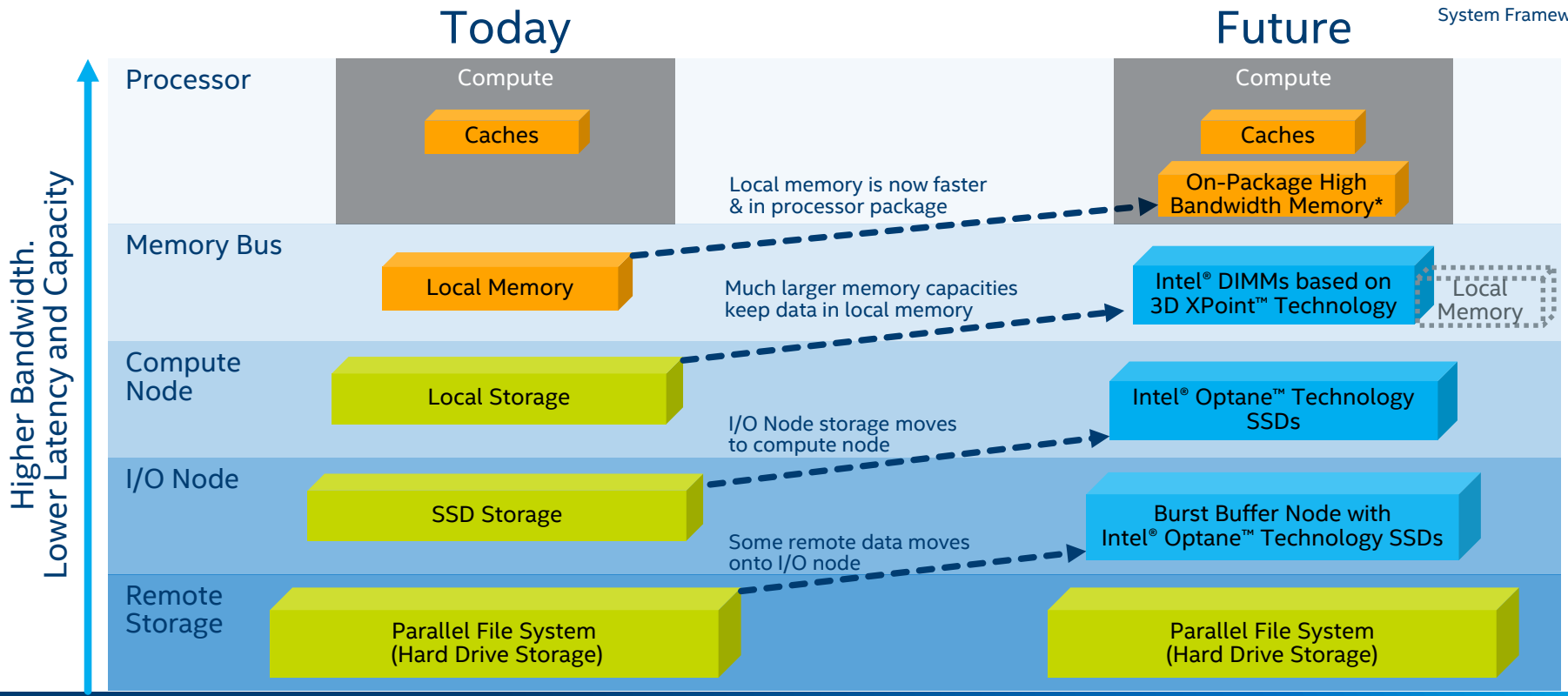
Reliability

Cost

Tighter System-Level Integration

Innovative Memory-Storage Hierarchy

Intel® Scalable System Framework



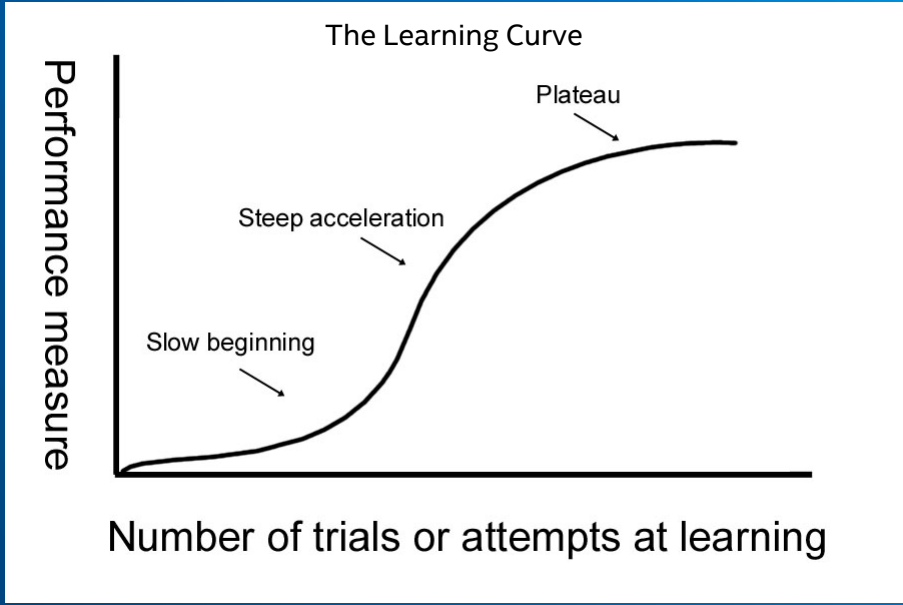
*cache, memory or hybrid mode

Compute Scalable Units for building Linux Clusters

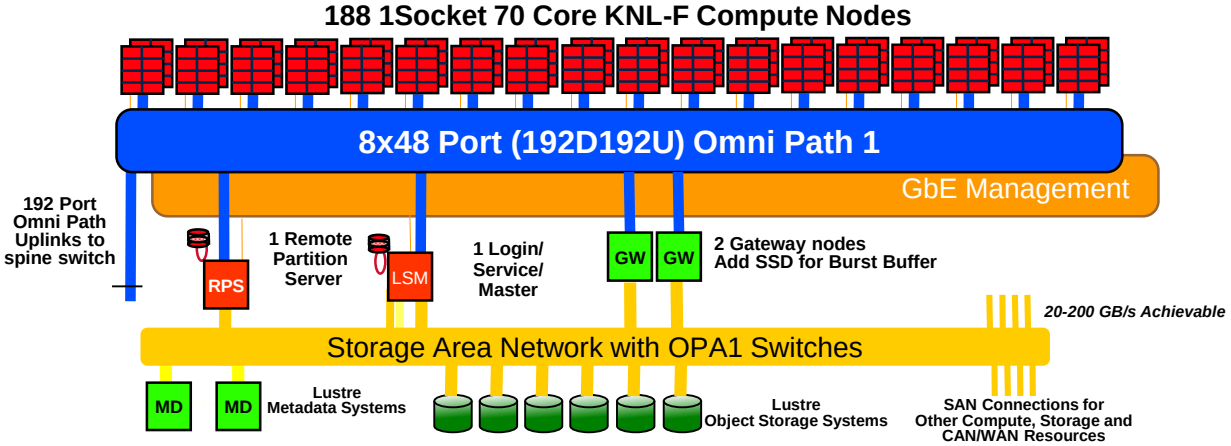
Learning Curve Observation: As one does an activity repeatedly, one's productivity improves and then plateaus

Don't just build one large Linux cluster, build many of them configured on the customer budget and workload requirements

Requires a scalable unit approach to building Linux Clusters



KNL+STL1 Compute Intensive SU Configuration



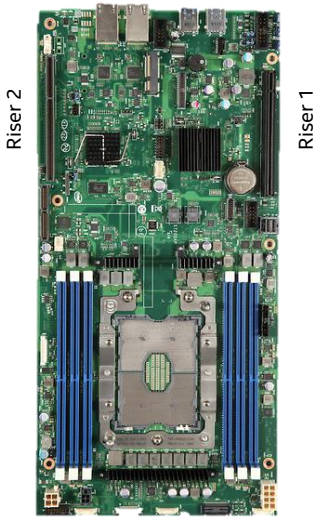
System Parameters: ~564 TF/s SU, 39 TB DRAM

- Intel KNL 68 C @ 3.05 TF/s nodes; 16 GB MCDRAM, 96-192 GB DDR4
  - 16/32 GB DDR4-2133 SDRAM
- Single Rail Omni Path1; 12.5+12.5 GB/s Bandwidth
- Built from 48-Port L1 and 288 or 1,152 Spine Omni Path switches
- Compute and gateway nodes. Remote boot from RPS nodes
- GW IO Bandwidth 10-20 GB/s delivered parallel I/O performance
- Software for build and acceptance Sandstone Peak V1
- May have SSD in GW nodes for Burst Buffer to accelerated checkpoint/restart capabilities

Intel® Server board S7200AP(Adams Pass)Product Family

Efficient, high-density solution with optimized memory performance for density optimized value in high performance computing (HPC) deployments.

- The Intel® Server Board S7200AP is specifically designed for parallelized workflows in the HPC market. It features support for Intel® Xeon® Phi™ processors (Bootable Knights Landing), with 6 DIMMs (1DPC) and optional support for Intel® Omni-Path Fabric Technology.
- Customizable as a 2U, four node system, it features easy serviceability and high availability, with hot-swappable compute modules, 2.5” or 3.5” drive bays, and redundant power supply modules.



Board Features

- **Intel® Xeon® Phi™ processor (Bootable Knights Landing)**
  - Up to 215W TDP support; 230W TDP for KNL-F
  - Intel® C-612 chipset: Intel Wellsburg Platform Controller Hub (PCH)
  - 4 ports to bridge board
  - 4 ports to miniSAS connector on motherboard
  - 1 port to mSATA connector on motherboard
- **Board Form factor: 6.8”W x 14.2”L**
- **6 x DDR4 DIMMs, 1SPC, 6 x native channels/system**
  - Supported speeds: 1866, 2133, 2400MT/s Registered/LRDIMM ECC
- **Manageability:**
  - Pilot 3 BMC with optional advanced features via RMM4-lite module;
- **KNL Integrated PCI-E Gen 3 I/O Configuration:**
  - Riser 1 – PCIe Gen3 x 16
  - Riser 2 – PCIe Gen3 x 20 (x16 or x4)
- **LOM: Ethernet**
  - 2x Intel i210 (Springville 1GbE) Controllers
- **External I/O**
  - (2) USB 3.0
- **Fabric Support**
  - Dual-port Intel® Omni-Path Fabric (StormLake) with KNL-F/QSFP Carrier in Riser 1
  - Intel® Omni-Path Low-profile PCIe Adapter
- **Chassis: 2U/4Node Bobcat Peak w/2130W PSU**

Intel Server board S7200AP (Adams Pass) in Node Tray

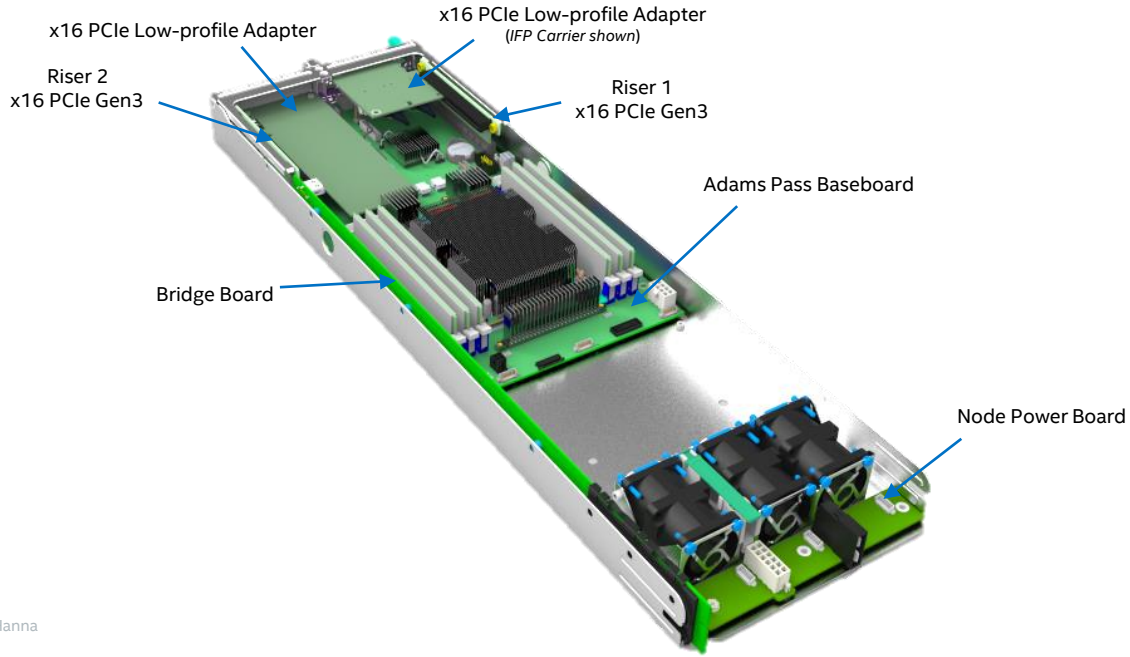


Image: Glen Hanna

Intel Server Chassis H2000XXLR2

Adams Pass in Bobcat Peak-G Chassis

Front View

Rear View

Features

- 4-Node System with Adams Pass Half-width Board
- (8) I/O PCIe x16 LP cards
- 16 x 2.5” (H2216XXLR2) or 12 x 3.5” (H2312XXLR2) SAS/SATA Hot-swap HDDs
- 2U x 30” Deep
- 2130W Redundant PSU

Images: Glen Hanna

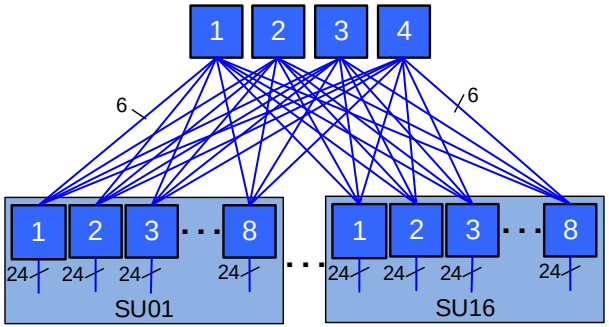
Assume 450W per node ➔ 1800W per chassis

KNL+STL 192 Node SU fits in 3 racks

552 TF/s SU with Single Rail STL1

- 8 groups of 24 nodes, each with a Omni Path 1 48 port switch and Ethernet management switch
- Standard Front2Back Air Cooled 42U rack with PDU and cable management
- Rack1,2 @ 32.4kW
- Rack 3 @ 21.6kW

Building a Beowulf cluster with High Radix 768 port Spine Switches in Fat Tree Topology



Two stage full bi-section bandwidth Fat-Tree with optical connections between Leaf and Spine Switches

- Compute, memory and IO bandwidth scales as more SU are added
- Can start small and add SU as needed (e.g., more microscopes added)

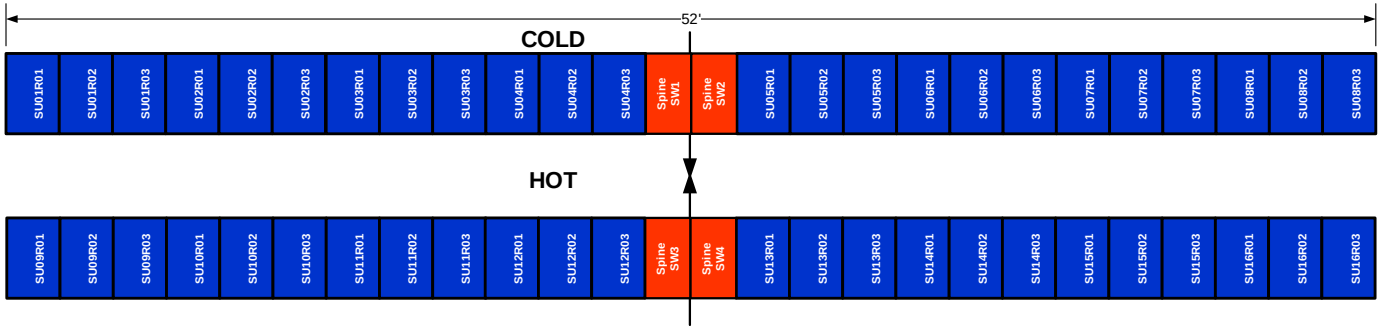
16 SU fully configures 4 spine switches

TOP50 16SU Cluster Floor Layout

System Parameters

Compute portion only

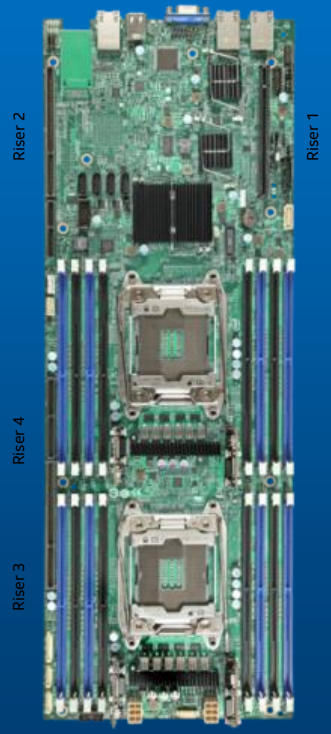
- 48 compute and 4 STL1 racks, 16 SU with 3 racks per SU
- Two row layout with 8 SU and 2 STL1 switch per row
- Does not include SAN nor Lustre MDS or OSS
- 52’ wide by 12’ depth with 2’x4’ standard 42U racks
- Estimate 1,385 kW system power



Intel® Server board S2600TP Product Family

A high-density, performance solution for Data Analytics, Storage and Cloud requiring large memory and maximum I/O capacity

- Bringing higher memory and I/O capacity together with the power and performance of dual Intel Xeon® Processors E5-2600 v3, the Intel® Server Board S2600TP features 16 DIMMs and 4 riser slots utilizing all 80 lanes of PCI Gen3 available in the chipset with optional FDR Infiniband and Intel® Remote Management Module support, making the S2600TP the half-width choice for demanding Data Analytics, Storage, and Cloud applications.

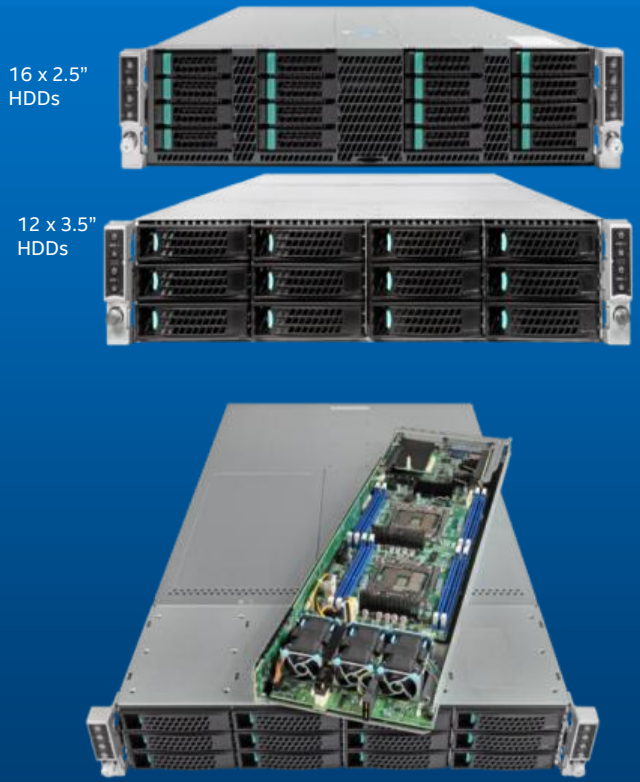


Board Features

- **Dual Intel® Xeon® processor E5-2600 v4: Broadwell-EP**
  - Up to 145W TDP support; **160W support for board only**
- **Intel® C-610 chipset: Wellsburg Server South Bridge**
  - 5 x SATA 6 Gbps ports (4 + 1 SATADOM) onboard, 5 x SATA 6Gbps ports (4 + 1 SATADOM) optional through bridge board
  - 2 USB 2.0 ports, 1 USB 2.0 port optional through bridge board
- **Board Form factor:**
  - 6.8"W x 18.9"L
- **FDR IB Connect IB x8 for IB sku only**
  - 1 QSFP+ port
- **16 DDR4 DIMMs, 2SPC, 8x native channels/system**
  - Haswell supported speeds: 1333, 1600, 1866 (2DPC), **2133** (1DPC) up to **2400MT/s**
  - **Memory VRDs Down**
  - **NVDIMM support**
- **Manageability:**
  - Pilot 3 BMC with optional advanced features via Intel® RMM4 Lite module; **onboard dedicated management NIC**
- **Haswell-EP Integrated PCIe* Gen 3 I/O Configuration:**
  - **Riser 1 – PCIe Gen3 x 16 from CPU1**
  - **Riser 2 – PCIe Gen3 x 24 (x 8+x 16) from CPU1**
    - **X 8 PCIe Gen3 for onboard IB for IB sku**
    - **X 16 PCIe Gen3 for IO Module need additional rIOM riser card**
  - **Riser 3 – PCIe Gen3 x 24 (x8+x16) from CPU2**
  - **Riser 4 – PCIe Gen3 x 16 slot from CPU2**
- **LOM: Ethernet w/ Intel® Virtualization Technology**
  - Dual 1GbE (Powerville)

Intel® Server Chassis H2000G Product Family

Efficient, high-density solution with optimized memory performance for preferred value in high performance computing (HPC) deployments



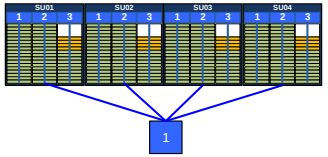
Chassis Features

- **Chassis Dimensions**
  - L x W x H = 28.86" x 17.24" x 3.460" (2.5" SKU)
  - L x W x H = 30.35" x 17.24" x 3.460" (3.5" SKU)
- **2U 4-Node System**
  - Kennedy Pass 8-DIMM, -EP
  - Taylor Pass 16-DIMM, -EP
  - Nodes are hot-swap and rear-accessible
- **Storage**
  - 16 x 2.5" 12 Gb/s SAS/SATA HDDs
  - 12 x 3.5" 12 Gb/s SAS/SATA HDDs
  - Hot-swap, cable-less HDD support
- **System Cooling**
  - (3) 40x56 mm dual-rotor fans/node
- **I/O Capability**
  - (1) IOM x16 and x8 PCIe Gen3 module
  - (1) Low-profile x16 PCIe Gen3 adapter
- **Power**
  - 1600 W redundant CRPS PSU
- **Independent Front Panel controls for each Individual Node**

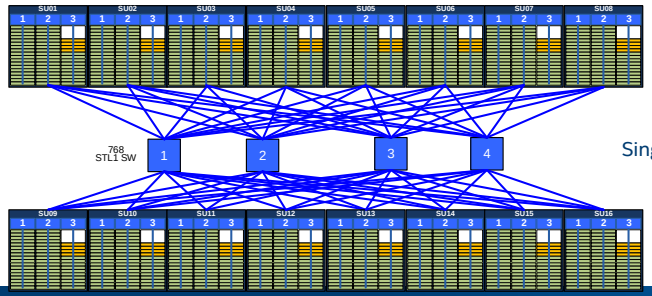
Flexibility of the Scalable Unit concept allows for a variety of scale up KNL+Xeon cluster sizes

SU	2nd SW	CN	KNL Cores@68	KNL Peak	KNL Linpack	KNL Mem (TiB)
1	0	188	12,784	564	338	39.00
4	1	752	51,136	2,256	1,354	156.00
8	2	1,504	102,272	4,512	2,707	312.00
12	3	2,256	153,408	6,768	4,061	468.00
14	4	2,632	178,976	7,896	4,738	546.00
16	4	3,008	204,544	9,024	5,414	624.00
18	5	3,384	230,112	10,152	6,091	702.00
20	5	3,760	255,680	11,280	6,768	780.00
24	6	4,512	306,816	13,536	8,122	936.00
32	8	6,016	409,088	18,048	10,829	1,248.00
48	12	9,024	613,632	27,072	16,243	1,872.00

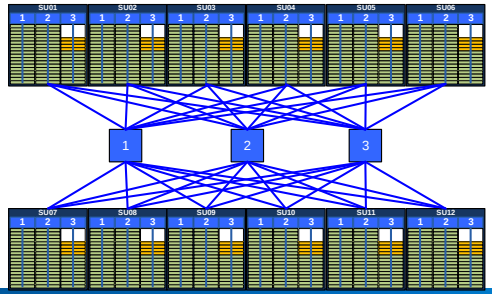
Would rank #10 on June 2015 TOP500 List, #50 projected for June 2017



1,385 kW system power @ 16 SU



Single Rail STL1

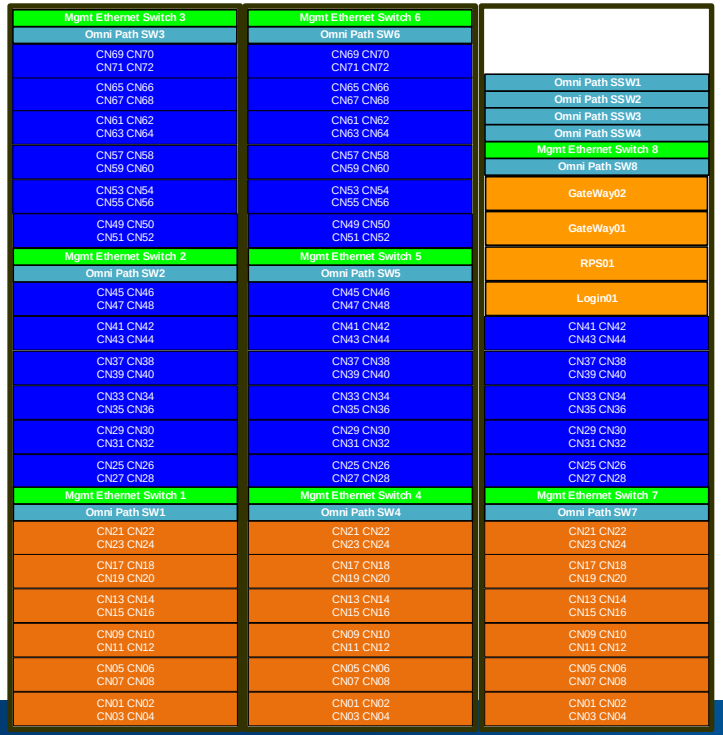


SU Concept can be scaled down as well

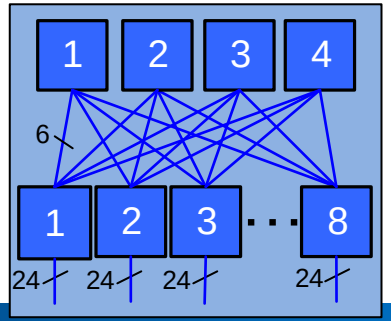
- The following show how to scale down the SU to one or two or racks and mix Xeon and Xeon Phi roughly 1:2 ratio
- 1 & 2 Rack design uses 1/10 Gb Eth SAN and repurposes 2 Xeon CN nodes as LN and RPS
- Direct STL1 cables between Leaf SW
- 3 Rack design uses 2U Xeon for LN, RPS and GW and 4x 48P SW as Spine SW

		1Rack	2Rack	3Rack
Xeon Phi		48	96	122
Node (TF/s)	3	144	288	366
MCDRAM (GB)	16	144	288	366
DDR4 (TB)	192	144	288	366
Xeon		22	46	66
Node (TF/s)	0.59	12.98	27.14	38.94
MCDRAM (GB)	0	0	0	0
DDR4 (TB)	512	11.26	23.55	33.79
Total				
SU (TF/s)		157	315.14	404.94
SU MCDRAM (GB)		144	288	366
SU DDR4 (TB)		155.3	311.55	399.79

1xSU design. 48P Spine SW. Use Eth for SAN and Xeon for LN and RPS nodes



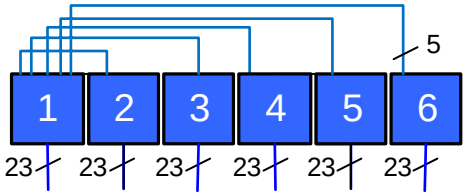
- Xeon Phi CN
  - 116 nodes
- Xeon CN
  - 72 nodes
- Xeon Service Nodes and SAN
  - 1 LN, RPS and 2 GW (2U1N)
  - STL SAN from GW



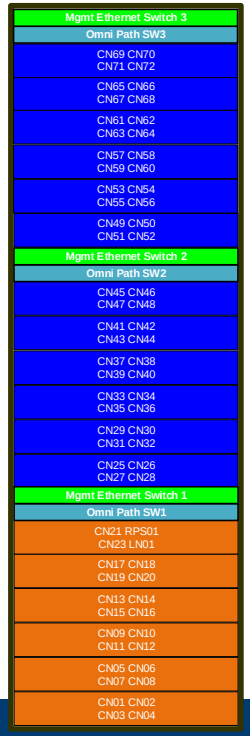
Two Rack SU. Use Eth for SAN and Xeon for LN and RPS nodes

Mgmt Ethernet Switch 3	Mgmt Ethernet Switch 6
Omni Path SW3	Omni Path SW6
CN69 CN70 CN71 CN72	CN69 CN70 CN71 CN72
CN65 CN66 CN67 CN68	CN65 CN66 CN67 CN68
CN61 CN62 CN63 CN64	CN61 CN62 CN63 CN64
CN57 CN58 CN59 CN60	CN57 CN58 CN59 CN60
CN53 CN54 CN55 CN56	CN53 CN54 CN55 CN56
CN49 CN50 CN51 CN52	CN49 CN50 CN51 CN52
Mgmt Ethernet Switch 2	**Mgmt Ethernet Switch 5**
Omni Path SW2	Omni Path SW5
CN45 CN46 CN47 CN48	CN45 CN46 CN47 CN48
CN41 CN42 CN43 CN44	CN41 CN42 CN43 CN44
CN37 CN38 CN39 CN40	CN37 CN38 CN39 CN40
CN33 CN34 CN35 CN36	CN33 CN34 CN35 CN36
CN29 CN30 CN31 CN32	CN29 CN30 CN31 CN32
CN25 CN26 CN27 CN28	CN25 CN26 CN27 CN28
Mgmt Ethernet Switch 1	**Mgmt Ethernet Switch 4**
Omni Path SW1	Omni Path SW4
CN21 RPS01 CN23 LN01	CN21 CN22 CN23 CN24
CN17 CN18 CN19 CN20	CN17 CN18 CN19 CN20
CN13 CN14 CN15 CN16	CN13 CN14 CN15 CN16
CN09 CN10 CN11 CN12	CN09 CN10 CN11 CN12
CN05 CN06 CN07 CN08	CN05 CN06 CN07 CN08
CN01 CN02 CN03 CN04	CN01 CN02 CN03 CN04

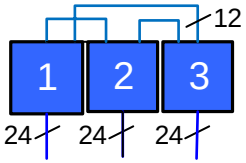
- Xeon Phi CN
  - 92 nodes
- Xeon CN
  - 44 nodes
- Xeon Service Nodes and SAN
  - 1 LN and RPS (in 4U4N)
  - 1/10 GbE Ethernet from every node



One Rack SU. Use Eth for SAN and Xeon for LN and RPS nodes



- Xeon Phi CN
  - 48 nodes
- Xeon CN
  - 22 nodes
- Xeon Service Nodes and SAN
  - 1 Xeon LN and RPS (in 2U4N)
  - 1/10 GbE Ethernet from every node



System Software Ecosystem

OpenHPC & HPC Orchestrator

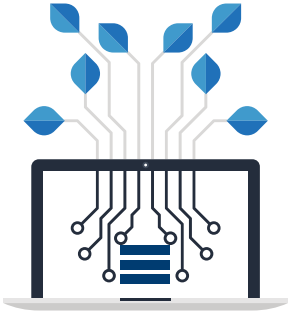
CURRENT STATE OF SYSTEM SOFTWARE EFFORTS IN HPC ECOSYSTEM

Fragmented efforts across the ecosystem – “Everyone building their own solution.”

With system margins under pressure, unwillingness to invest in system software

A desire to get exascale performance & speed up software adoption of HW innovation

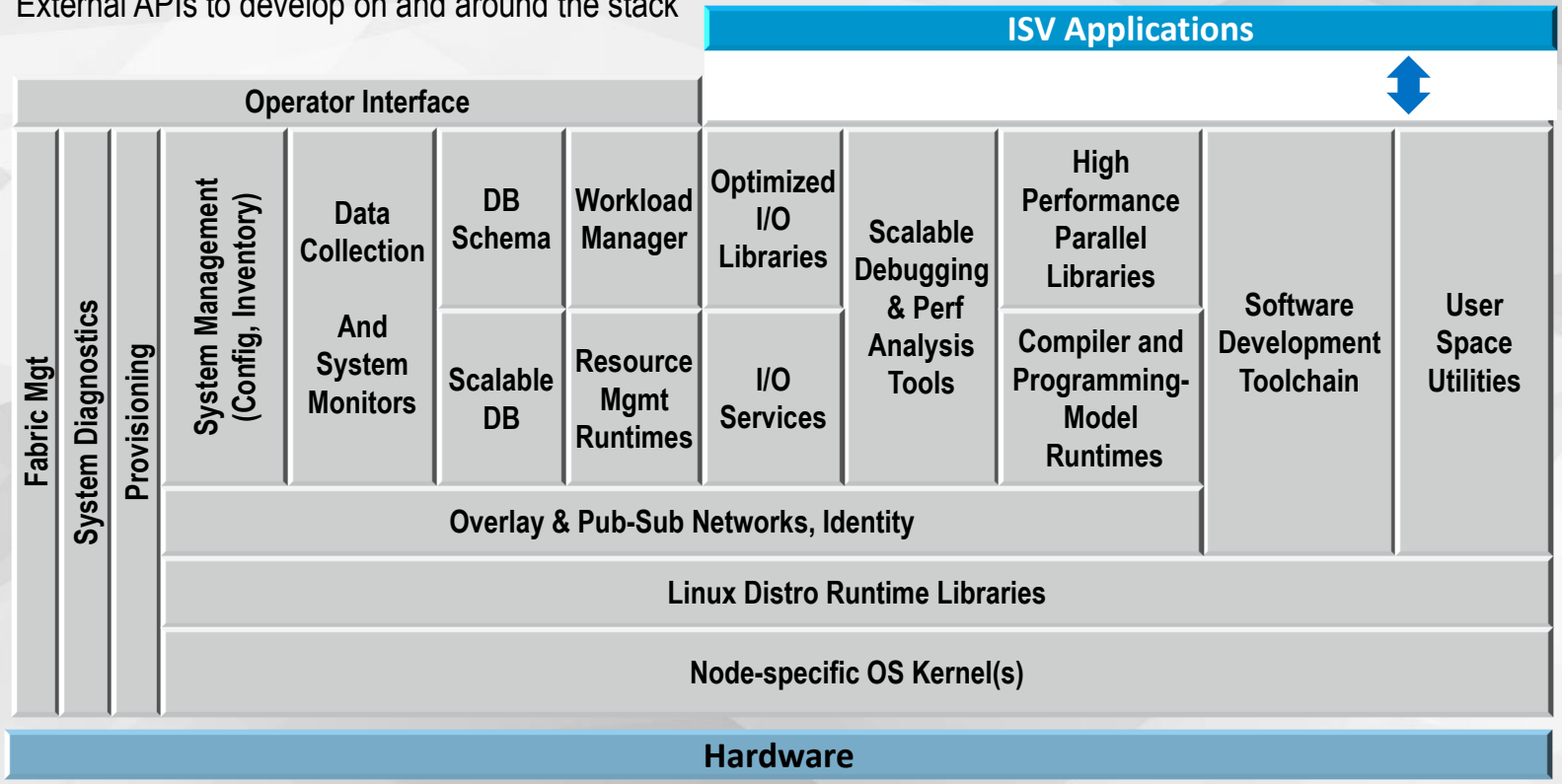
New complex workloads (ML, Big Data, etc) drive more complexity into the software stack



THE REALITY: We will not be able to get where we want to go without a major change in system software development

HPC STACK COMPONENT VIEW

- Intra-stack APIs to allow for customization/differentiation
- External APIs to develop on and around the stack



DESIRED FUTURE STATE

- **Stable HPC Platform Software that:**
  - Fuels a vibrant and efficient HPC software ecosystem
  - Takes advantage of hardware innovation & drives revolutionary technologies
  - Eases traditional HPC application development and testing at scale
  - Extends to new workloads (ML, analytics, big data)
  - Accommodates new environments (i.e. cloud)

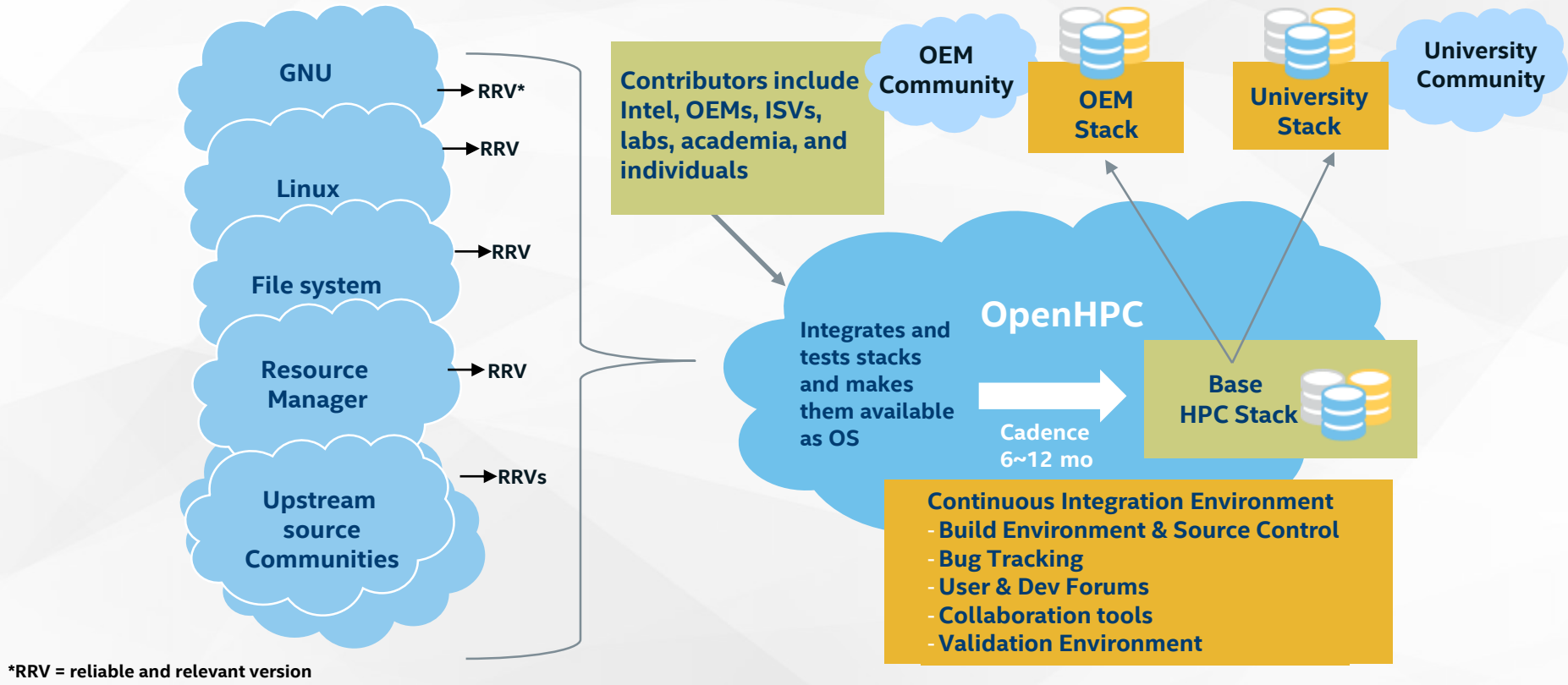
PARTICIPANTS AS OF 3/31/2016

OEMs

Users

ISV-OSV

OPENHPC COMMUNITY WORKFLOW



*RRV = reliable and relevant version

OpenHPC++ - Potential future efforts

Functional Areas	Components	Contributions by:
Base OS	CentOS 7.2, SLES12SP1, McKernel, Kitten, mOS	RIKEN, Sandia, Intel
AdminTools	Conman, Ganglia, Intel Cluster Checker**, Lmod, LosF, Nagios, pdsh, prun, EasyBuild, Spack, genders, mrsh, clustershell, ORCM	Intel
Provisioning	Warewulf, xCAT	Community
Resource Mgmt.	SLURM, Munge, ParaStation mgmt, PMIx, PBS Pro	ParTec, community, Altair
Cross Cutting	OpenStack HPC suitable components	Cray
Runtimes	OpenMP, OCR, OmpSs	BSC
I/O Services	Lustre client (community version), shine, Lustre server	Community
Numerical/SciLib	Boost, GSL, FFTW, Metis, PETSc, Trilinos, Hypre, SuperLU, Mumps, Intel MKL**	
I/O Libraries	HDF5 (pHDF5), NetCDF (including C++ and Fortran interfaces), Adios	
Compiler Families	GNU (gcc, g++, gfortran), Intel Parallel Studio XE (icc,icpc,ifort)**	
MPI Families	MVAPICH2, OpenMPI, Intel MPI**, MPICH, ParaStation MPI	Argonne, ParTec
Development Tools	Autotools (autoconf, automake, libtool), Valgrind,R, SciPy/NumPy, Intel Inspector **	
Performance Tools	PAPI, Intel IMB, mpiP, pdtoolkit TAU, Intel Advisor**, Intel Trace Analyzer and Collector**, Intel Vtune Amplifier**, Paraver, Scalasca, Cube	BSC, Jülich

** Bring your own license model

OPENHPC TO INTEL HPC SW STACK TO SSF

openHPC

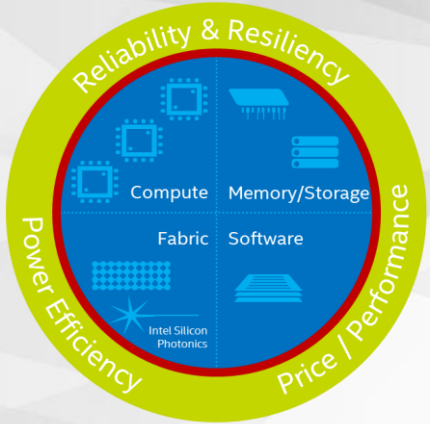
An open source community for HPC software

Intel seeded the community with pre-integrated, pre-tested and validated HPC software stack & will continue contributions along with other members of the community

Intel will offer Intel-supported products based on the open source stacks

Performance Peak Family of Products

- *Turnkey*
- *Advanced*
- *Custom*



Intel products are the realization of the software portion of SSF

Intel's HPC Scalable System Framework

Intel® HPC Orchestrator

openHPC

Open source solution for HPC

Community led organization

World-wide participation

39 members since launch
(including non-IA based vendors)

Over 11,000 new worldwide visitors since launch

Version 1.1 now available

NEW

Intel® HPC Orchestrator

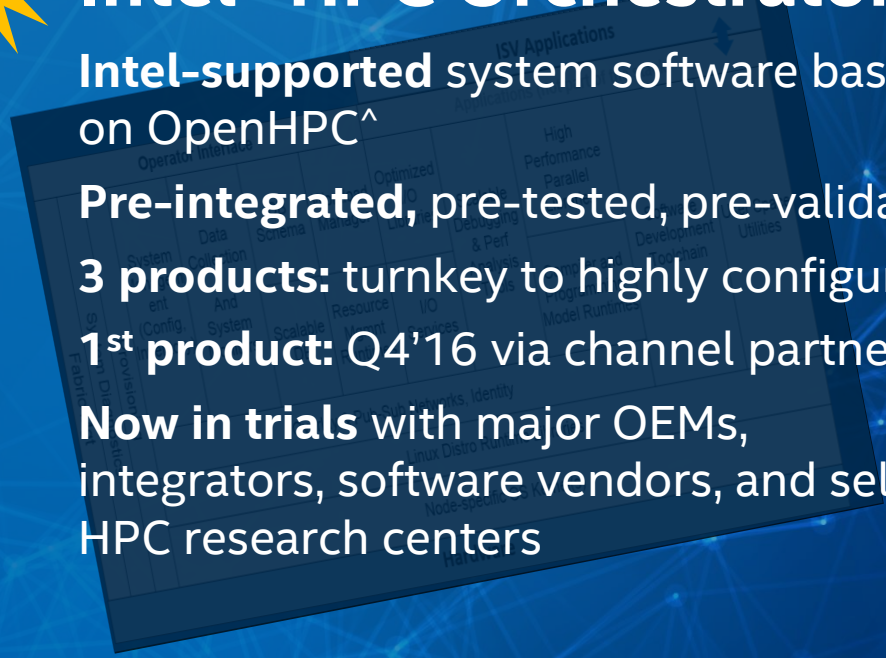
Intel-supported system software based on OpenHPC^

Pre-integrated, pre-tested, pre-validated

3 products: turnkey to highly configurable

1st product: Q4'16 via channel partners

Now in trials with major OEMs, integrators, software vendors, and select HPC research centers



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

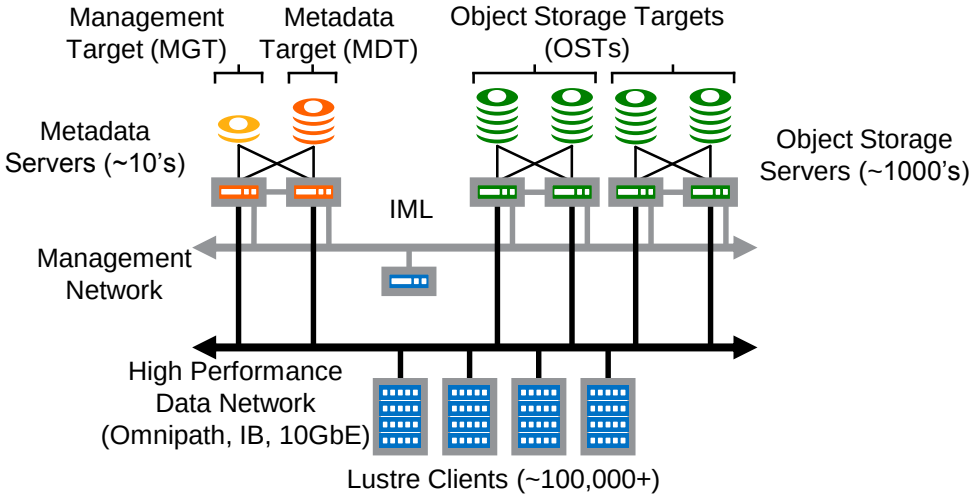
^ OpenHPC is an **open source HPC** community. For more information please visit: http://www.openhpc.community/

www.intel.com/hpcorchestrator

Storage Scalable Units for building Lustre parallel file systems

Lustre Architecture

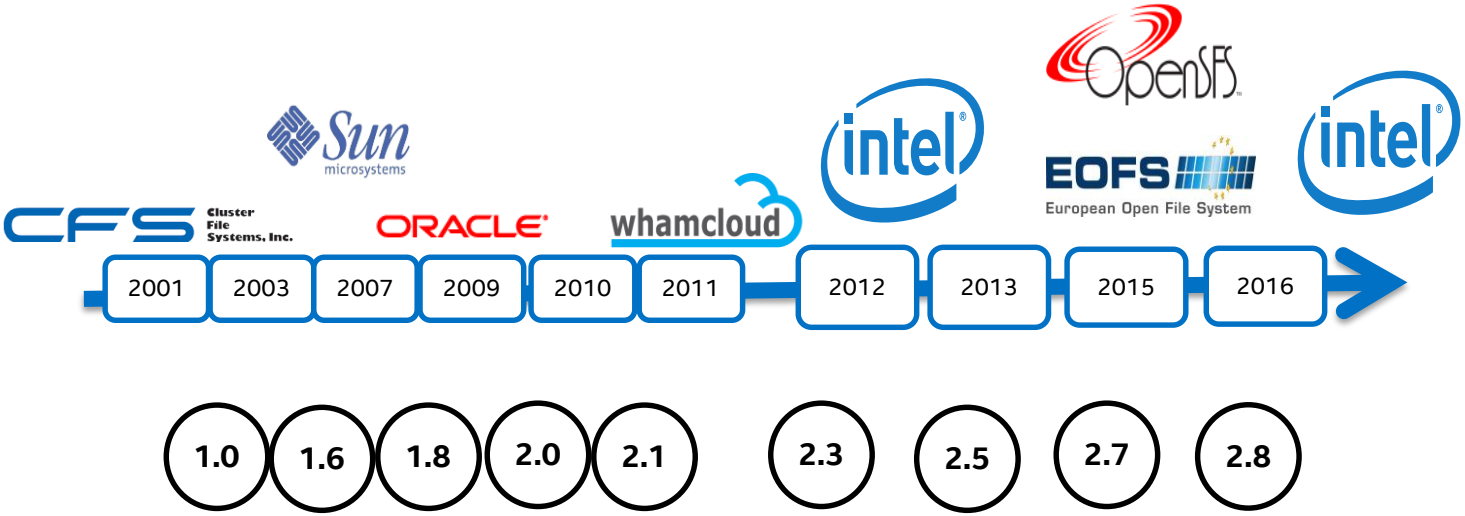
- Lustre software has four main components
  - Client – Runs on compute nodes and provides POSIX file system view of distributed storage resources
  - Metadata Target – Runs on MDS and provides mapping from POSIX name space to objects on OST
  - Object Storage Target – Runs on OSS provides scale out object storage
  - LNET Router – Runs on SU gateway (not shown)
- Lustre management infrastructure includes
  - Management Target – Runs on MDS passive fail over node
  - Intel Manager for Lustre – GUI interface and API for managing and monitoring Lustre environment



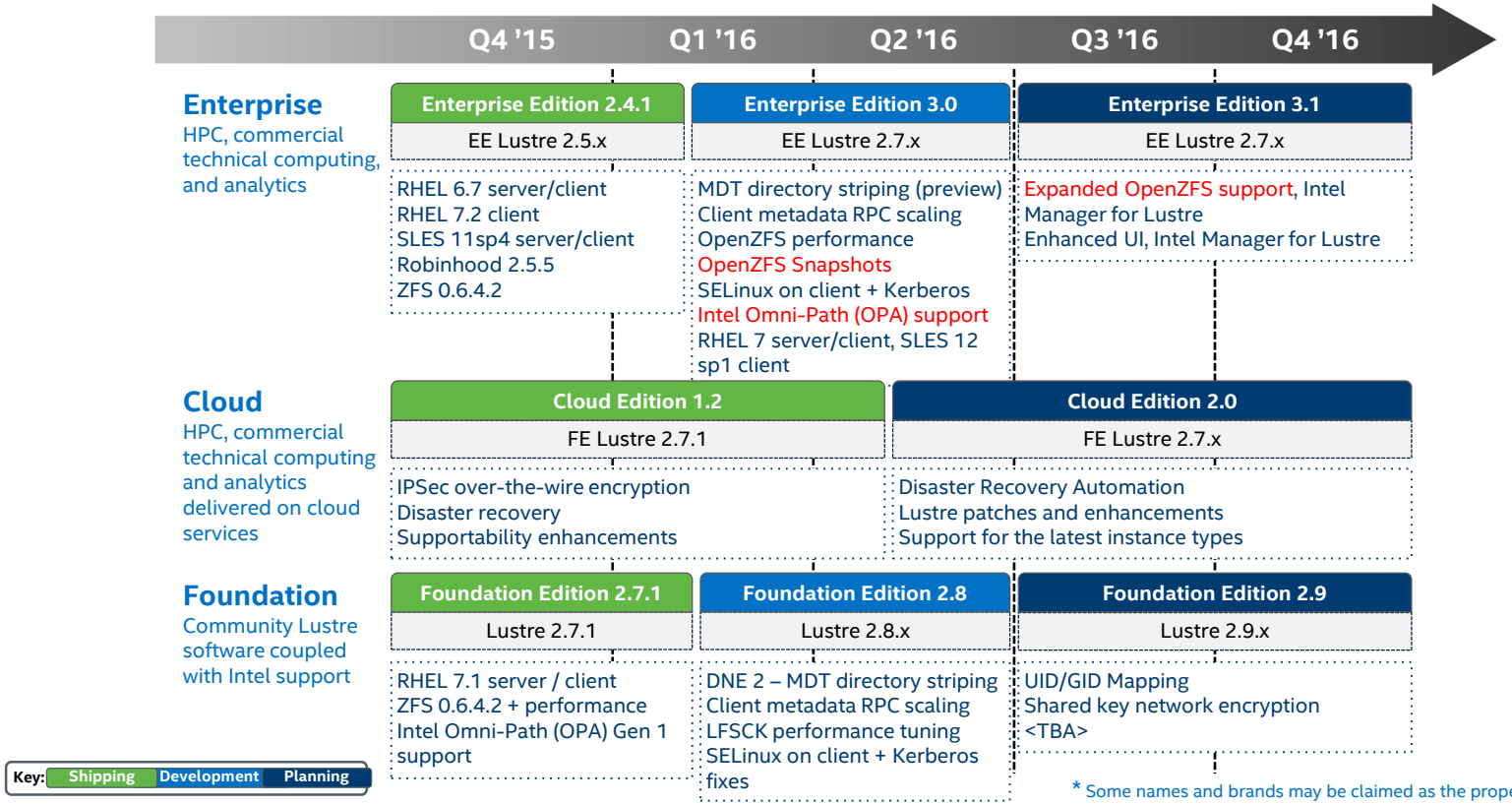
Lustre* is the leading scale-out parallel posix/file solution for High Performance Computing.

A Brief History of Lustre*

(WHO DOES THE RELEASES)



Intel® Solutions for Lustre* Roadmap

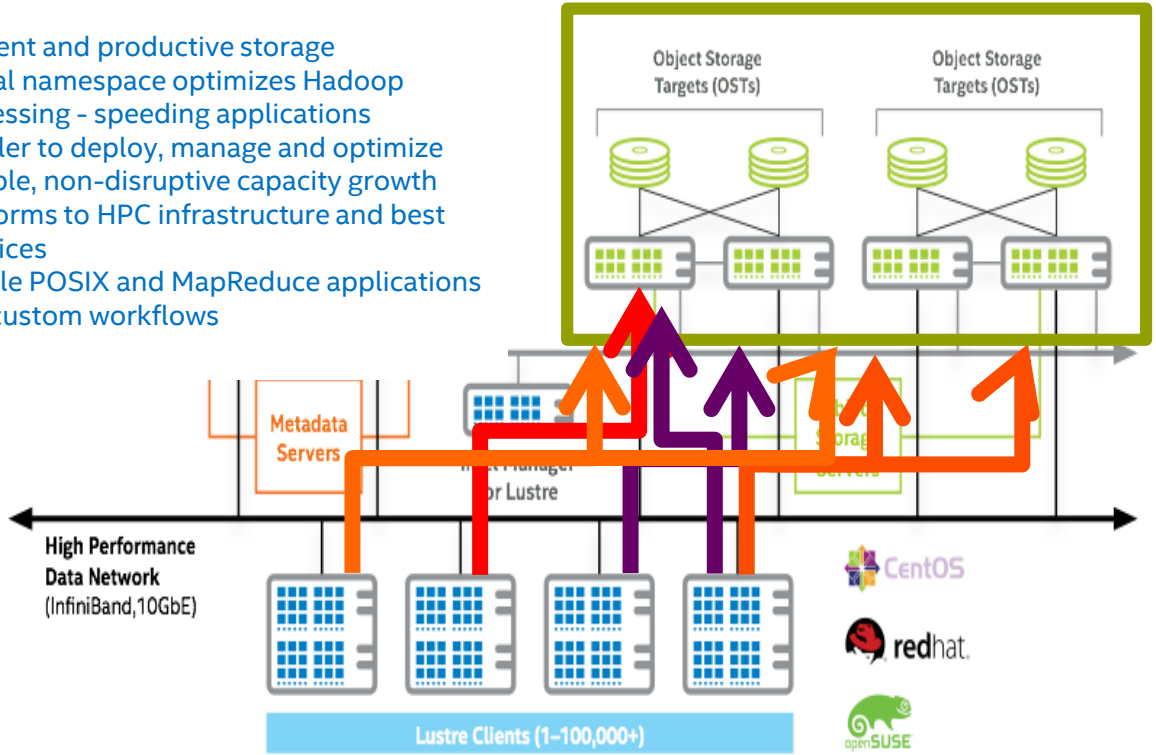


* Some names and brands may be claimed as the property of others.

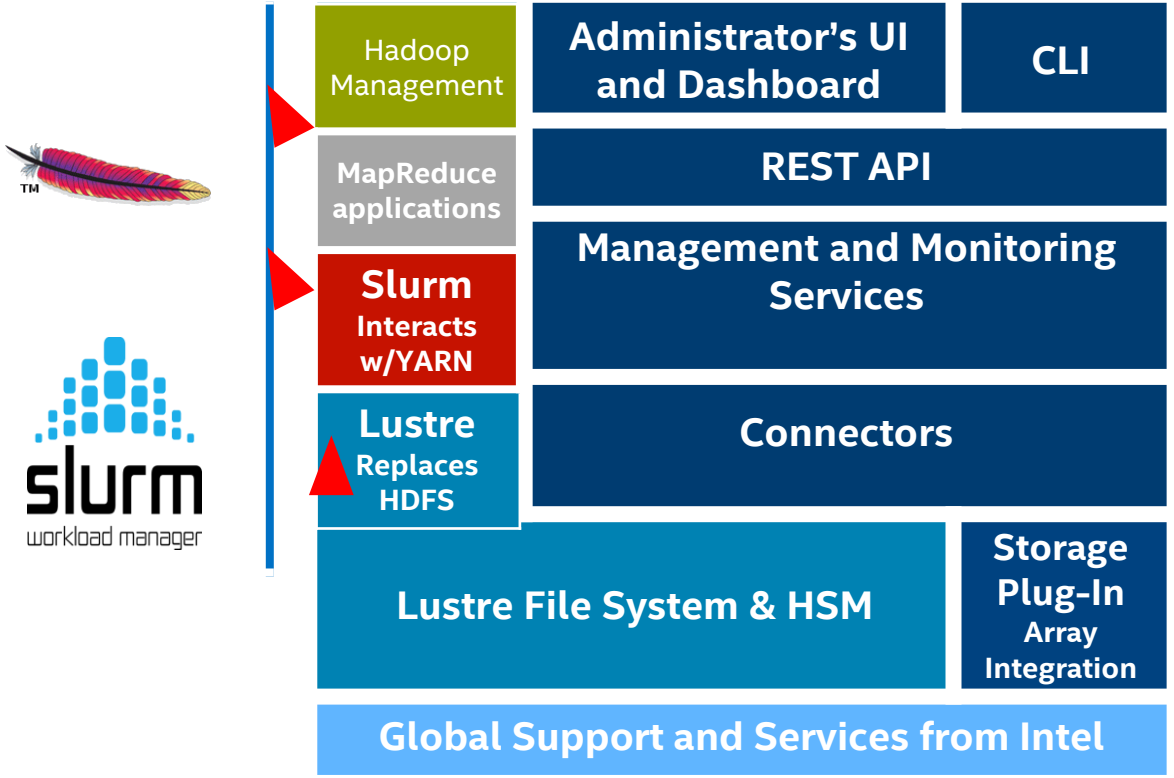
Product placement not representative of final launch date within the specified quarter

Powerful Benefits for Hadoop on HPC Resources

1. Efficient and productive storage
2. Global namespace optimizes Hadoop processing - speeding applications
3. Simpler to deploy, manage and optimize
4. Flexible, non-disruptive capacity growth
5. Conforms to HPC infrastructure and best practices
6. Couple POSIX and MapReduce applications into custom workflows

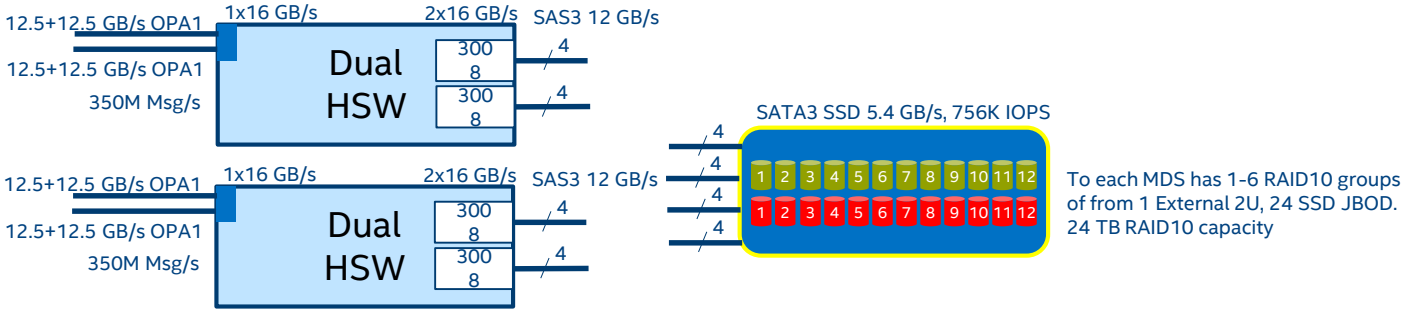


Integrated Software Stack for MapReduce



Next Generation Lustre/OpenZFS MDS with RAIDZ and JBOS

Assumptions for 1Q2016 config:
Use Dual/Single Xeon E5-2667v3 (3.2GHz, 8 Core, 20MB LLC)
SATA3 2.5” 2.0 TB, SSD 450 MB/s and 63K IOPS
2U External 24 Drive JBOD enclosure
PCIe Gen 3 channel throughput is 1GB/s
SAS3 8-port RAID10 controller throughput is 4.8 GB/s
Omni Path 1 = 2(12.5+12.5 GB/s, 10+10 GB/s), 350M Msg/s delivered



To each MDS has 1-6 RAID10 groups of from 1 External 2U, 24 SSD JBOD. 24 TB RAID10 capacity

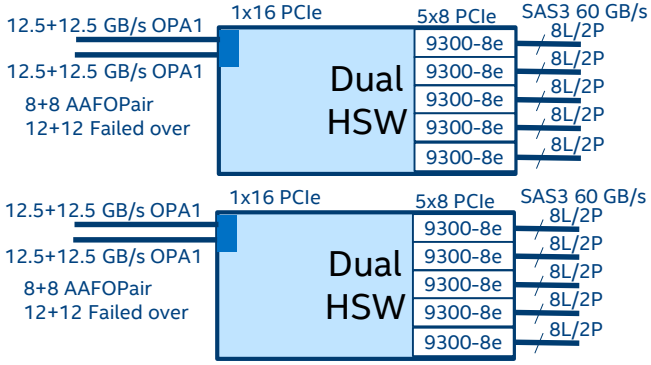
700M Msgs/s OPA1, 5.4 GB/s, 756K IOPs IO bandwidth, 24TB Capacity in RAID10

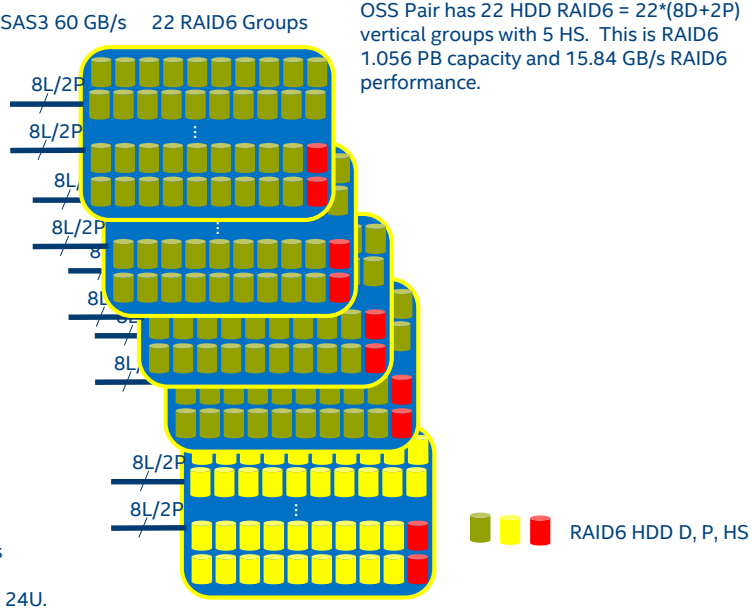
MDS SSU = 1 Fail over pair 4U + 2U JOBD Chassis = 6U.

Price estimate = $10K+3K+24*1K+150) = $37K

Next Generatioin Lustre/OpenZFS OSS with RAIDZ and JBOD

Assumptions for 1Q2016 config:
Use Dual Xeon E5-2667v3 (3.2GHz, 8 Core, 20MB LLC)
HDD throughput 45MB/s
PCIe Gen 3 channel throughput is 0.8 GB/s
Use Avago/LSI SAS 9300-8e dual port PCIe x8 HBA > 6.4 GB/s
Omni Path 1 = 2(12.5+12.5 GB/s, 10+10 GB/s) delivered





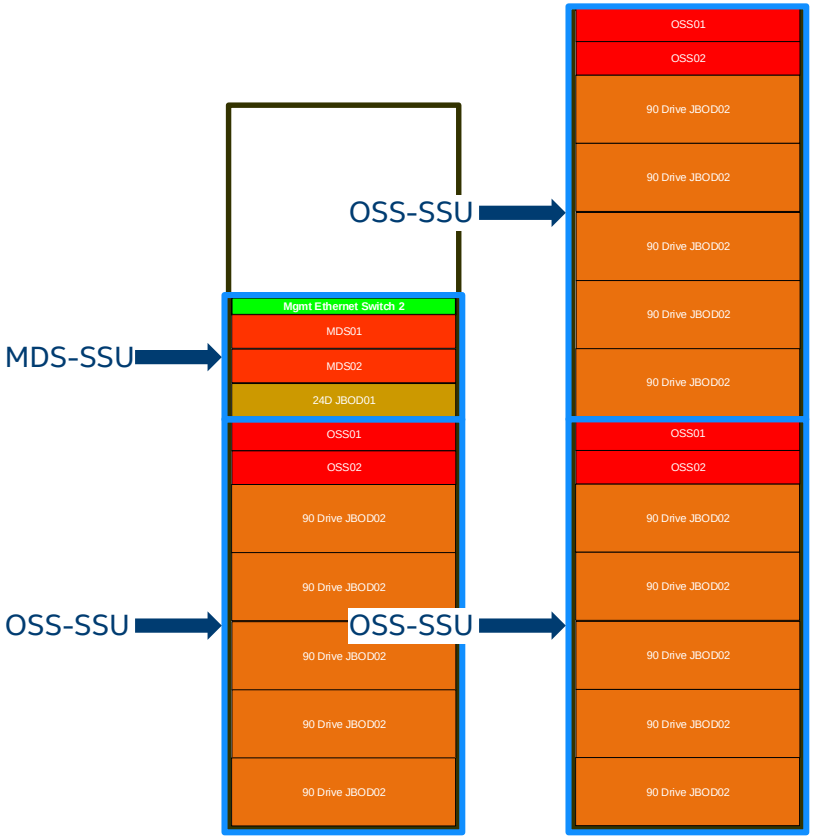
To each OSS 11 groups of HDD pairs from each of 5 JBOD in 11*(8D+2P)+5HS vertical RAID6 groups.

OSS Pair has 22 HDD RAID6 = 22*(8D+2P) vertical groups with 5 HS. This is RAID6 1.056 PB capacity and 15.84 GB/s RAID6 performance.

8+8 GB/s Lustre delivered bandwidth
1.056 PB RAID6 (8+2P) capacity with 6TB SAS3 Drives

OSS SSU = 1 Fail over pair 4U + 5x4U JOBD Chassis = 24U.

Two SSU from MDS-SSU and one or more OSS-SSU



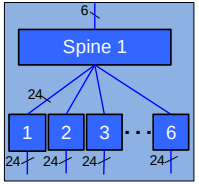
Single MDS-SSU can support multiple OSS-SSU

- MDS-SSU = 6U

OSS-SSU = 24U, 1.056 PB and 8 GB/s delivered Lustre Bandwidth

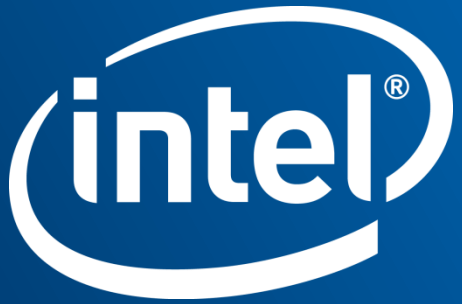
Second rack with 2OSS-SSU adds 2.1 PB and 16 GB/s bw

Combined Compute and IO Rack SU design with 7U 192 port Spine SW



CN Racks	KNL	BDW
1	48	24
2	96	48
3	144	72
4	192	96
8	384	192

- Eliminates SAN and GW nodes by connecting 6 IO Rack nodes directly to Spine SW
  - 10/40 GbE SAN can be supported by repurposing RPS
- Scales from 1 CN rack to 4 CN racks
  - Spine switch can incrementally add 32 ports up to 192 ports
  - 1 & 2 CN Racks full bisection BW
  - Bisection BW for 3 & 4 racks is 83% & 62.5%
- Can scale 5-8 CN racks by adding 2nd IO Rack
- Suitable for DTS, prototype and collaborator systems


intel®