Architetture innovative per il calcolo e l'analisi dati

Marco Briscolini

Workshop di CCR
La Biodola, Maggio 16-20. 2016

Agenda

- Cooling the infrastructure
- Managing the infrastructure
- Software stack

Target Segments - Key Requirements

Cloud Computing

Key Requirements:
- Mid-high bin EP processors
- Lots of memory (>256GB/node) for virtualization
- 1Gb / 10Gb Ethernet
- 1-2 SS drives for boot

Data Analytics

Key Requirements:
- Mid-high bin EP processors
- Lots of memory (>256GB per node)
- 1Gb / 10Gb Ethernet
- 1-2 SS drives for boot

High Performance Computing

Key Requirements:
- High bin EP processors for maximum performance
- High performing memory
- Infiniband
- 4 HDD capacity
- GPU support

Data Center Infrastructure

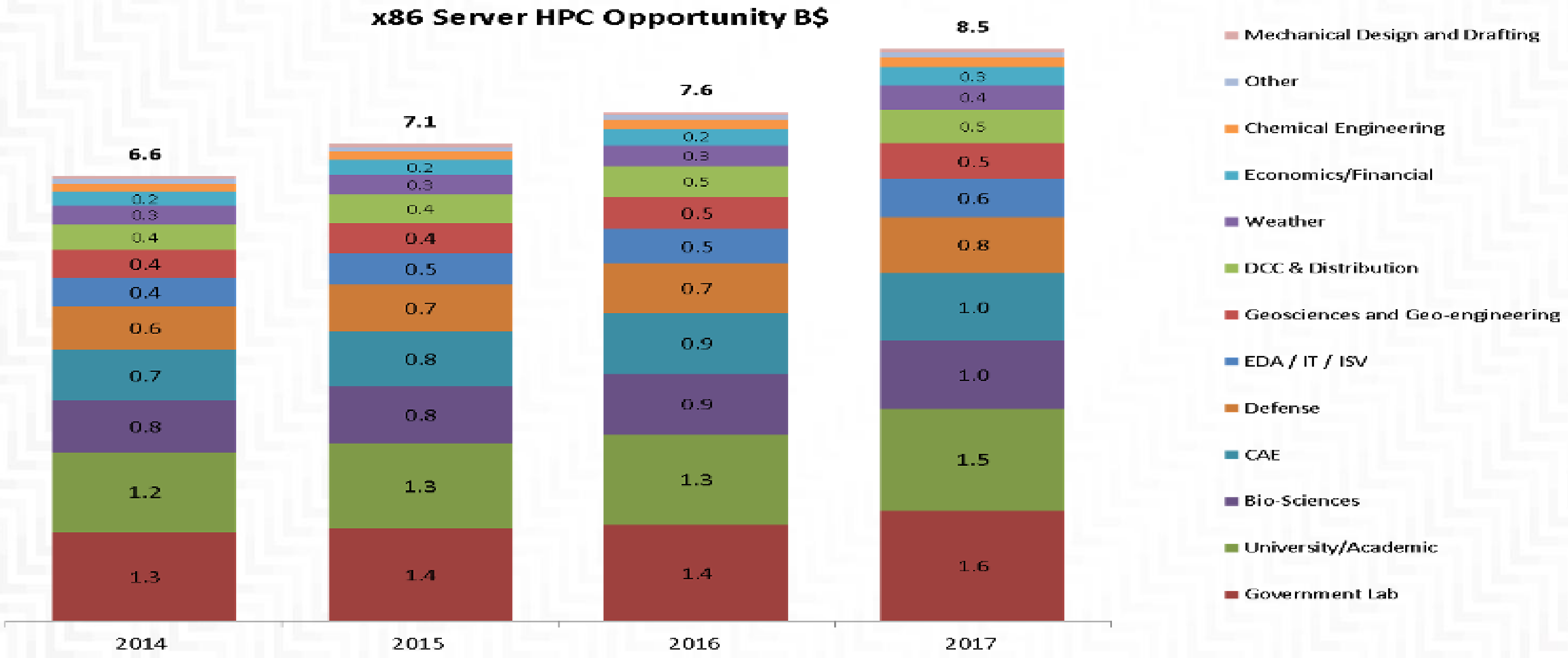
Key Requirements:
- Low-bin processors (low cost)
- Smaller memory (low cost)
- 1Gb Ethernet
- 2 Hot Swap drives(reliability)

Virtual Desktop

Key Requirements:
- Lots of memory (> 256GB per node) for virtualization
- GPU support

A LOOK AT THE X86 MARKET BY USE

- HPC is ~6.6 B$ growing 8% annually thru 2017



Some recent "HPC" EMEA Lenovo 2015 Client wins

Lenovo

This year wins and installations in Europe

1152 nx360M5 DWC
EDR

1512 nx360M5 BRW
3600 Xeon KNL
GPFS, GSS 10 PB
OPA

252 nx360M5 nodes
IB FDR14, Fat Tree
GPFS, 150 TB, 3 GB/s

392 nx360M5 nodes
IB FDR14 3D Torus
GPFS

312 nx360M5 DWC
6 nx360M5 with GPU
GPFS, IB FDR14, 2 GSS24

36 nx360M5 nodes
1 x 3850X6
IB FDR, GPFS

2 X 3 PFlops SuperMUC systems at LRZ Phase 1 and Phase 2

Phase 1

- Fastest Computer in Europe on Top 500, June 2012
  - 9324 Nodes with 2 Intel Sandy Bridge EP CPUs
  - HPL = 2.9 PetaFLOP/s
  - Infiniband FDR10 Interconnect
  - Large File Space for multiple purpose
    - 10 PetaByte File Space based on IBM GPFS with 200GigaByte/s I/O bw
- Innovative Technology for Energy Effective Computing
  - Hot Water Cooling
  - Energy Aware Scheduling
- Most Energy Efficient high End HPC System
  - PUE 1.1
  - Total Power consumption over 5 years to be reduced by ~ 37% from 27.6 M€ to 17.4 M€

Ranked 20 and 21 in Top500 June 2015

Phase 2

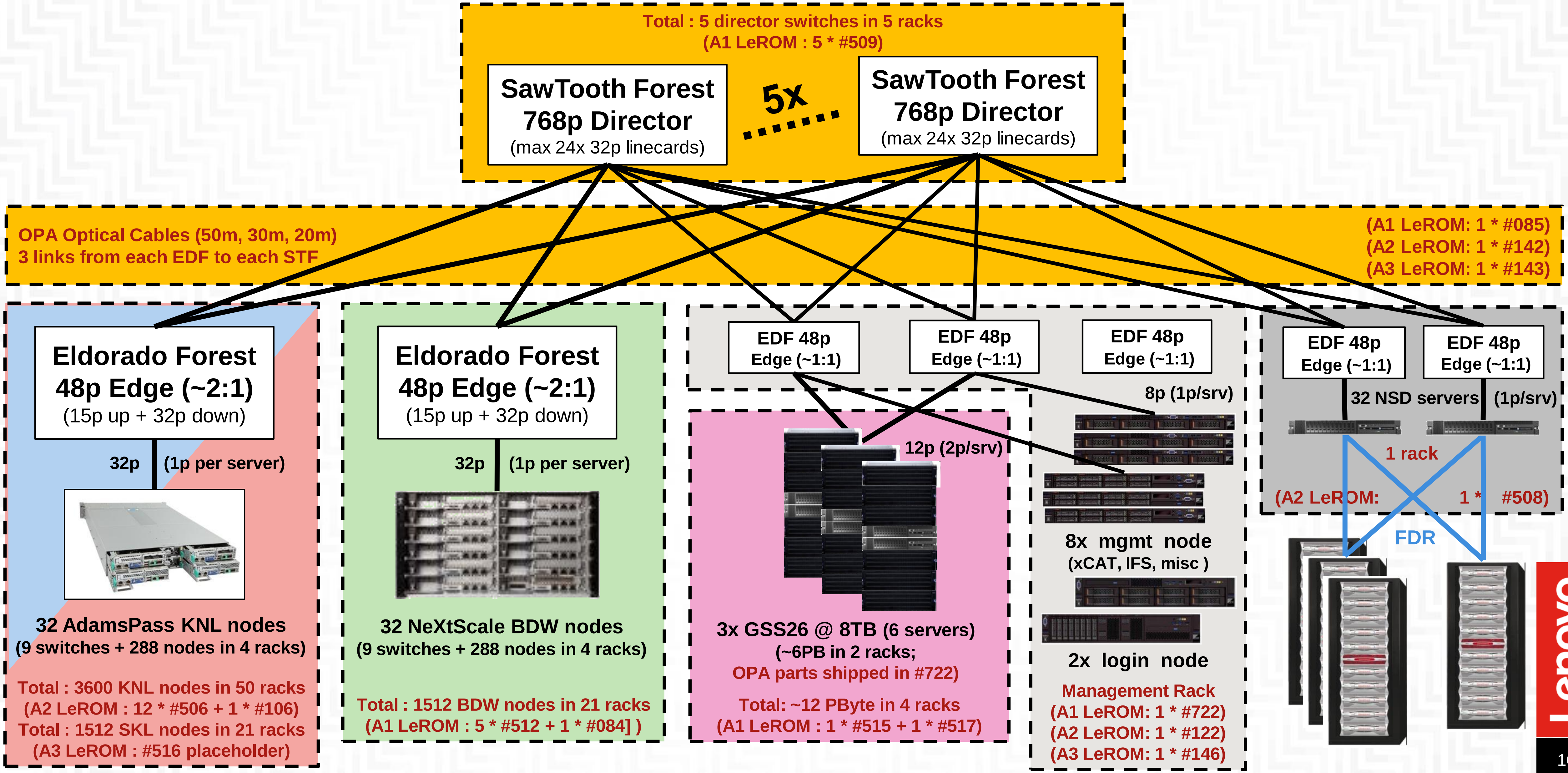
- Acceptance completed
  - 3096 nx360m5 compute nodes Haswell EP CPUs
  - HPL = 2.8 PetaFLOP/s
  - Direct Hot Water Cooled, Energy Aware Scheduling
  - Infiniband FDR14
  - GPFS, 10 x GSS26, 7.5 PB capacity , 100 GB/s IO bw

- System A:
- 1512 Lenovo nx360M5 (2 Petaflops)
  - 21 racks
  - 126 NeXtScale WCT Chassis
  - 3,024 Intel Broadwell-EP E5-2697v4 (2.3GHz, 145W)
  - 54.432Processor Cores
  - 12.096 16GB DIMMs
- 3600 Adamspass KNL nodes (11 Petaflops)
  - 50 Racks with 72 KNL nodes in Each Rack
  - 3.600 120GB SSD‘s
  - 244.800 cores
  - 345.600 GB RAM in 21.600 16GB DIMMs
  - 1.680 Optical cables
- 1512 Stark nodes (>4 Petaflops)
  - 21 racks
  - 3,024 Intel SkyLake
- Over 60.000m Optical Cables
- 6 GSS26 16PB raw in total
  - >100GB/s

BRW vs. KNL vs. SKL (based on Cineca)

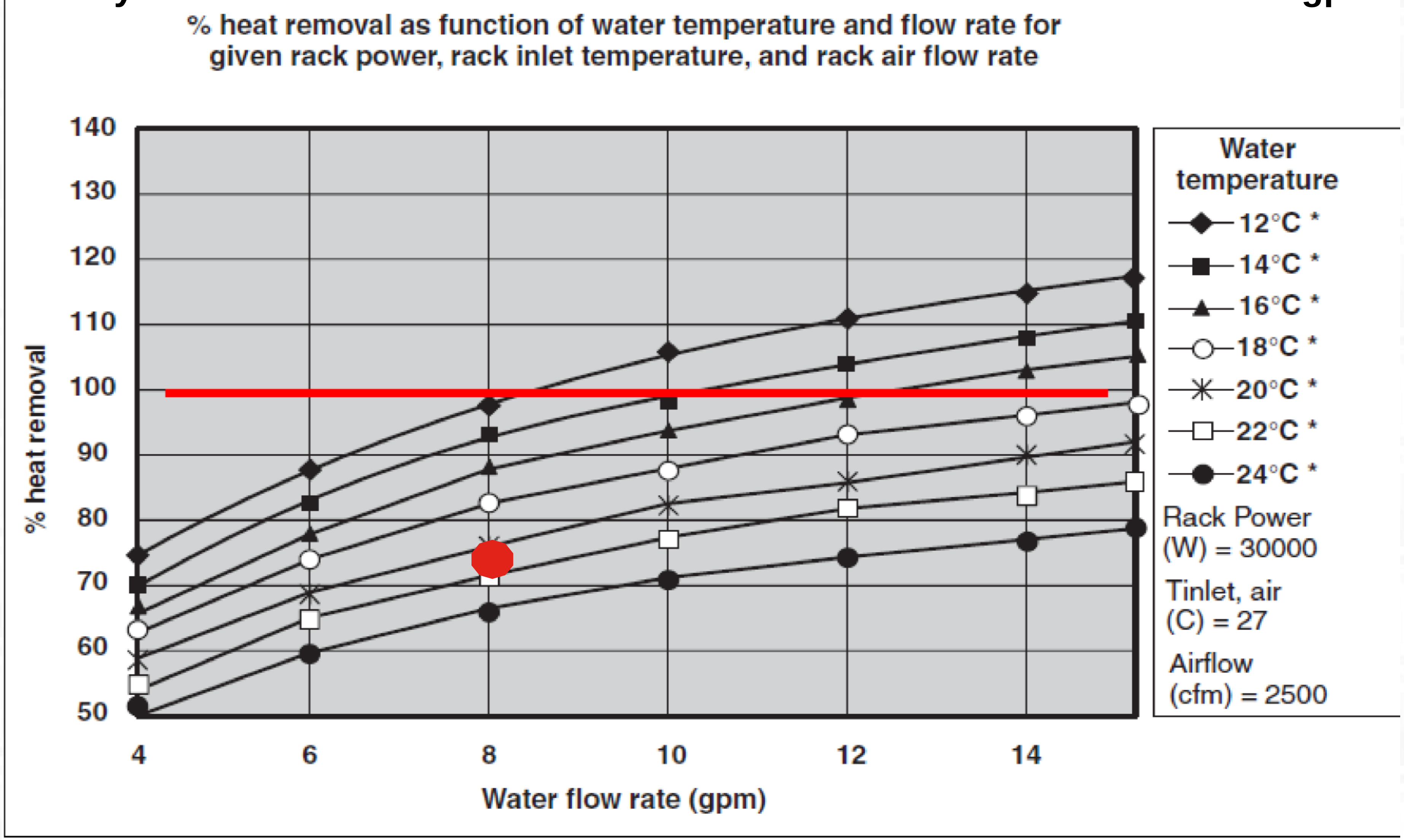
	BRW (2PFL)	KNL (11PFL)	SKL (4PFL)
Nodes	1512	3600	1512
CPU/node	2	1	2
TFlop/node	1.3	3	2.6 - 3.2
Price/node			
CPU	E5-2697v4	Bin1 (68c 1.4GHz)	SKP1
TFlop/Socket	0.65	3	1.3 - 1.6

CINECA – OMNI-PATH FABRIC ARCHITECTURE (SINGLE FABRIC, WITH 32:15 BLOCKING)



Power cooling using RDHX

Power cooling with hybrid W+A solution: Tinlet air 25°C and water on RDHX at 20°C and 8gpm



PPUE ESTIMATE

$$pPUE = \frac{IT + Cooling}{IT}$$

- **IT** = 1270Kw x 8760hrs = **11.125.200 kWhe**
- **Cooling =** 1.600.000 + 578.000 + 280.320 = **2.458.320 kWhe**
- $$pPUE = \frac{11.125.200 + 2.458.320}{11.125.200} = 1{,}22 \text{ annual average}$$

SOLUTION COMPONENTS

Think even bigger

NeXtScale System

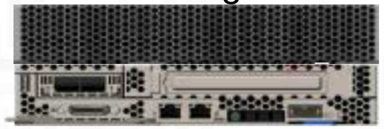
NEXTSCALE

Modular, high-performance system for scale-out computing

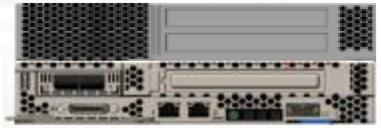
Chassis

Compute

Storage



Acceleration



Water Cooled Node

Primary Workloads

High Performance Computing

Public Cloud

Private Cloud

Scale Out Enterprise

The University of Wyoming
University of Michigan
Yale University
Honda
Tupolev Joint-Stock Company
J.P.Morgan
CINECA
LRZ
Skoltech Skolkovo Institute of Science and Technology
VISA
University of Victoria
Max-Planck-Gesellschaft
DTU Technical University of Denmark
MUREX
Cablevision
STFC Hartree Centre

Standard Rack

Lenovo

NeXtScale nx360 BDW Compute Node

For broadest workloads

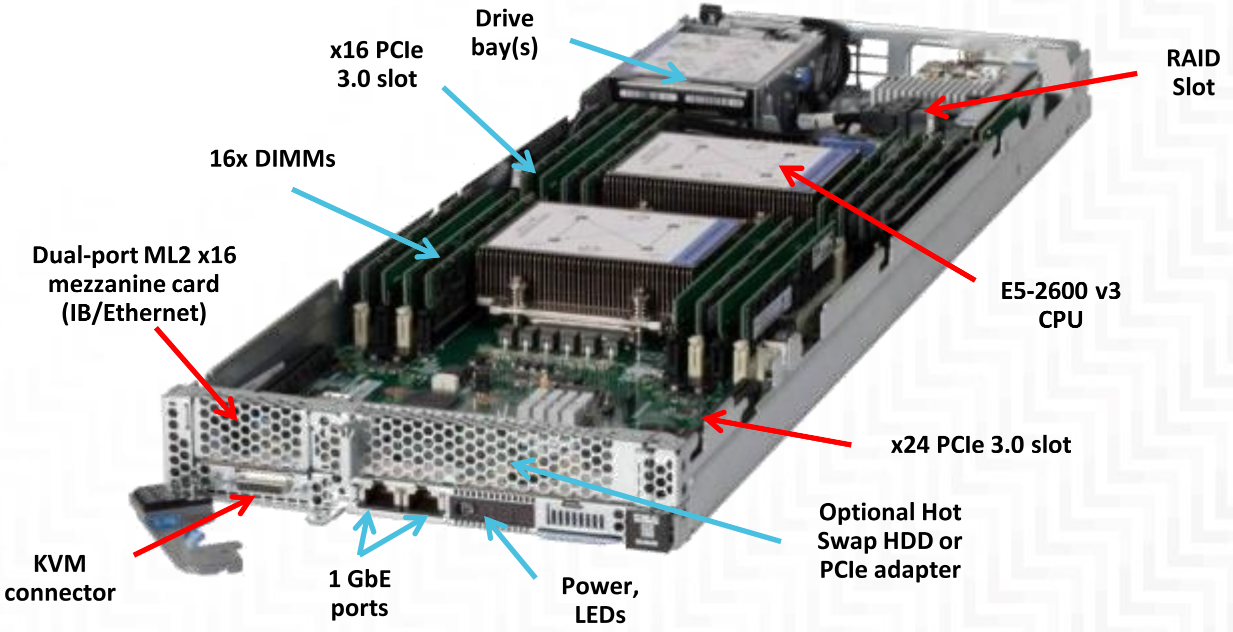
System infrastructure

Compute

- ½ Wide 1U, 2 socket server
- Intel E5-2600 v4 processors (up to 22C)
- 16x DIMM slots (DDR4, 2400MHz)
- 2 Front Hot-Swap HDD option (or std PCI slot)
- 4 internal HDD capacity
- Embedded RAID PCI slot
- ML2 mezzanine for x16 FDR and Ethernet

Key Differentiators:

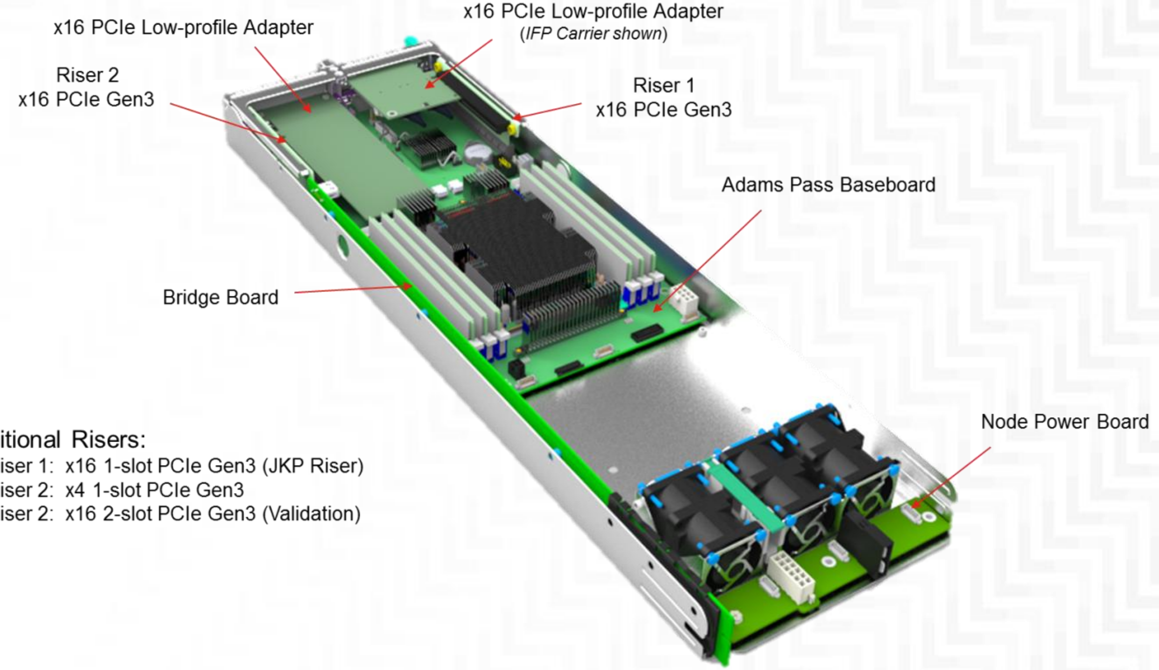
- Highest CPU support
- Support 2400MHz DDR4 series memory DIMM
- Flexible local storage options – choices of 3.5”, 2.5” HDD (SS & HS)
- Native expansion (NeX) support stoage and GPU/Phi



Intel Server board S7200AP(Adams Pass)

Board Features

- **Intel® Xeon® Phi™ processor (Bootable Knights Landing)**
  - Up to 200W TDP support;
  - **Intel® C-610 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



Additional Risers:

- Riser 1: x16 1-slot PCIe Gen3 (JKP Riser)
- Riser 2: x4 1-slot PCIe Gen3
- Riser 2: x16 2-slot PCIe Gen3 (Validation)

Chassis 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
- 2200W Redundant PSU

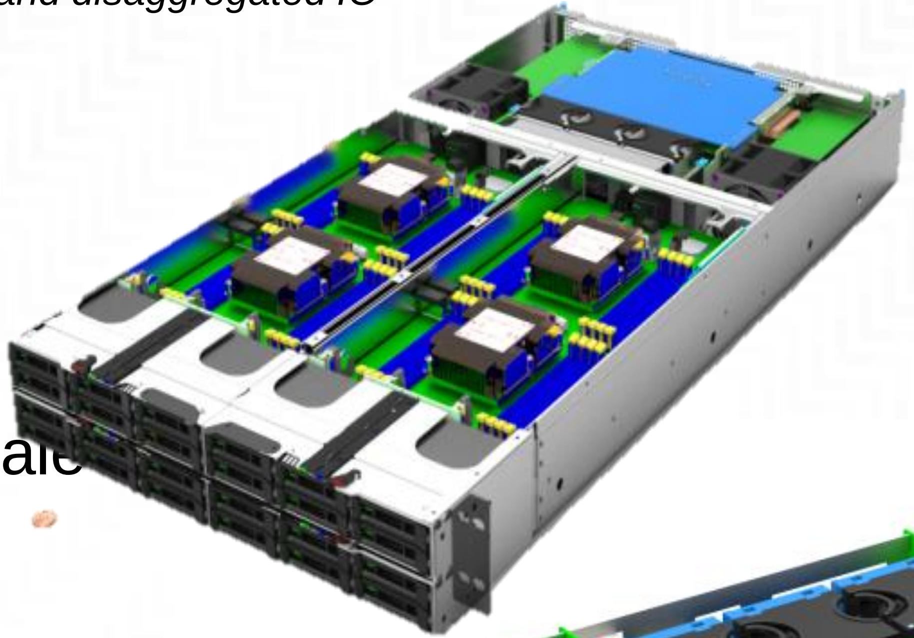
Rare view

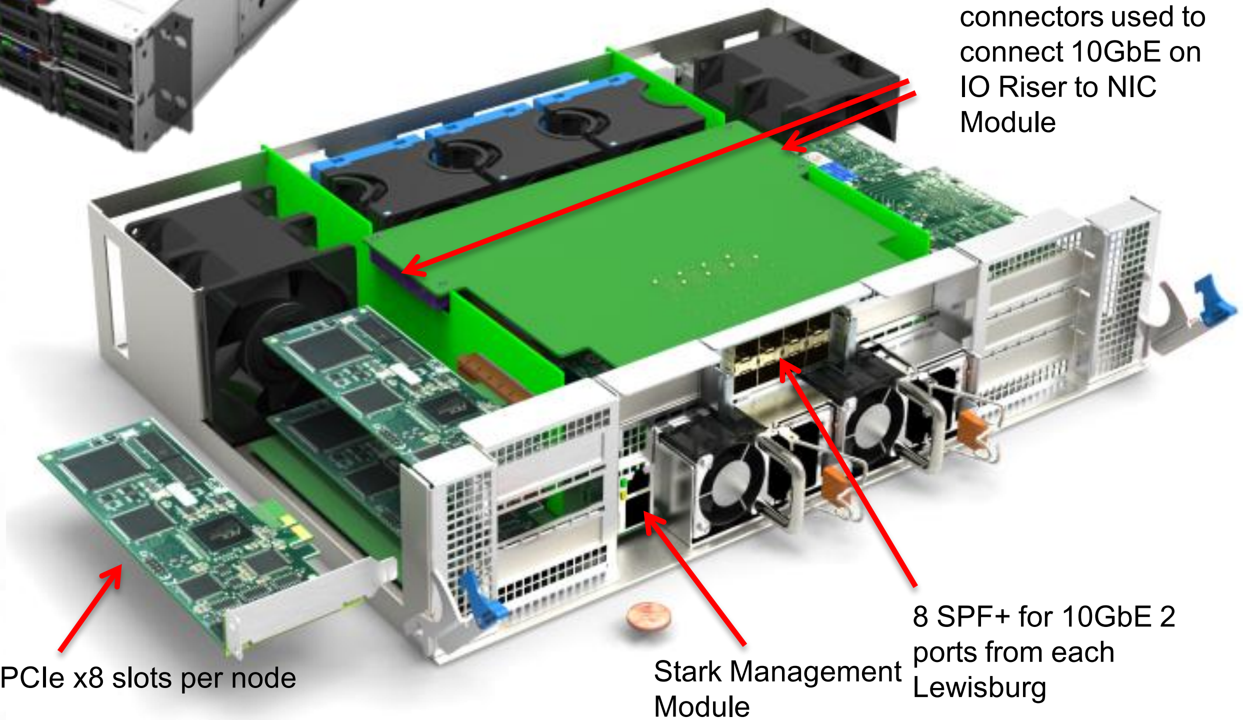
Front view

Stark – NeXt Generation Dense System

Stark Chassis – Flexible front installed nodes and disaggregated IO

- 2U Chassis
- 2P Purley Nodes (4x)
- Front accessible nodes
- 850mm depth
- Same density as NeXtScale
- Big increases in Storage
- More choice in IO
- 16 DIMM slots
- Apache Pass support
- Native 10Gb
- NVMe support
- M.2
- Common PSU across line up



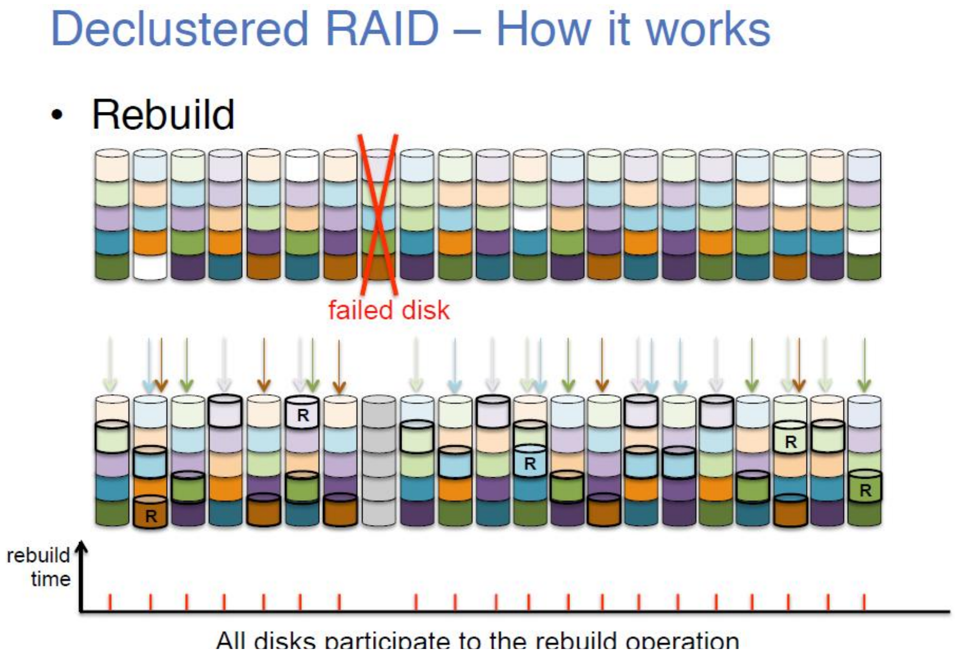


HPC Storage

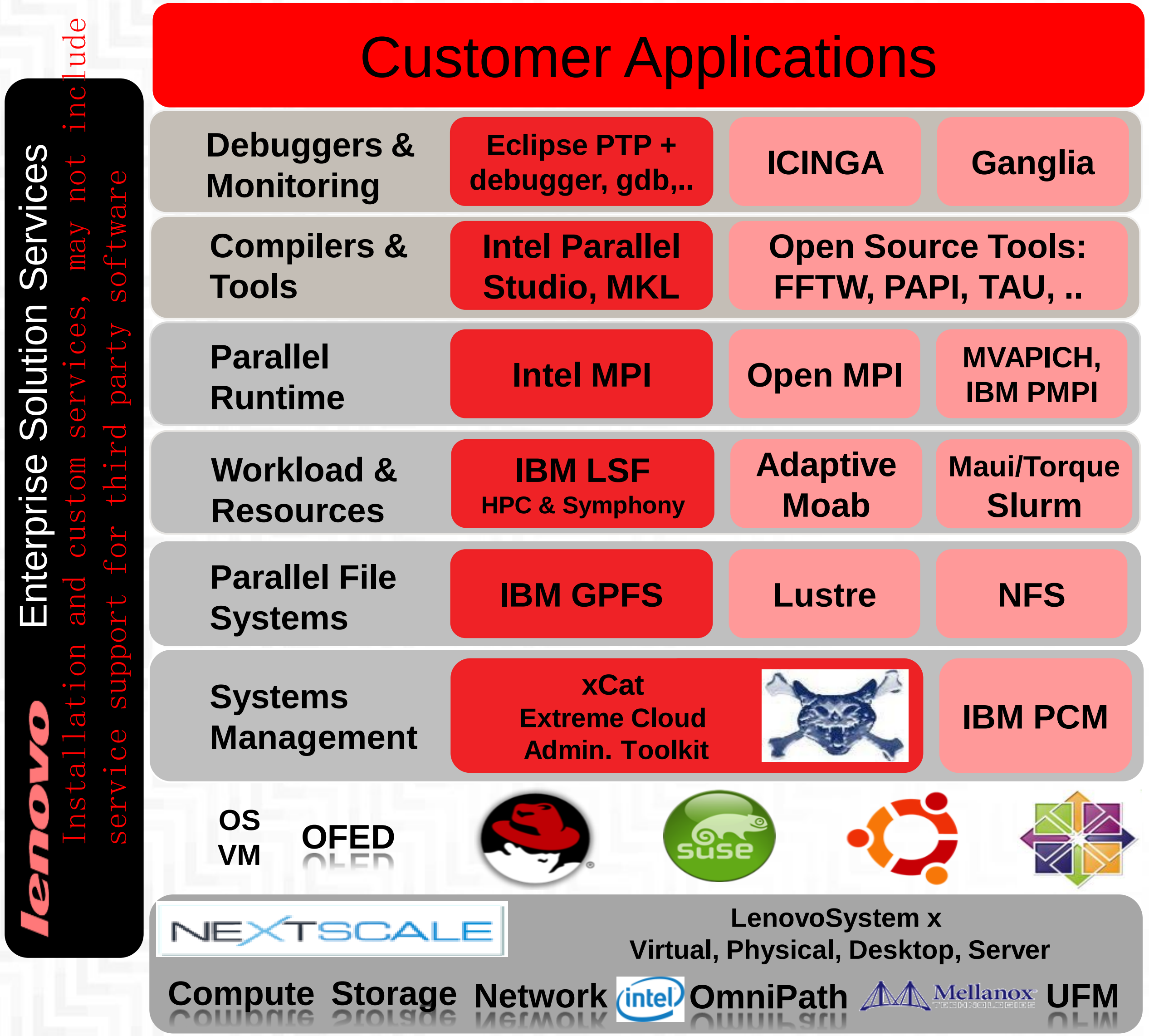
Lenovo GSS

Solution design

- **Embedded GPFS filesystem**
- **RAID support at filesystem level**
- **Fast data reconstruction by declustered RAID**
- **40GbE, FDR, EDR, OPA support**
- **Up-to 2.7PB raw in a system**
- **2 to 6 high density Jbod attached to two servers**
- **Reduced maintenance costs due to HW semplification**

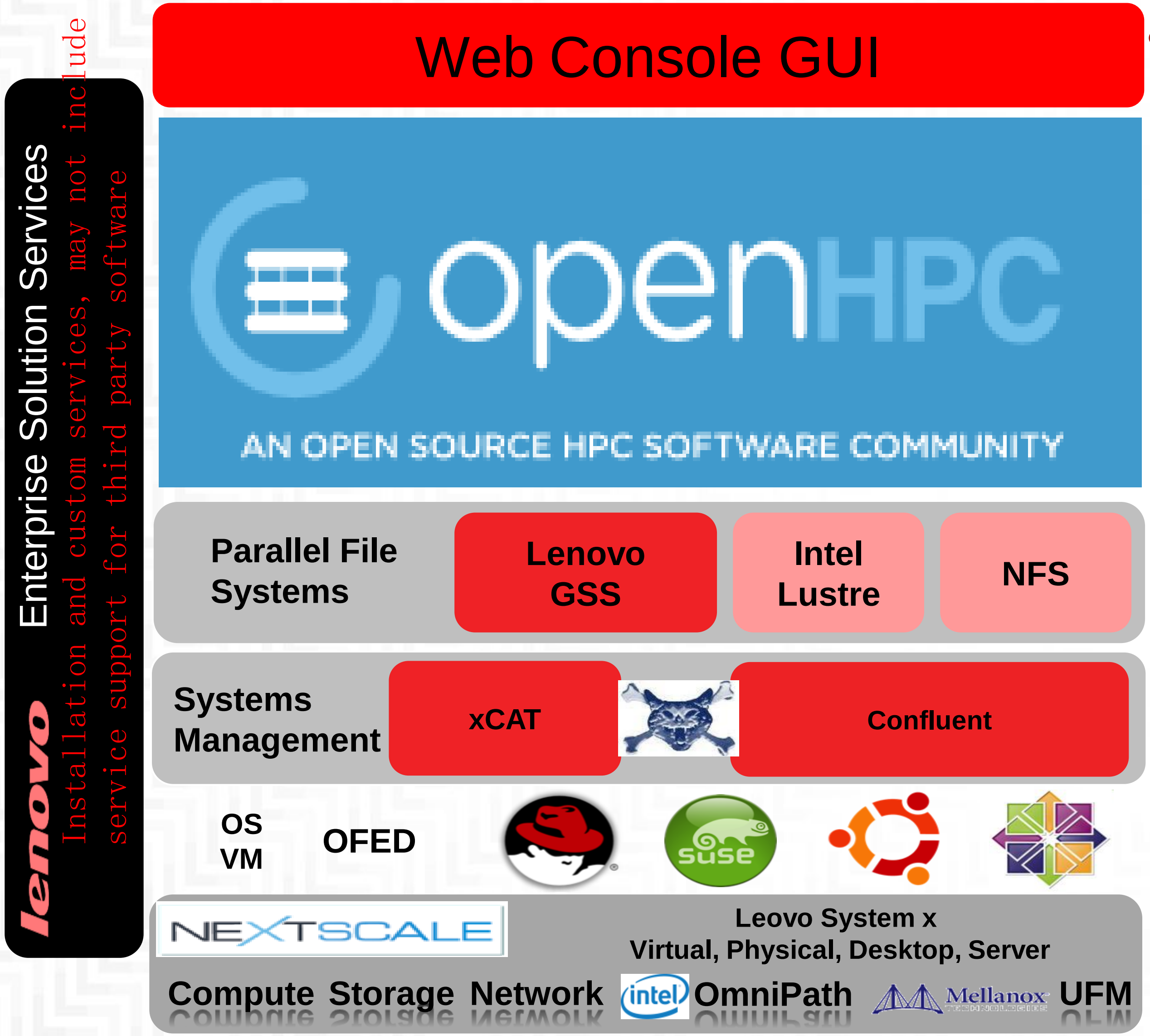


Current Lenovo HPC Software Solutions



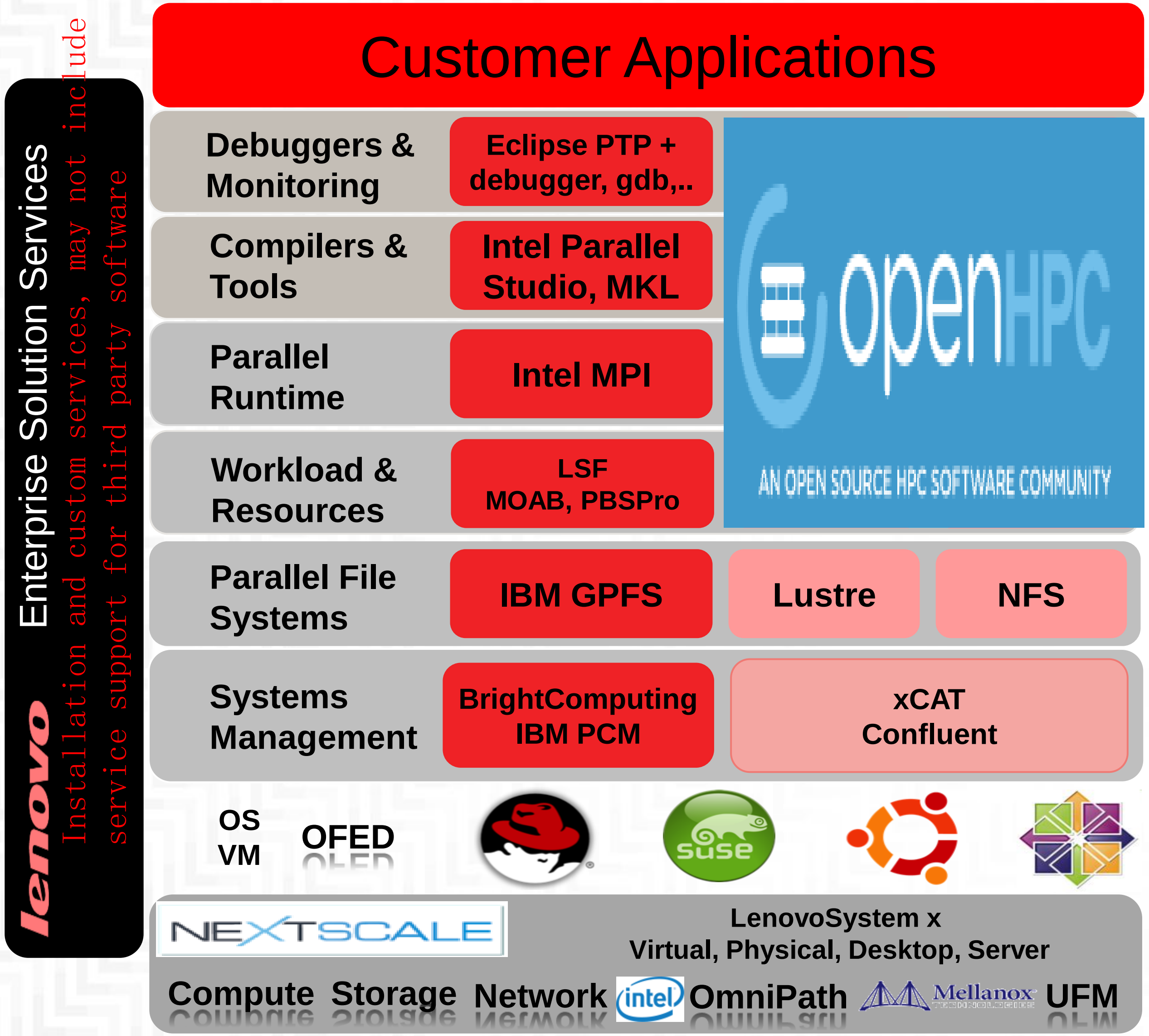
- **Building Partnerships to provide the "Best In-Class" HPC Cluster Solutions for our customers**
- Collaborating with software vendors to provide features that optimizes customer workloads
- Leveraging "Open Source" components that are production ready
- Contributing to "Open Source" (i.e. xCAT, Confluent, OpenStack) to enhance our platforms
- Providing "Services" to help customers deploy and optimize their clusters

Future HPC Open Source Management Stack



- **Adding new features to the stack**
  - **Web Console GUI**
  - **xCAT**
    - **Heat Map of servers/racks**
    - **Fluid Return Temperature /Flow rate of CDU**
  - **Energy Awareness**
    - **scheduler independent**

HPC Management Solutions through Partnerships



- **Provide an Open Source stack and a Commercial stack**
  - collaborating with software vendors to provide features optimized for Lenovo servers like Bright Computing and Altair

LENOVO HPC EMEA COMPETENCE CENTER

Think even bigger

HPC Innovation Center at Stuttgart, Germany

- System is Ready to do benchmarks

"Lenovo Means Business: New HPC Innovation Centre To Expand Enterprise Server Capabilities"
— Shannon Greenhalgh, Misco IT, UK, Mar 26 2015

"Chinese computing giant Lenovo has announced the opening of its first High-Performance Computing (HPC) Innovation Centre, located in Stuttgart, Germany"
— Gareth Halfacree, Bit-Tech, Mar 25 2015

"Lenovo creates an initial tellurian High Performance Computing (HPC) creation core in Stuttgart, Germany. This reaffirms their commitment to the space"
— Datacenter Management, Mar 26 2015

"Lenovo's energized Enterprise Systems Group comes out swinging"
— Charles King, Pund IT, Mar 25 2015

A Leadership Research and Development Center for Advancing HPC

- Bringing global talent together for accelerating HPC advancements
  - Strong core team with deep HPC skills
  - Core Center based in Stuttgart, Germany
  - Satellite centers hosted at customer locations and Lenovo Morrisville Benchmark Center
  - Creates a powerful innovation center with EU flavor and global reach

- Mission
  - Research, development and support to the Lenovo HPC business, our partners and clients
  - Active participation in the European HPC Community
  - Provides global benchmarking, proof of concept and solution demo capabilities

Lenovo

Core Client Partners

Technology Partners

Clients
BPs
SIs

Intel – Processing/Acceleration Innovation

Mellanox – High Speed Fabrics

Nvidia – Graphics Processing

IBM – Files System and Work Load

lenovo

HPC Innovation Center Europe

ISV

FZJ	LRZ	RZG	CINECA	BSC	STFC Hartree
HPC Storage	Energy Efficient Systems and Software	Advancing Material Science	Big Data and Many Cores	Extreme Application Scaling	HPC Software and Optimizatio

Bring together the newest technology and skills

Focused knowledge and deep skills advance the science of HPC

http://news.lenovo.com/news+releases/first-global-hpc-innovation-centre.htm

Built in Cooperation Between STFC Hartree Center, Lenovo and Cavium

- Deep server engineering skills
- Industry ecosystem relationships
- Clear desire to lead with new technology
- Driven to partner for the future
- NEXTSCALE

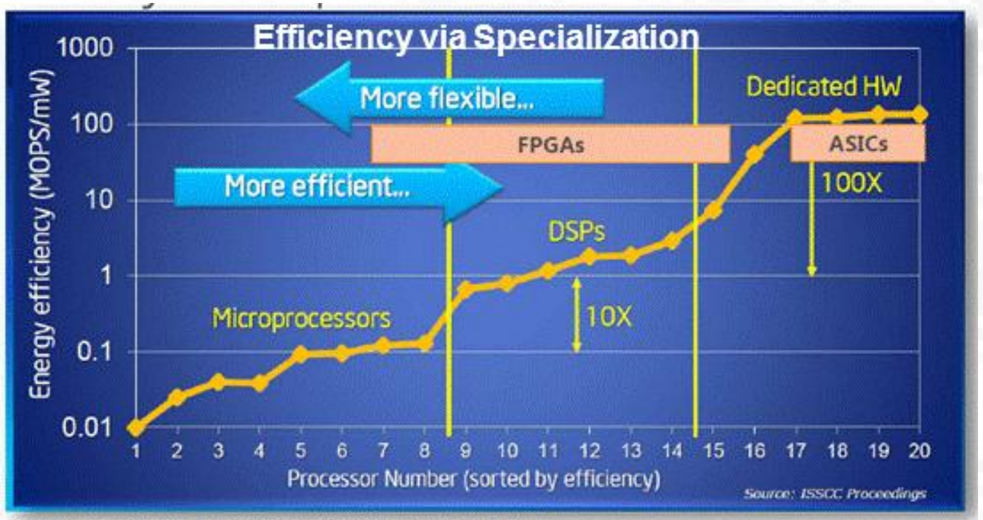
- Deep skills in software and code optimization
- Close partnerships with industry and commercial partners
- Focus on energy efficient computing

- Multi-generation and multi-core processor design experience
- Deep skills and IP in high performance SoCs including networking, accelerators and IO for ARM & MIPs

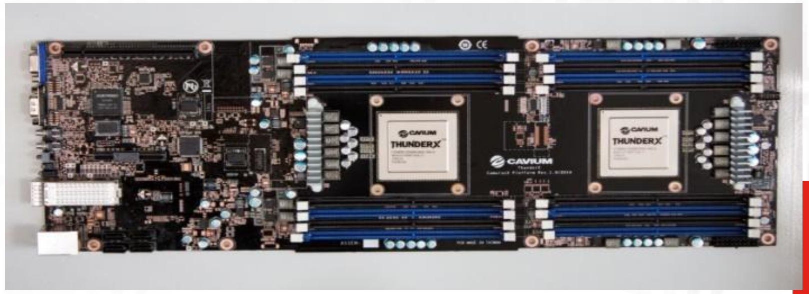
Workload Optimized Approach

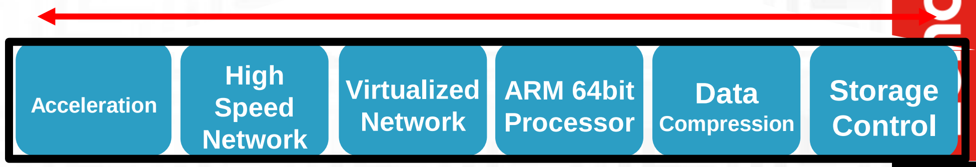
Workload Optimized

- Workload optimization brings together all functionality needed for a specific workload into one "foundation"
  - Software is the most flexible (general purpose computing)
  - FPGA / DSP can improve workload efficiency (today's workload optimization)
  - Dedicated hardware (in SoC) will give tremendous improvement (tomorrow's workload optimization)
- The benefits of this Workload Specific approach
  - Efficiency (performance, latency, power, and scalability)
  - All in one : customer can get packaged solution for the specific workload
- There are limitations – this is not for all workloads and all user
  - Balance between flexibility and efficiency for target segment
  - Highly targeted to specific applications



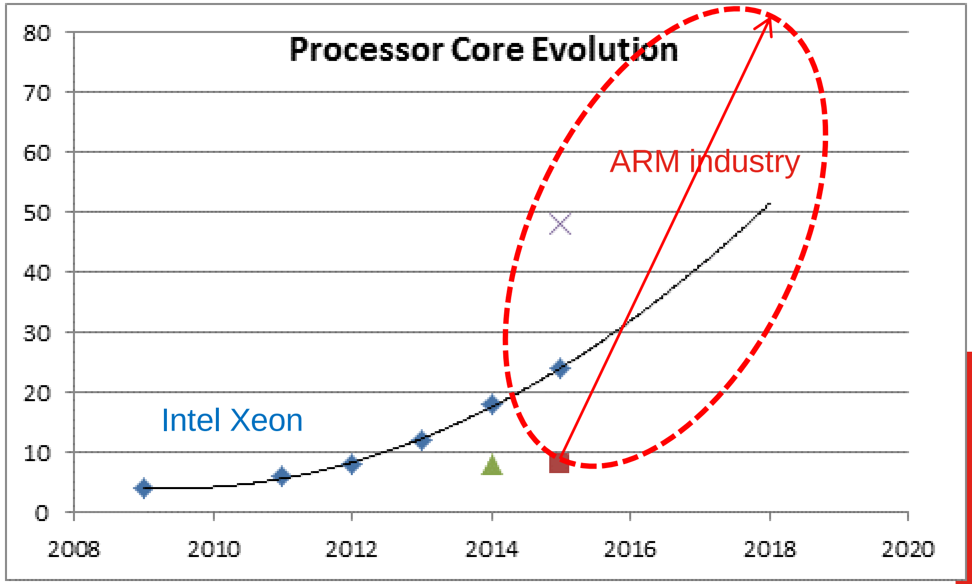
Source: Bob Broderson, Berkeley Wireless group





ARM64bit server toward Pre-EXA Scale

- ARM64bit = open architecture = innovation from industry
- ARM 64bit industry is accelerating to improve core performance and core count very aggressively.
- When 64 cores with 2SIMD/core @ 3GHz comes to market:
  - 64 core ARM 64bit @ 3.0GHz, 2 VFP SIMD / core = 3TFLOPS / 2S server (ARM SIMD : NEON (SP) 8 FLOPS / cycle, VFP (DP) 4 FLOPS / cycle)
- 1U packaging of ARM 2S + 2x GPGPU (3.5TFLOPS) can provide:
  - 10 TFLOPS (3 (ARM) + 3.5 (GPGPU) x 2) / U
  - 420TFLOPS / rack
- High bandwidth coherent interconnect integration is another technology focus area in ARM industry.



ARM64bit Partnership Landscape

Critical partnership engagement – Redhat, Mellanox, NVIDIA, PathScale, and PGI

Library

Starting new attention to library development.
First generation hardware in 2015 will accelerate the enablement

Compiler Optimization

GCC optimization is critical focus of both ARM and ARM SCO vendors

ARM & SOC vendors also engaging commercial compiler partners

IHV Support

Active enablement of HPC IHV under way.
More enablement will come in 2015 – 2016.

OS

Community

Enterprise

Started ARM64bit Early Access Program in 2014.

Lenovo Proposal

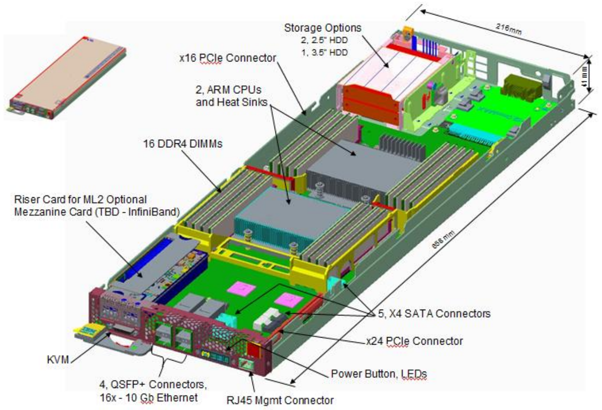
PHASE 1 (4Q 2015 – 1Q 2017)

ARM + GPGPU HPC early Investigation

- Cavium ThunderX (48core @ 2GHz, PCIe G3 8x)
- Mellanox ConnectX3Pro (FDR)
- NVIDIA K80 GPGPU x 2 (GPGPU tray)
- NeXtScale formfactor

- Commercial compiler (PathScale / PGI)

- Potential investigation to use integrated 10GbE for cluster communication



PHASE 2 (2H 2017 – 1H 2019)

HPC Optimized ARM server joint development

- 2nd Generation ARM 64bit (vendor TBD)
  32-64 cores, 2.8-3.0GHz, PCIe G3 16x (PCIe G4 TBD)
  NVLink 2.0 (TBD)
- Mellanox InfiniBand (EDR)
- NVIDIA Next Generation GPGPU (NVLink 2.0)

- Server formfactor (TBD)

- Potential investigation to use integrated high speed Ethernet for cluster communication

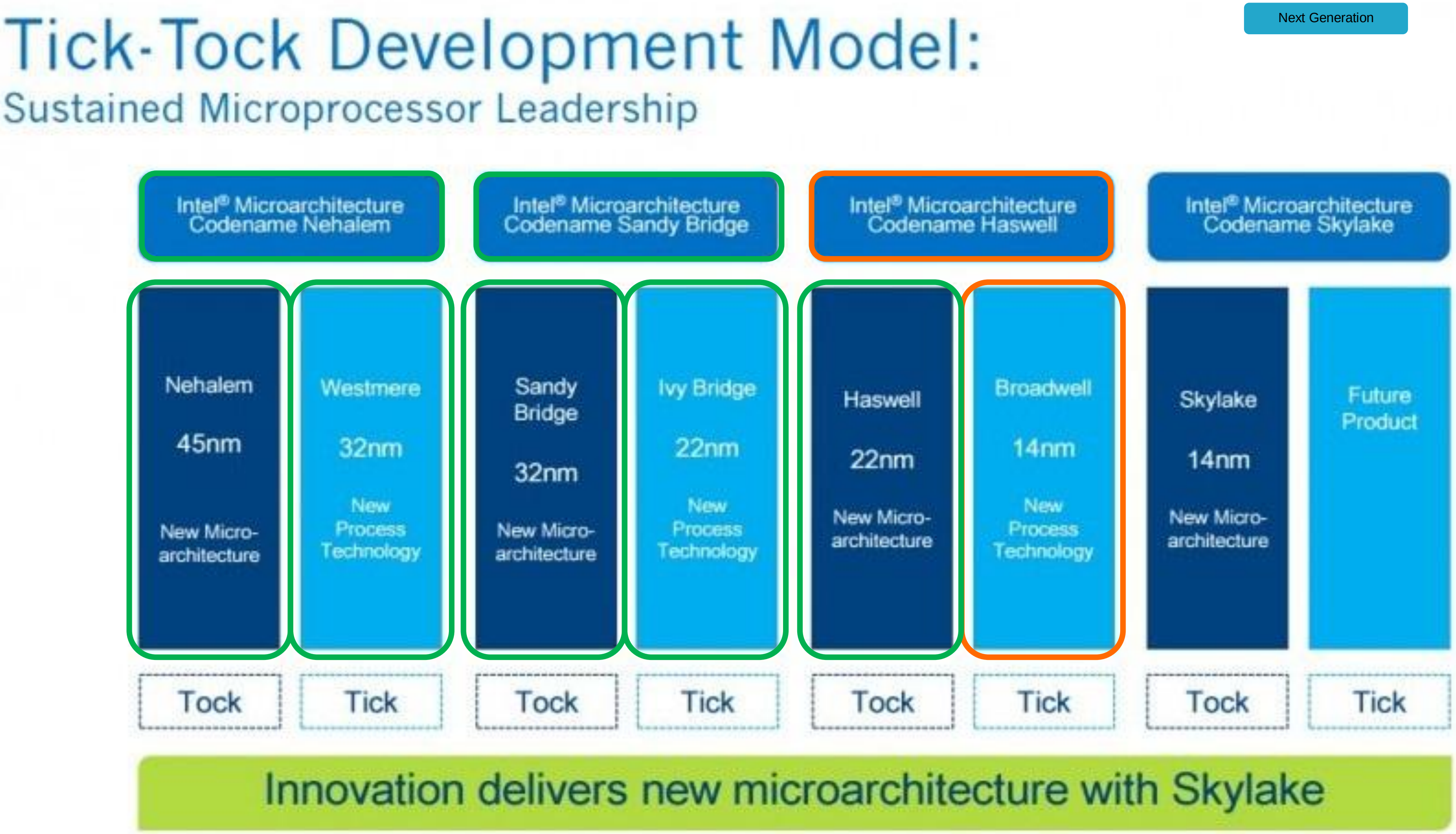
ARM innovation	Phase 1	Phase 2
core	48	32 (multi threads) - 64
GPGPU Link	PCIe G3 8x	PCIe G4 16x or NVLink 2.0
InfiniBand	FDR (PCIe G3 8x)	EDR (G3 16x or G4)
Compiler	Under optimization	Optimized (outlook)
Library	Under development	Matured (outlook)

TECHNOLOGY TRENDS

Think even bigger

Intel processors Development Model

Previous Generation
Current generation
Next Generation



INTEL® XEON® PROCESSOR E5-2600 V4 PRODUCT FAMILY:
PRELIMINARY 2S SERVER/WORKSTATION SKU LINE-UP

Intel® Xeon® processor E5-2600 v4 product family
Grantley Refresh Overview

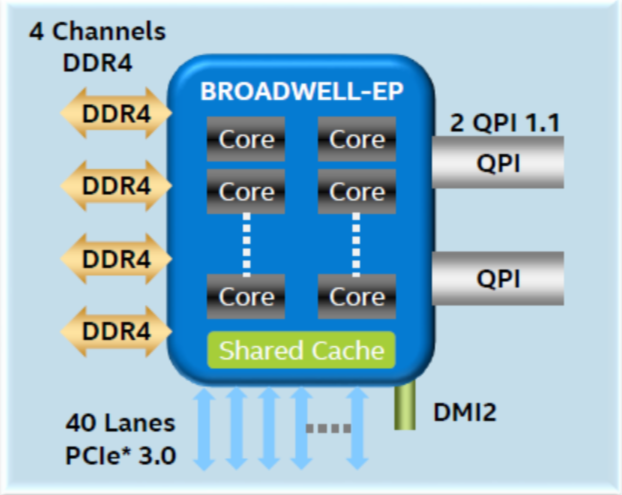
Broadwell microarchitecture

Built on 14nm process technology

Socket compatible# replacement for Intel® Xeon® processor E5-2600 v3 on Grantley

Several new features and capabilities

Feature	Xeon E5-2600 v3 (Haswell-EP)	Xeon E5-2600 v4 (Broadwell-EP)
Cores Per Socket	Up to 18	**Up to 22**
Threads Per Socket	Up to 36 threads	**Up to 44 threads**
Last-level Cache (LLC)	Up to 45 MB	**Up to 55 MB**
QPI Speed (GT/s)	2x QPI 1.1 channels 6.4, 8.0, 9.6 GT/s	
PCIe* Lanes/ Controllers/Speed(GT/s)	40 / 10 / PCIe* 3.0 (2.5, 5, 8 GT/s)	
Memory Population	4 channels of up to 3 RDIMMs or 3 LRDIMMs	**+ 3DS LRDIMM&**
Max Memory Speed	Up to 2133	**Up to 2400**
TDP (W)	160 (Workstation only), 145, 135, 120, 105, 90, 85, 65, 55	

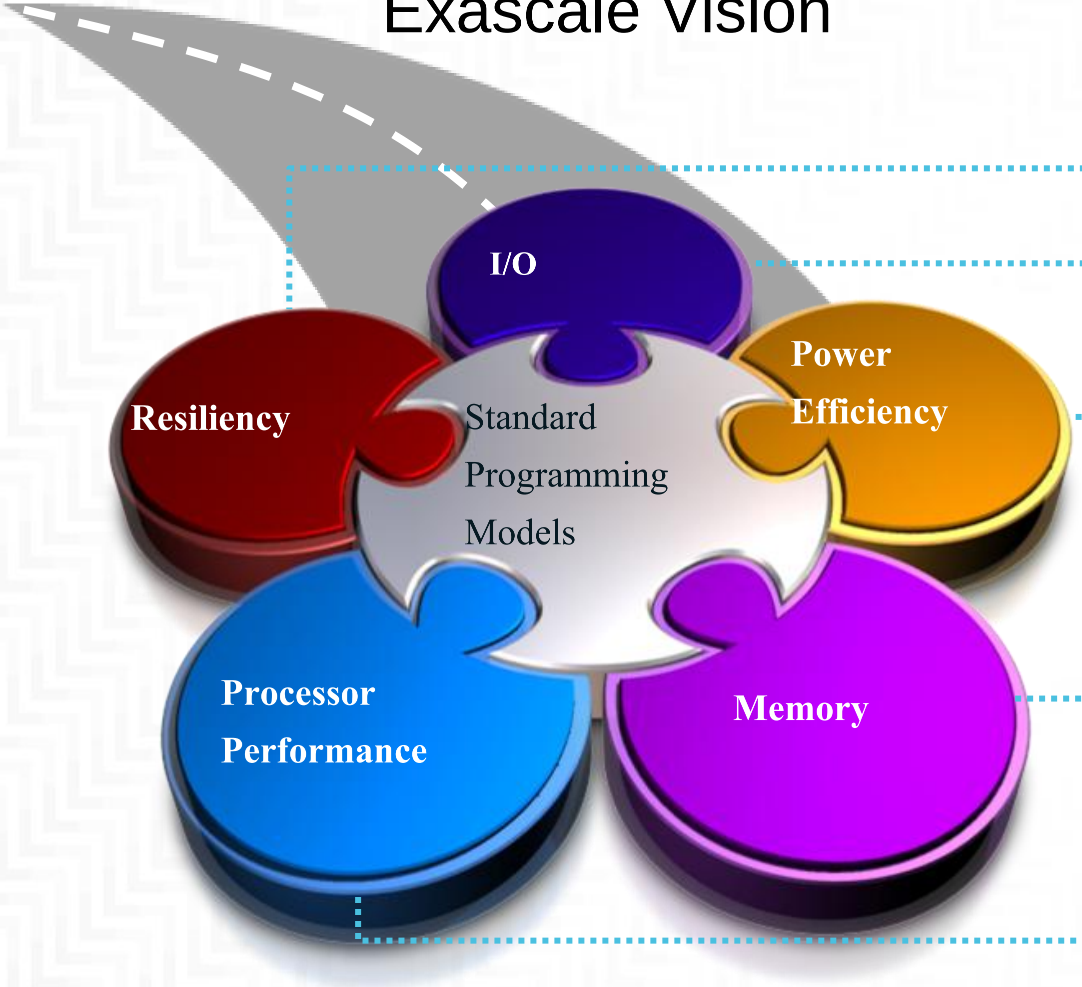


Requires BIOS and firmware update

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

Next Step on Intel's Path to Exascale Computing

Exascale Vision



Next Step: KNL

Systems scalable to >100 PFlop/s

Up to 100 Gb/s with Storm Lake integrated fabric

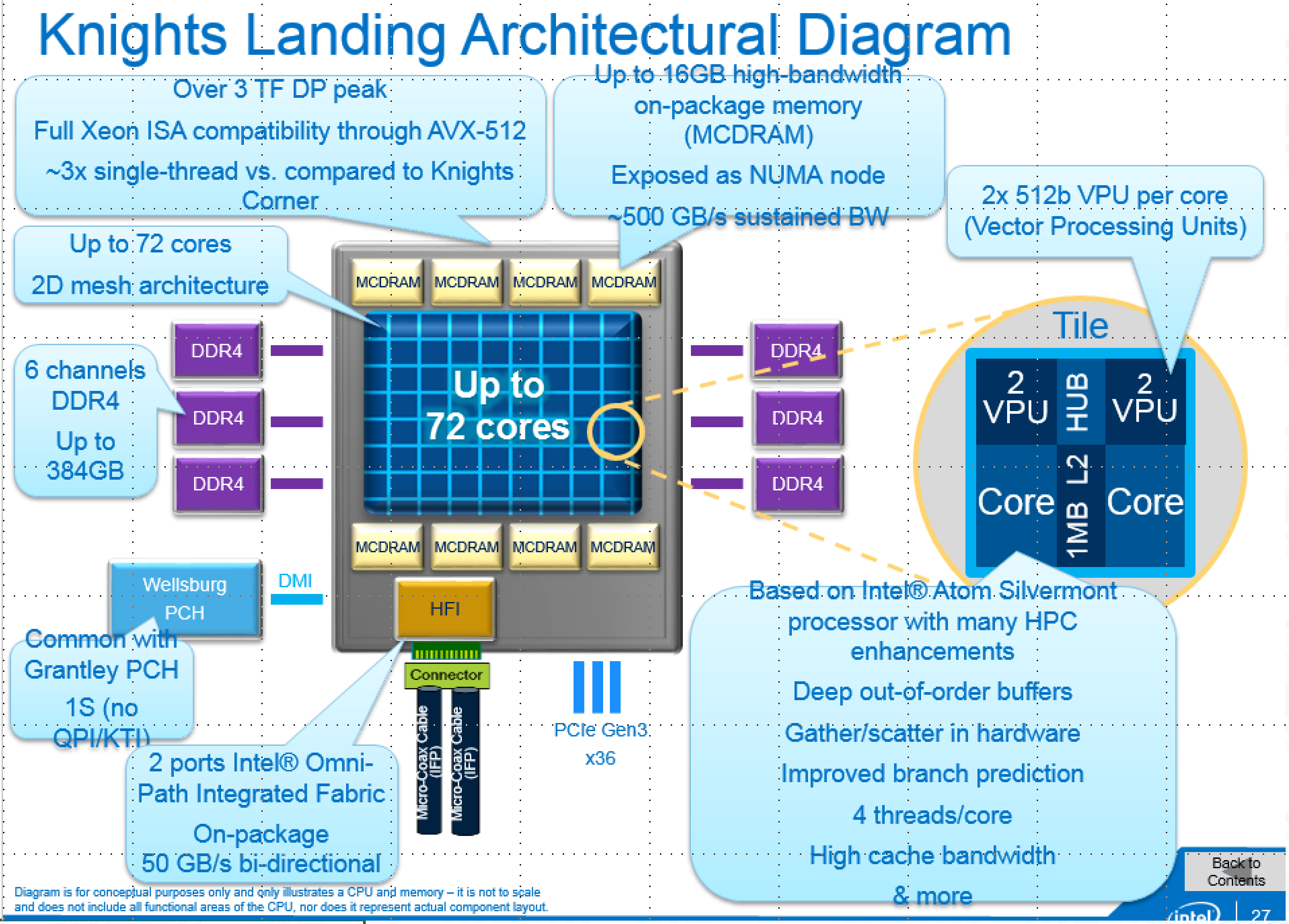
Over 15 GF/Watt[1]

~500 GB/s sustained memory bandwidth with integrated on-package memory

~3X Flops and ~3X single-thread theoretical peak performance over Knights Corner[1]

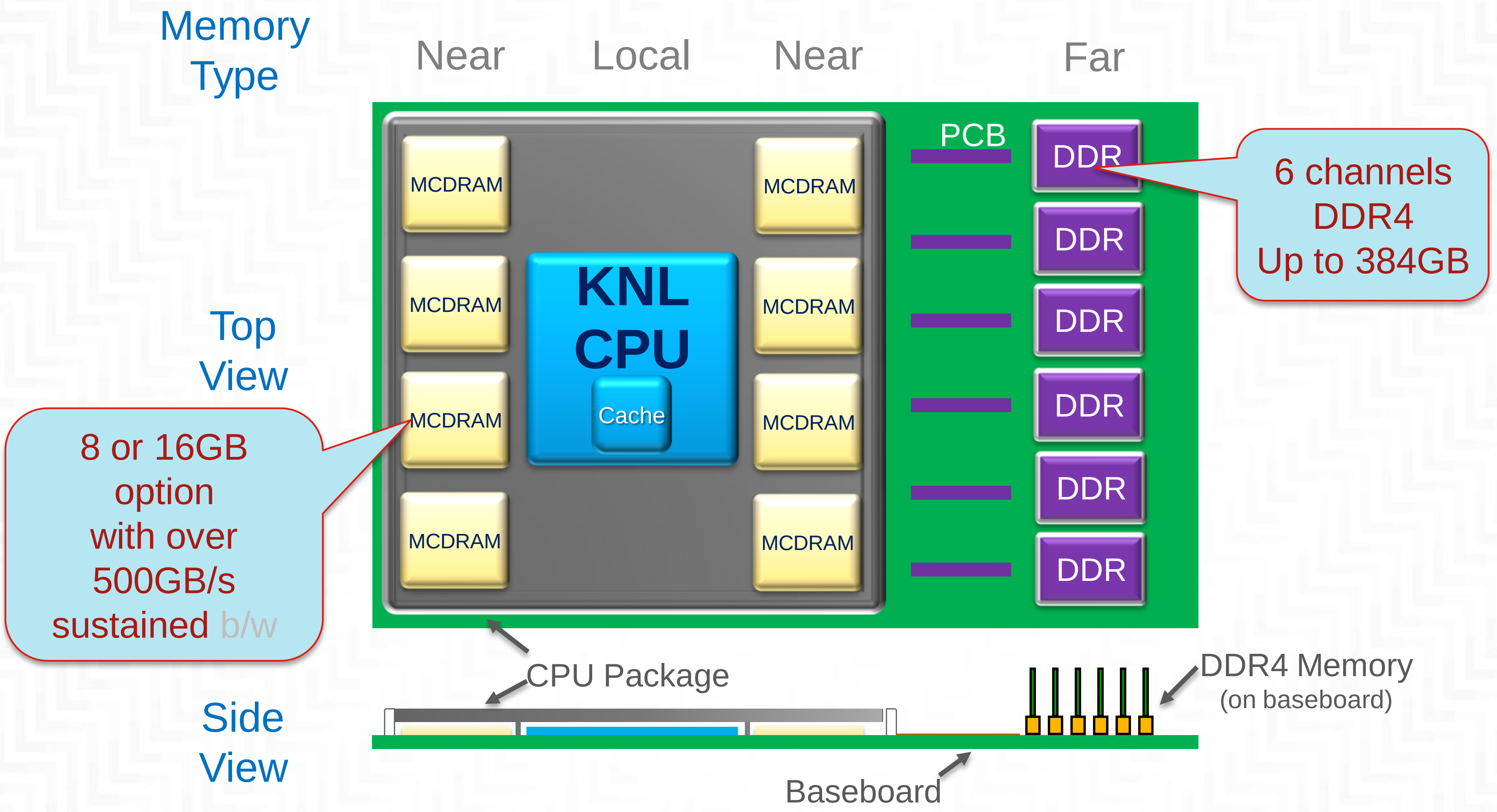
1 Projections based on internal Intel analysis during early product definition, as compared to prior generation Intel® Xeon Phi™ Coprocessors, and are provided for informational purposes only. Any difference in system hardware or software design or configuration may affect actual performance.

details



Knights Landing Architectural Diagram
Over 3 TF DP peak
Full Xeon ISA compatibility through AVX-512
~3x single-thread vs. compared to Knights Corner
Up to 16GB high-bandwidth on-package memory (MCDRAM)
Exposed as NUMA node
~500 GB/s sustained BW
2x 512b VPU per core (Vector Processing Units)
Up to 72 cores
2D mesh architecture
MCDRAM
DDR4
6 channels DDR4
Up to 384GB
Up to 72 cores
Tile
2 VPU
HUB
2 VPU
Core
1MB L2
Core
Wellsburg PCH
DMI
Common with Grantley PCH
1S (no QPI/KTI)
HFI
Connector
Micro-Coax Cable (IFP)
Micro-Coax Cable (IFP)
PCIe Gen3 x36
2 ports Intel® Omni-Path Integrated Fabric
On-package
50 GB/s bi-directional
Based on Intel® Atom Silvermont processor with many HPC enhancements
Deep out-of-order buffers
Gather/scatter in hardware
Improved branch prediction
4 threads/core
High cache bandwidth
& more
Diagram is for conceptual purposes only and only illustrates a CPU and memory – it is not to scale and does not include all functional areas of the CPU, nor does it represent actual component layout.
Back to Contents
27

KNIGHTS LANDING INTEGRATED ON-PACKAGE MEMORY



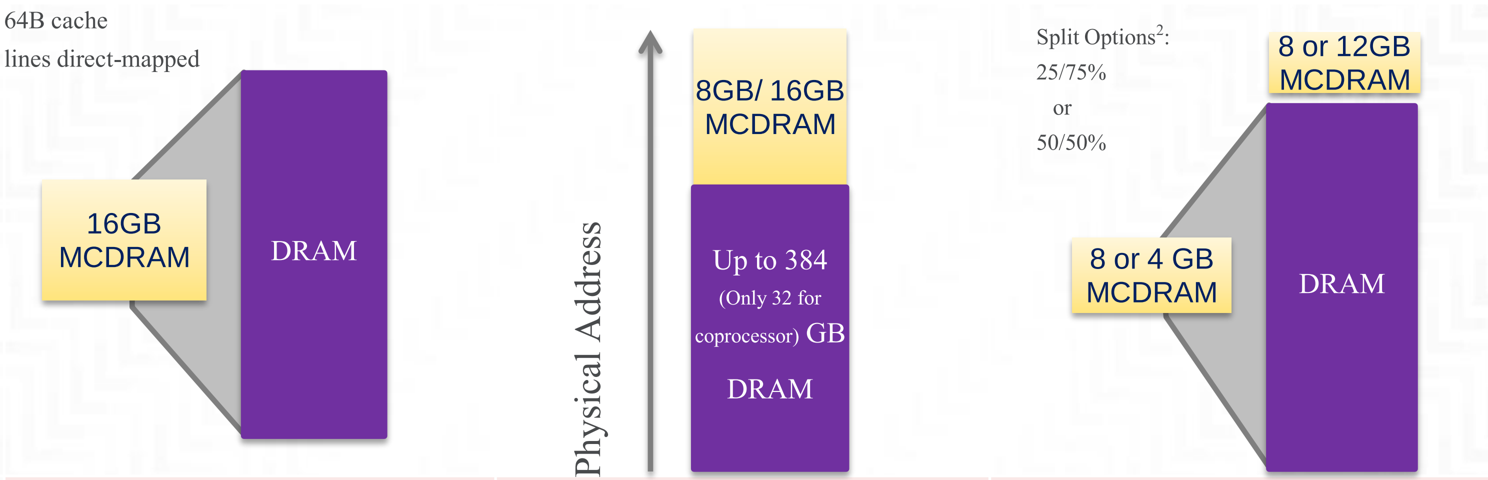
Integrated on-package MCDRAM brings memory nearer to CPU for higher memory bandwidth and better performance

Diagram is for conceptual purposes only and only illustrates a CPU and memory – it is not to scale and does not include all functional areas of the CPU, nor does it represent actual component layout.

Integrated On-Package Memory Usage Models

Model configurable at boot time and software exposed through NUMA[1]

	Cache Model	Flat Model	Hybrid Model



	Cache Model	Flat Model	Hybrid Model
Description	Hardware automatically manages the MCDRAM as a "L3 cache" between CPU and ext DDR memory	Manually manage how the app uses the integrated on-package memory and external DDR for peak perf	Harness the benefits of both Cache and Flat models by segmenting the integrated on-package memory
Usage Model	▪ App and/or data set is very large and will not fit into MCDRAM ▪ Unknown or unstructured memory access behavior	▪ App or portion of an app or data set that can be, or is needed to be "locked" into MCDRAM so it doesn't get flushed out	▪ Need to "lock" in a relatively small portion of an app or data set via the Flat model ▪ Remaining MCDRAM can then be configured as Cache

1. NUMA = non-uniform memory access
2. As projected based on early product definition

Back to Contents

Intel® Xeon Phi™ Product Family

Based on Intel® Many Integrated Core (MIC) Architecture

2013

Knights Corner

Intel® Xeon Phi™
x100 Product Family

22 nm process
Coprocessor
Over 1 TF DP Peak
Up to 61 Cores
Up to 16GB GDDR5

2016

Knights Landing

Intel® Xeon Phi™
x200 Product Family

14 nm process
Processor & Coprocessor
Over 3 TF DP Peak
Up to 72 Cores
On Package High-Bandwidth Memory
3X Single-Thread
Out-of-order core

Future

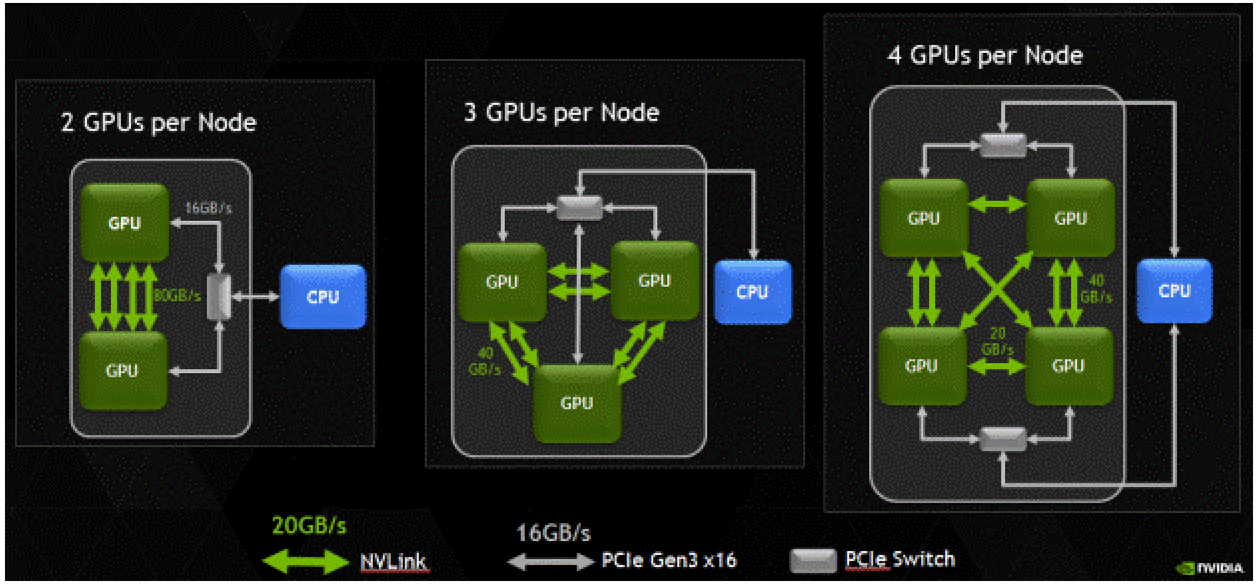
Knights Hill

Next generation of the Intel® MIC Architecture Product Line

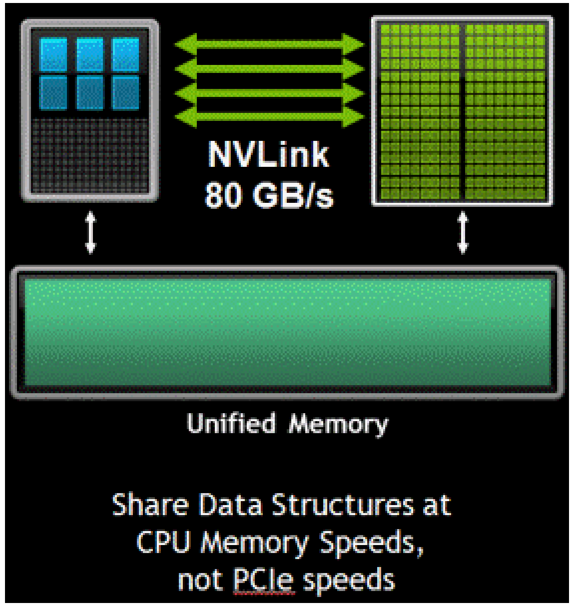
10 nm process
2nd Generation Integrated Intel® Omni-Path
… In planning …

NVLink

1st Generation



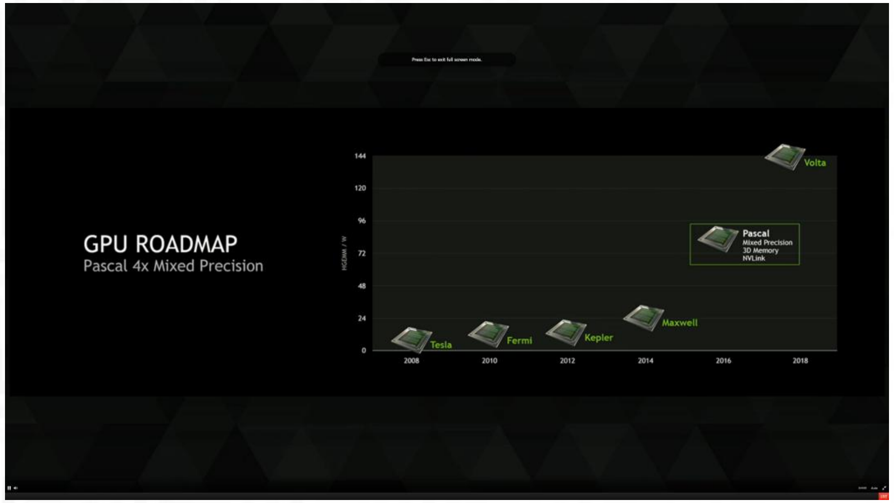
2nd Generation

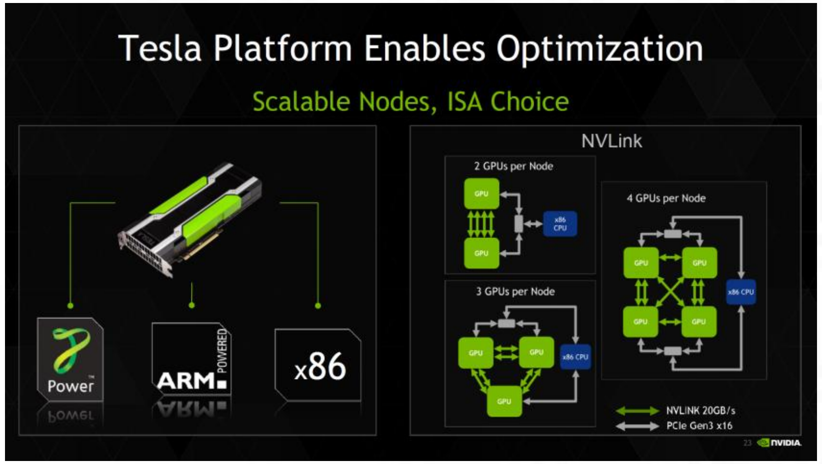


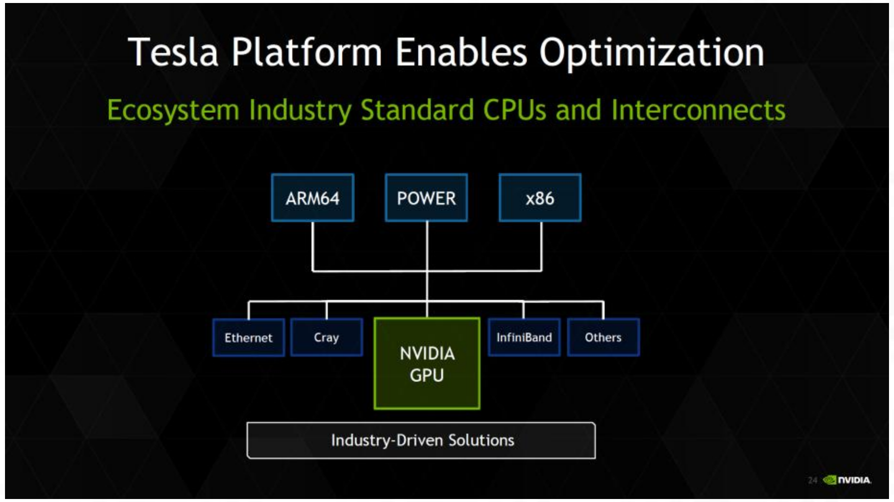
http://devblogs.nvidia.com/parallelforall/how-nvlink-will-enable-faster-easier-multi-gpu-computing/

NVIDIA trends

NVIDIA generations







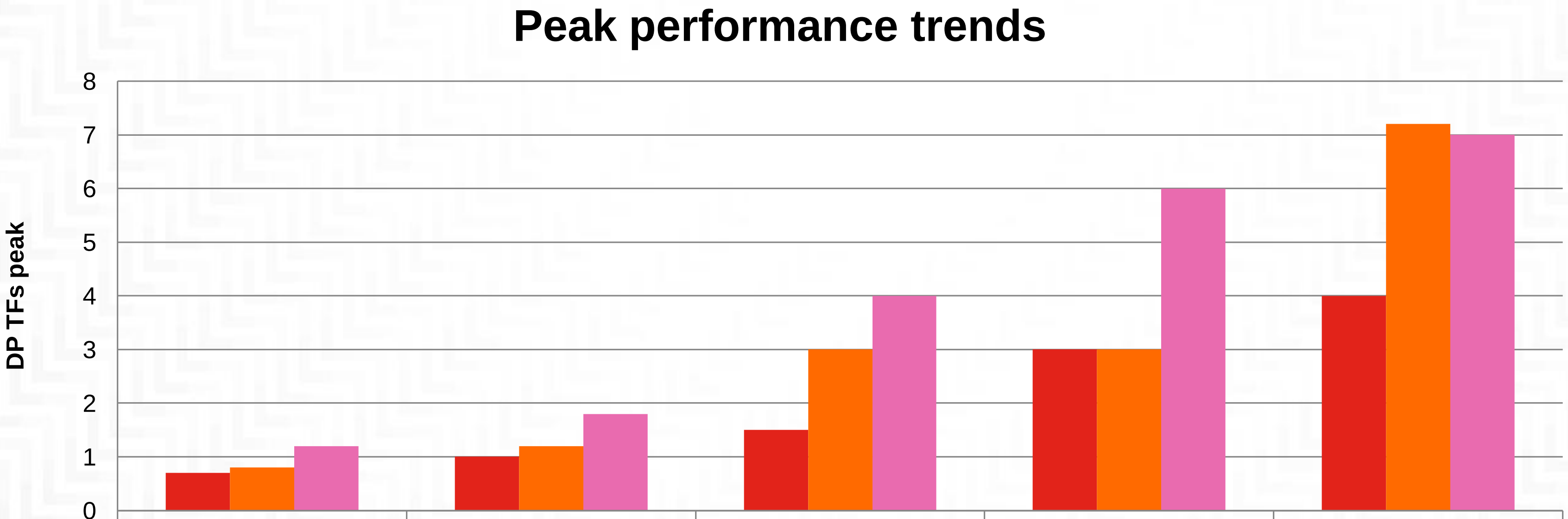
1Q17 Pascal ~4TFs NVLINK.1 ~10GFs/W

2Q18 Volta ~6TFs NVLINK.2 ~17GFs/W

1Q19 "Volta+" ~8-10TFs ~25GFs/W

Technology trends toward 2018 and beyond

- **Technology evolution determines a significant performance growth in the next 3yrs**
- **From 2015 to 2018 peak performances double at least on x86, X-Phi, GPUs**
- **Technology solutions to hundreds of PFs is not so evident and will depend by several conditions:**
  - **Peak performance vs cost**
  - **Peak performance vs power consumption (GFs/W)**
  - **Sustained performances vs power consumption and TCO**



	2014	2015	2016	2017	2018
x86 2p	0,7	1	1,5	3	4
X-Phi	0,8	1,2	3	3	7,2
GPUs	1,2	1,8	4	6	7

COOLING TECHNOLOGY AND TCO

Think even bigger

How to measure Power Efficiency

- PUE

$$PUE = \frac{\text{Total Facility Power}}{\text{IT Equipment Power}}$$

  - **Power usage effectiveness** (**PUE**) is a measure of how efficiently a computer data center uses its power;
  - PUE is the ratio of total power used by a computer facility to the power delivered to computing equipment.
  - Ideal value is 1.0
  - It does not take into account how IT power can be optimised
- ITUE

$$ITUE = \frac{(\text{IT power + VR + PSU + Fan})}{\text{IT Power}}$$

  - **IT power effectiveness** (ITUE) measures how the node power can be optimised
  - Ideal value if 1.0
- ERE

$$ERE = \frac{\text{Total Facility Power} - \text{Treuse}}{\text{IT Equipment Power}}$$

  - **Energy Reuse Effectiveness** measures how efficient a data center reuses the power dissipated by the computer
  - ERE is the ratio of total amount of power used by a computer facility to the power delivered to computing equipment.
  - An ideal ERE is 0.0. If no reuse, ERE = PUE

How Direct Water Cooling Makes Your Data Center More Efficient and Lowers Costs

- **Chillers not required** for most geographies
  - Due to inlet water temperature of 18°C to 45°C
  - Reduce CAPEX for new data centers

- **40% energy savings** in datacenter due to no fans or chillers.

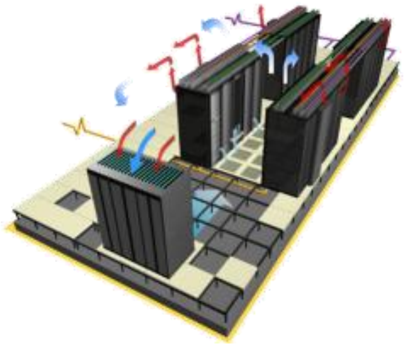
- Compute node **power consumption reduced ~ 10%** due to lower component temperatures (~5%) and no fans (~5%)

- Power Usage Effectiveness P_{Total} / P_{IT}: **PUE ~ 1.1** possible with NeXtScale WCT
  - 1.1 PUE achieved at LRZ installation
  - 1.5 PUE is typical of a very efficient air cooled datacenter.

- **85-90% Heat recovery** is enabled by the compute node design
  - Heat energy absorbed may be reutilized for heating buildings in the winter
  - Energy Reuse Effectiveness (P_{Total} – P_{Reuse}) / P_{IT}: **ERE ~ 0.3**

Choice of Cooling

Air Cooled



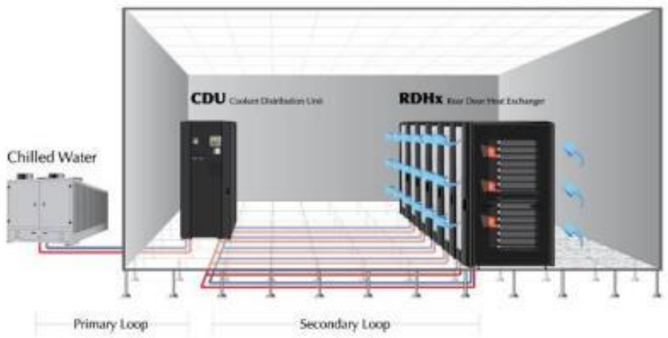
- Standard air flow with internal fans
- Fits in any datacenter
- Maximum flexibility
- Broadest choice of configurable options supported
- Supports Native Expansion nodes (Storage NeX, PCI NeX)

PUE ~1.5

ERE ~ 1.5

Choose for broadest choice of customizable options

Air Cooled with Rear Door Heat Exchangers



- Air cool, supplemented with RDHX door on rack
- Uses chilled water with economizer (18C water)
- Enables extremely tight rack placement

PUE ~1.2

ERE ~ 1.2

Choose for balance between configuration flexibility and energy efficiency

Direct Water Cooled

- Direct water cooling with no internal fans
- Higher performance per watt
- Free cooling (45C water)
- **Energy re-use**
- Densest footprint
- Ideal for geos with high electricity costs and new data centers
- Supports highest wattage processors

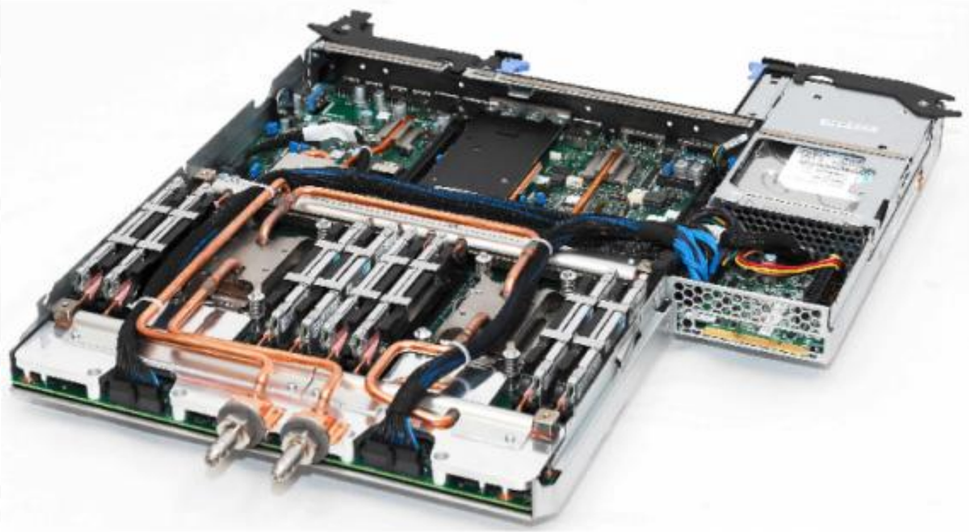
PUE <= 1.1

ERE ~ 0.3 with hot water

Choose for highest performance and energy efficiency

iDataplex dx360M4 (2010-2013)

- iDataplex rack with 84 dx360M4 servers
- dx360 M4 nodes, 2xCPUs (130W, 115W), 16xDIMMS (4GB/8GB), 1HDD/2SSD, network card.
- ~85% Heat Recovery, Water 18°C-45°C, 0.5 lpm / node.



dx360M4 Server

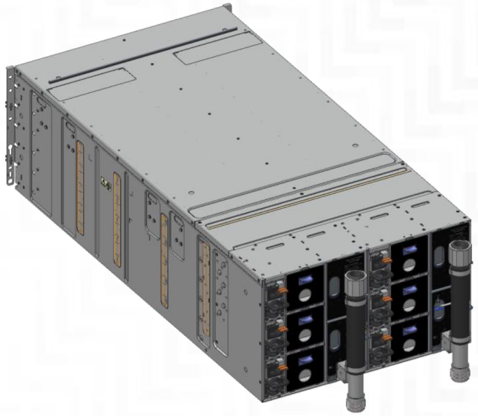
iDataplex Rack

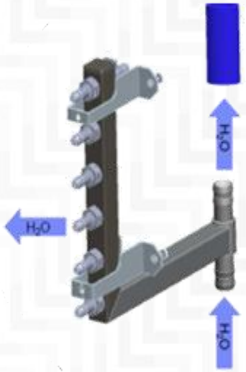
NextScale nx360M5 WCT (2013-2015)

- NextScale Chassis 6U/12Nodes , 2 nodes / tray.
- nx360M5 WCT 2xCPUs (up to 165W), 16xDIMMS (8GB/16GB/32GB), 1HDD/2SSD, 1 ML2 or PCIe Network Card.
- ~85% Heat Recovery, Water 18°C-45°C (and even upto 50°C), 0.5 lpm / node.

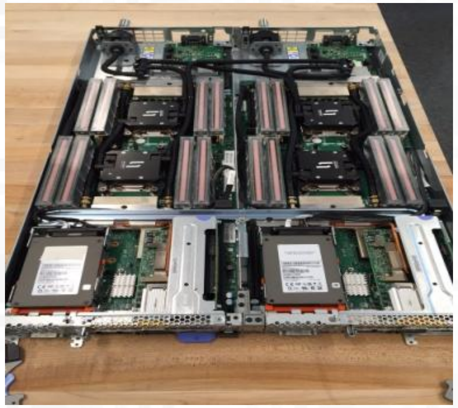
2 Nodes of nx-360M5 WCT in a Tray

NextScale Chassis

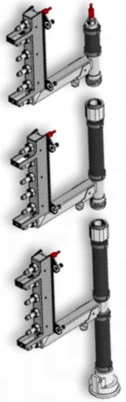




Single Manifold Drop (1 per chassis)



nx360M5 with 2 SSDs



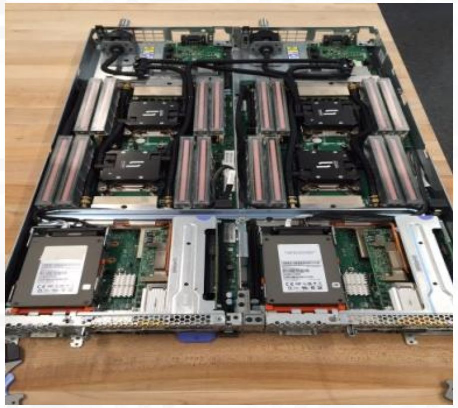
Scalable Manifold

6 drop Manifold

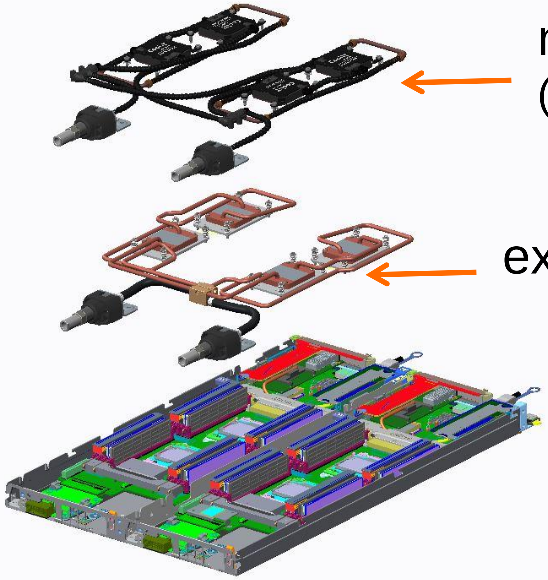
NextScale nx360M5 WCT (2016)

- NextScale Chassis 6U/12Nodes , 2 nodes / tray.
- nx360M5 WCT 2xCPUs (up to 165W), 16xDIMMS (8GB/16GB/32GB), 1HDD/2SSD, 1 ML2 or PCIe Network Card.
- ~85% Heat Recovery, Water 18°C-45°C, 0.5 lpm / node.

2 Nodes of nx-360M5 WCT in a Tray



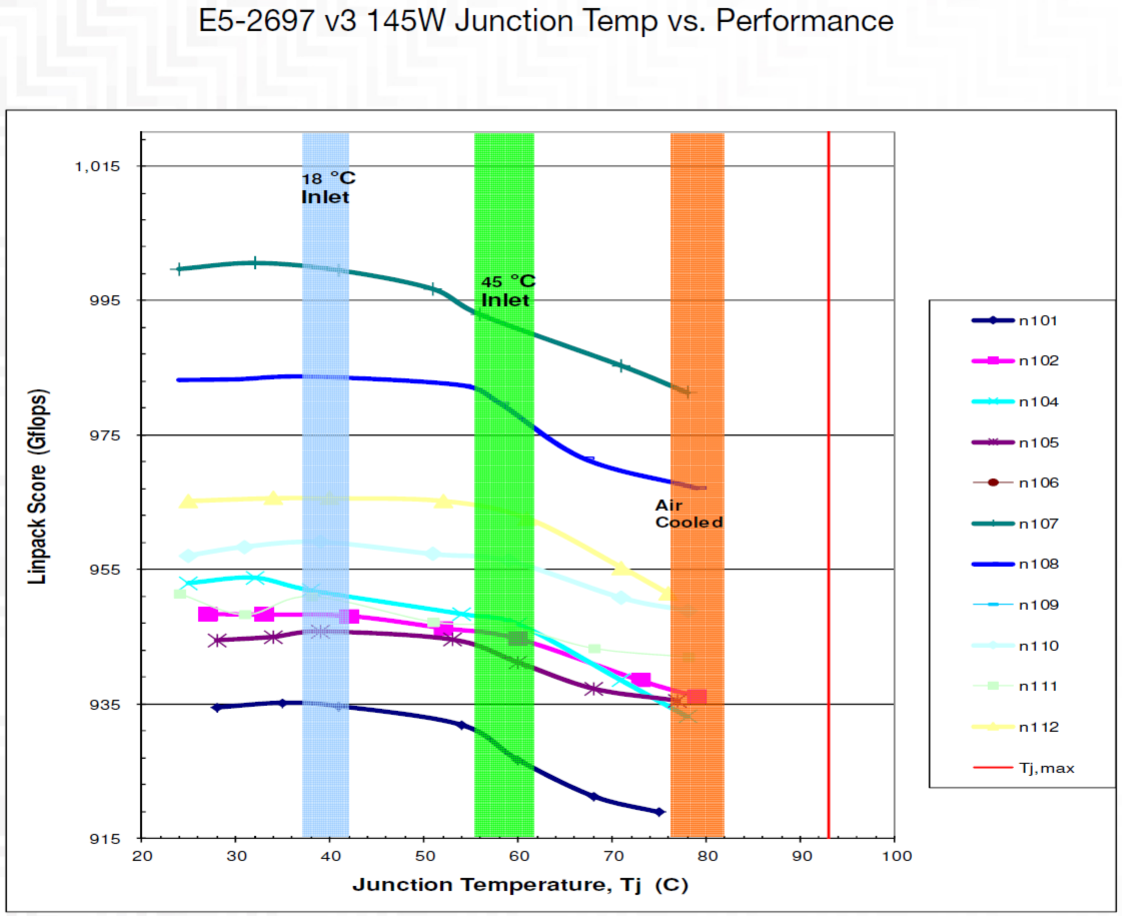
nx360M5 with 2 SSDs



new cost reduced FEP waterloop
(Fluorinated Ethylene Propylene)

existing copper waterloop

Power consumption, Junction Temperature and Cooling

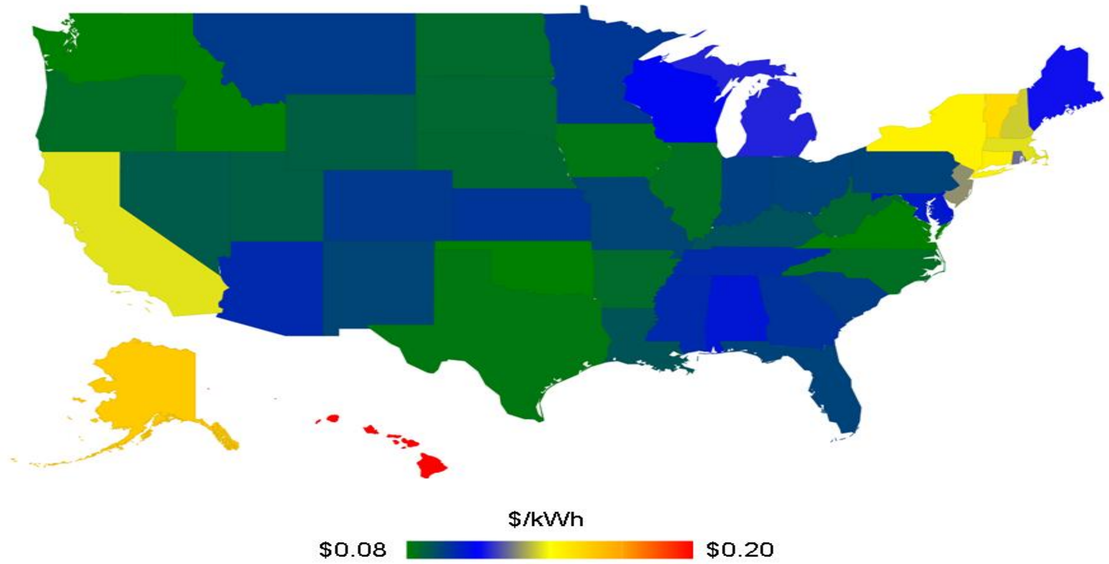


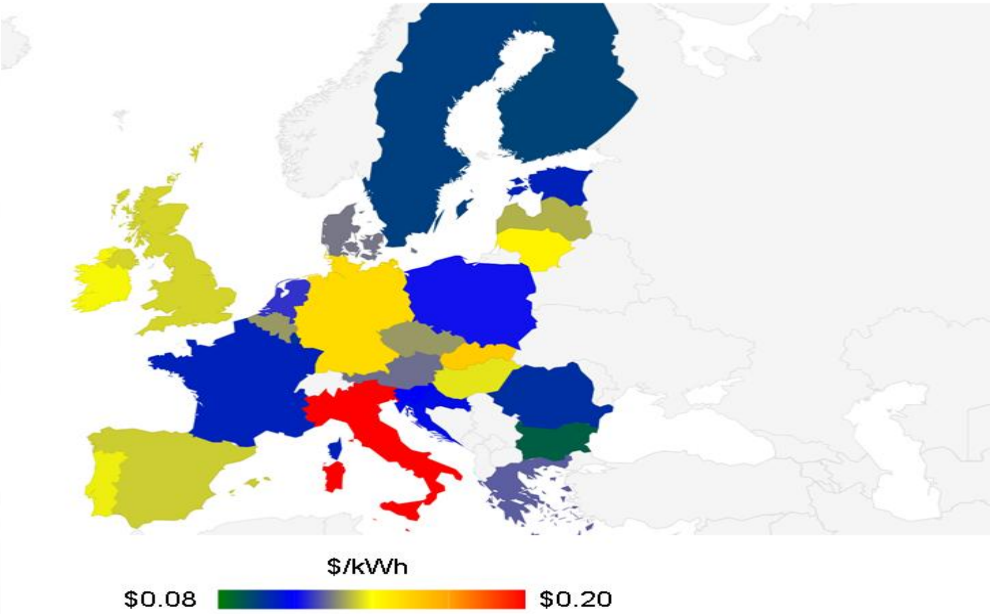
- Example: HPL scores across a range of temperatures:
  - 12 sample processors running on NeXtScale System WCT
- HPL scores remain mostly flat for junction temperatures in the range that water cooling operates.
- The HPL scores drop significantly when junction temperature is in range that air cooling operates.
- Conclusion: Direct Water Cooling lowers processor power consumption by about 6% or enables highest performance

* Vinod Kamath

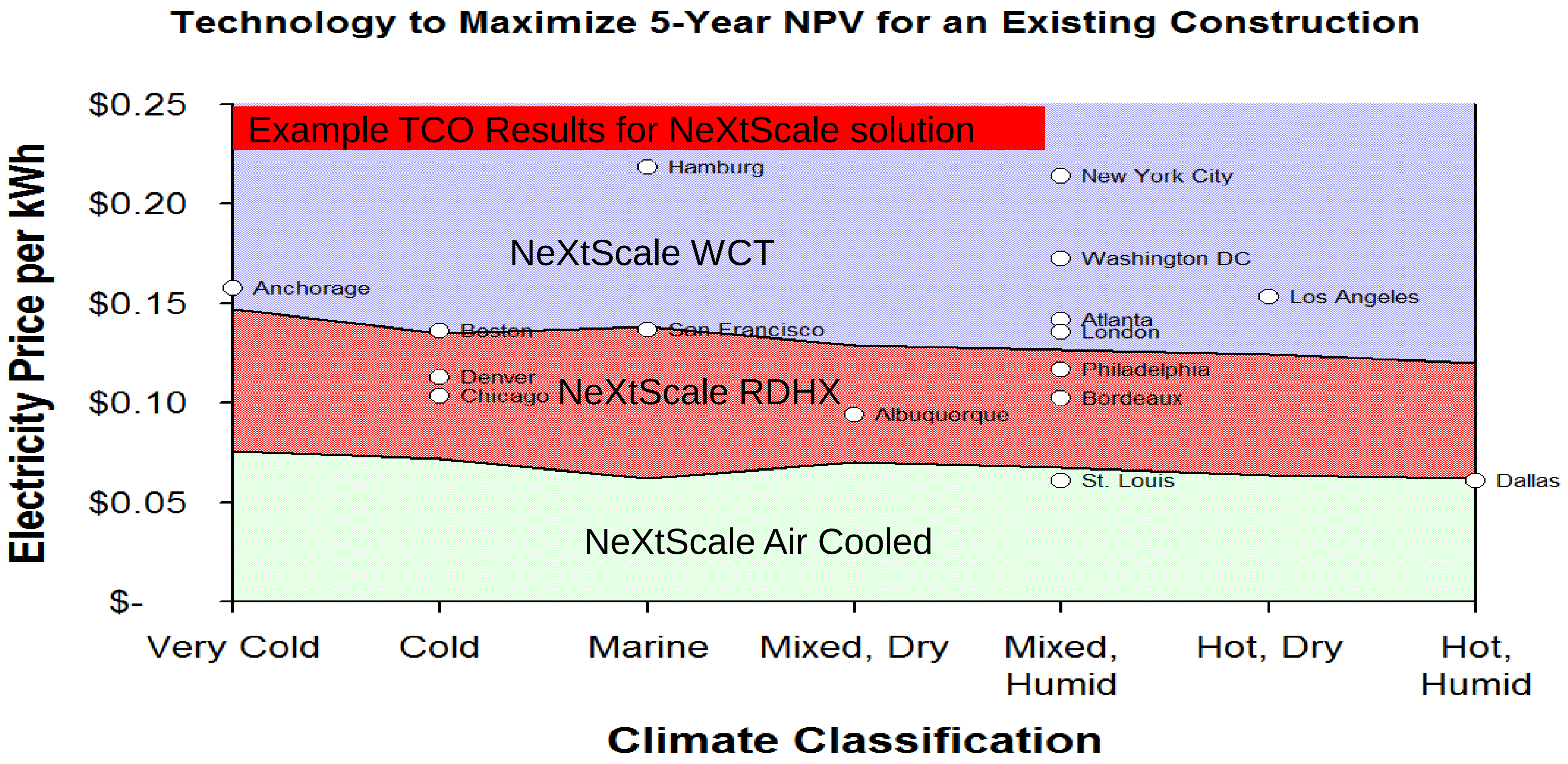
Goals of HPC Total Cost of Ownership Study

- Show the return on investment (ROI) for customers based on their geographic location, weather and energy cost.
- Compare and present data that shows the value of 3 cooling technologies, Air, RDHx and Direct Water Cooling (DWC) based on above parameters.
- Compare the ROI for three cooling technologies when installed in existing conventional data centers (Brown Field) and when installed into data centers with all new infrastructure (Green Field).

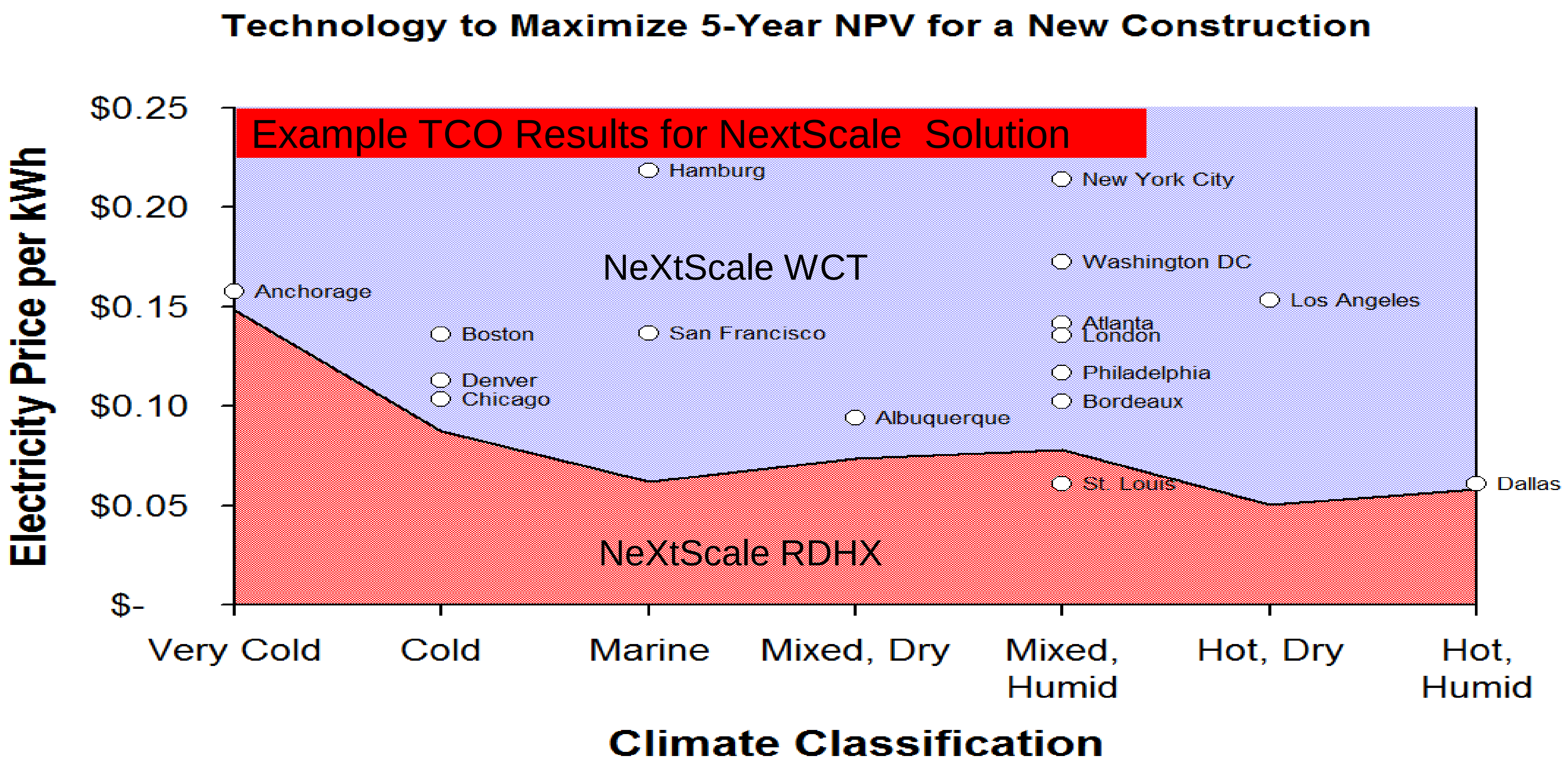




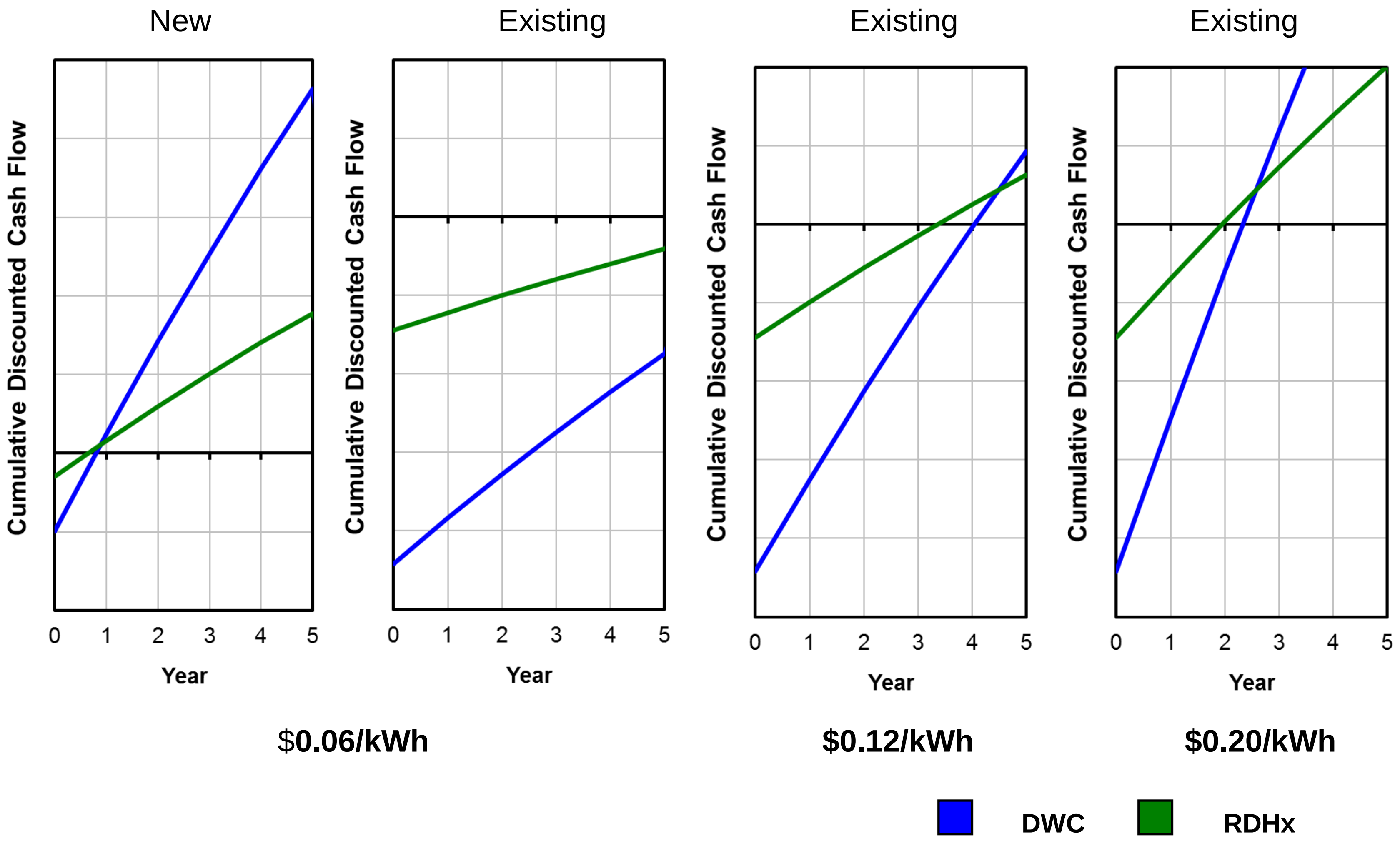
Technology Selection for an Existing Data Center Installation



Technology Selection for a New Data Center Installation



Payback period for DWC vs Air-Cooled w/RDHx



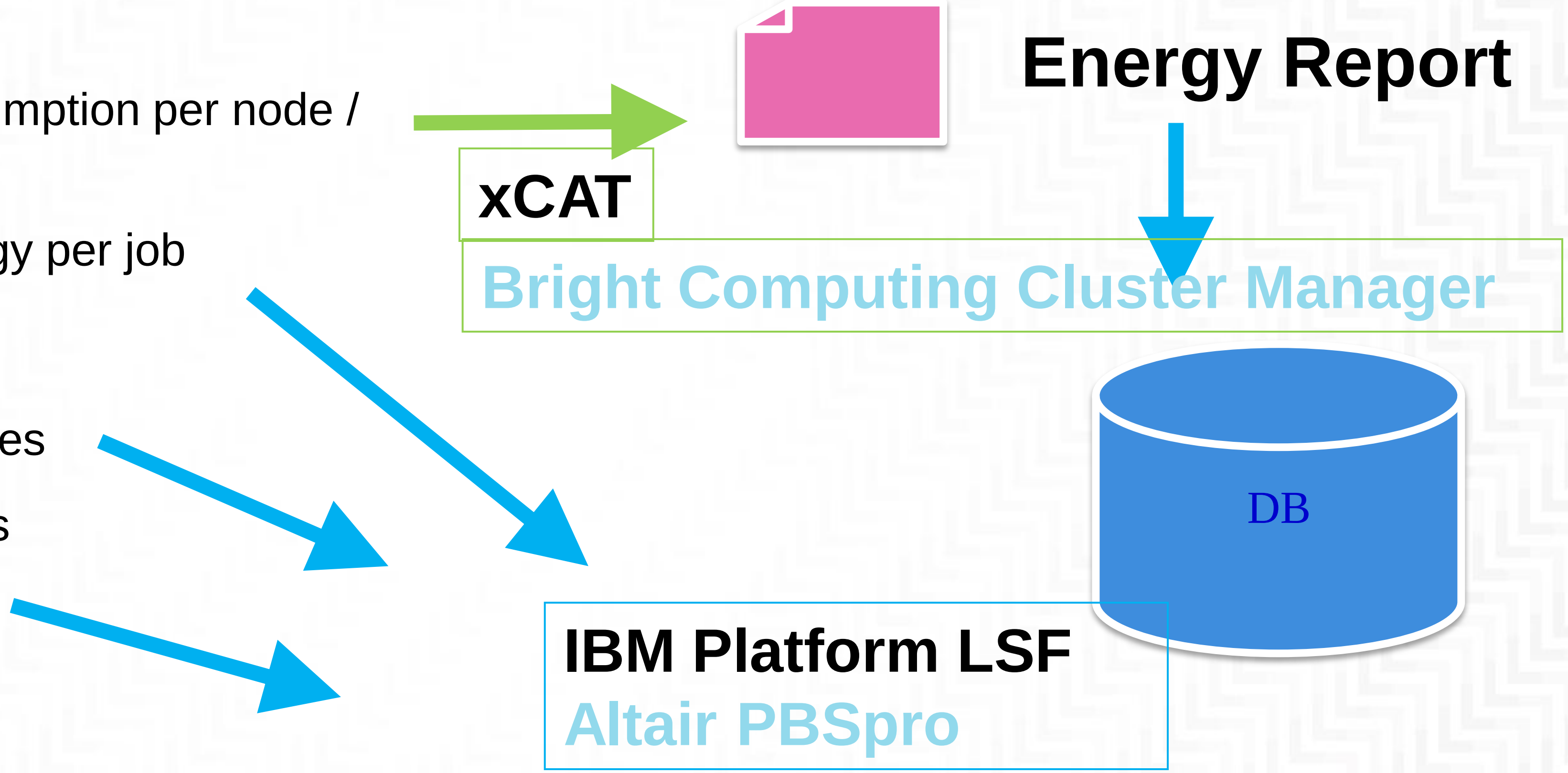
- New data centers: Water cooling has **immediate** payback.
- Existing air-cooled data center payback period strongly depends on electricity rate

How to manage power

- Report
  - temperature and power consumption per node / per chassis
  - power consumption and energy per job
- Optimize
  - Reduce power of inactive nodes
  - Reduce power of active nodes

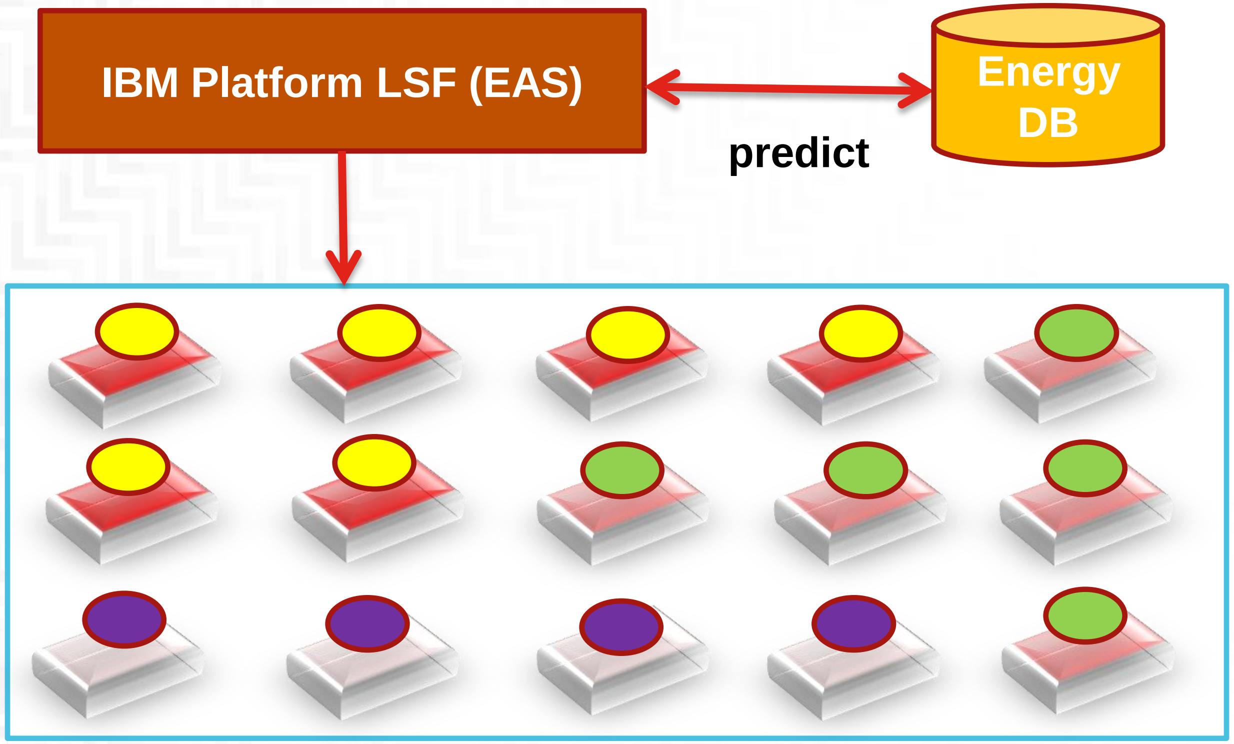
How to manage power

- Report
  - temperature and power consumption per node / per chassis
  - power consumption and energy per job
- Optimize
  - Reduce power of inactive nodes
  - Reduce power of active nodes



How LL/LSF Automatically Select Optimal CPU frequency

- Step I: Learning
  - LSF evaluates the power profile of all nodes
  - calculates coefficients factors
  - save them in the energy database

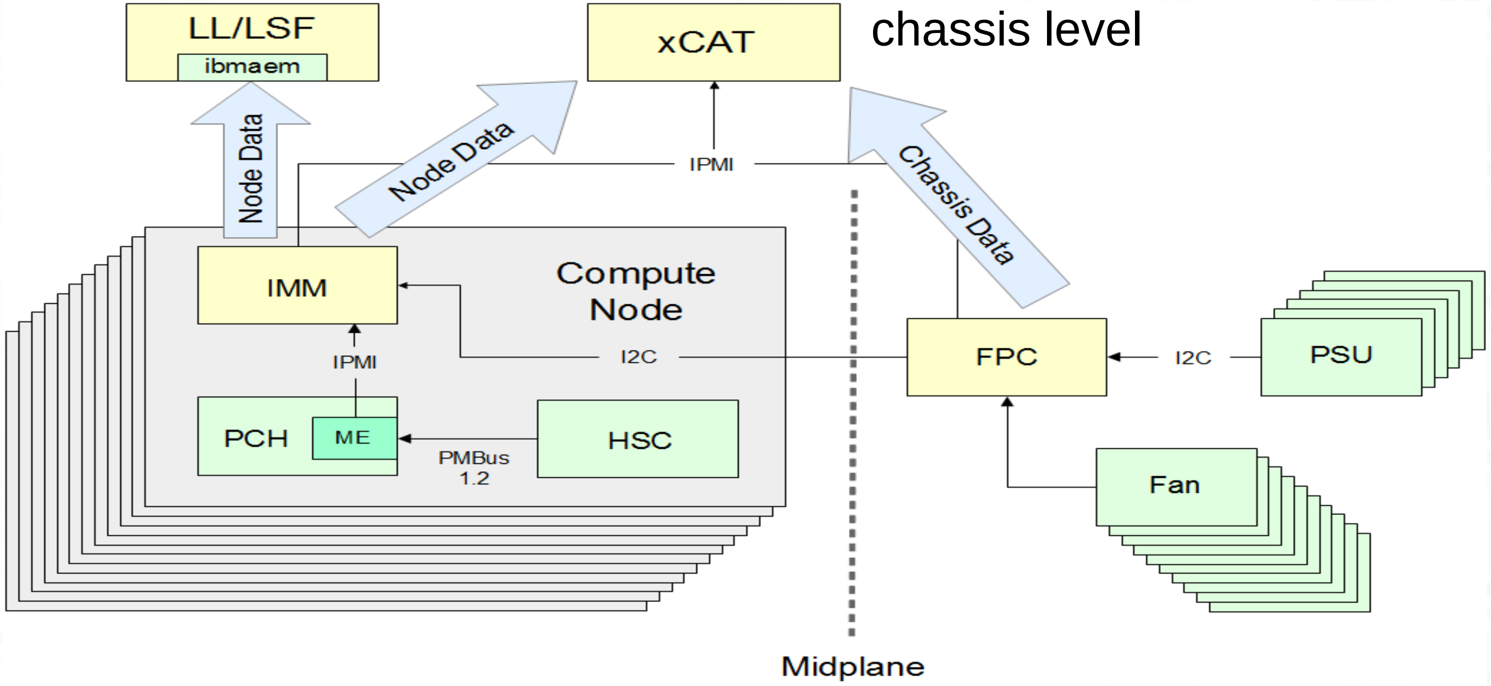


- Step II: Set Default Frequency
  - System administrator defines cluster default cpu frequency (nominal or lower frequency)
- Step III: Tag the job first time
  - User submits the application with a tag
  - runs the job under default frequency
  - LSF collects and reports energy consumption, runtime, performance metrics (cpi, gbs)rs
  - Generates predication result and saves in DB
- Step IV: Use predication
  - User re-submits the same application with the same tag and specifies energy policy
  - LSF selects the optimal cpu frequency for application based on predication result and policy setting.
  - Run the application under selected frequency

Power Management on NeXtScale

- IMM = Integrated Management Module (Node-Level Systems Management)
  - Monitors DC power consumed by node as a whole and by CPU and memory subsystems
  - Monitors inlet air temperature for node
  - Caps DC power consumed by node as a whole
  - Monitors CPU and memory subsystem throttling caused by node-level throttling
  - Enables or disables power savings for node

- FPC = Fan/Power Controller (Chassis-Level Systems Mgmt
  - Monitors AC and DC power consumed by individual power supplies and aggregates to chassis level
  - Monitors DC power consumed by individual fans and aggregates to chassis level



PCH = Platform Controller Hub (i.e., south bridge)
ME = Management Engine (embedded in PCH, runs Intel NM firmware)
HSC = Hot Swap Controller (provides power readings)

xCAT

- Query or set the power capping status – whether or not power capping is currently being enforced (Watt)
- Query or set the power capping value – the permitted max wattage per motherboard (Watt)
- Query the minimum power cap the system can guarantee (Watt)
- Query the cumulative kWh used per motherboard – AC & DC (kWh)
- Query from the PSU the recent average AC wattage used (Watt)
- Query recent histogram data of wattage usage – how many 1 second intervals it has operated in each wattage range (Watt)
- Query current ambient temperature and CPU exhaust temperature (Centigrade)
- S3 Suspend and Resume (on given configuration)

Thank You