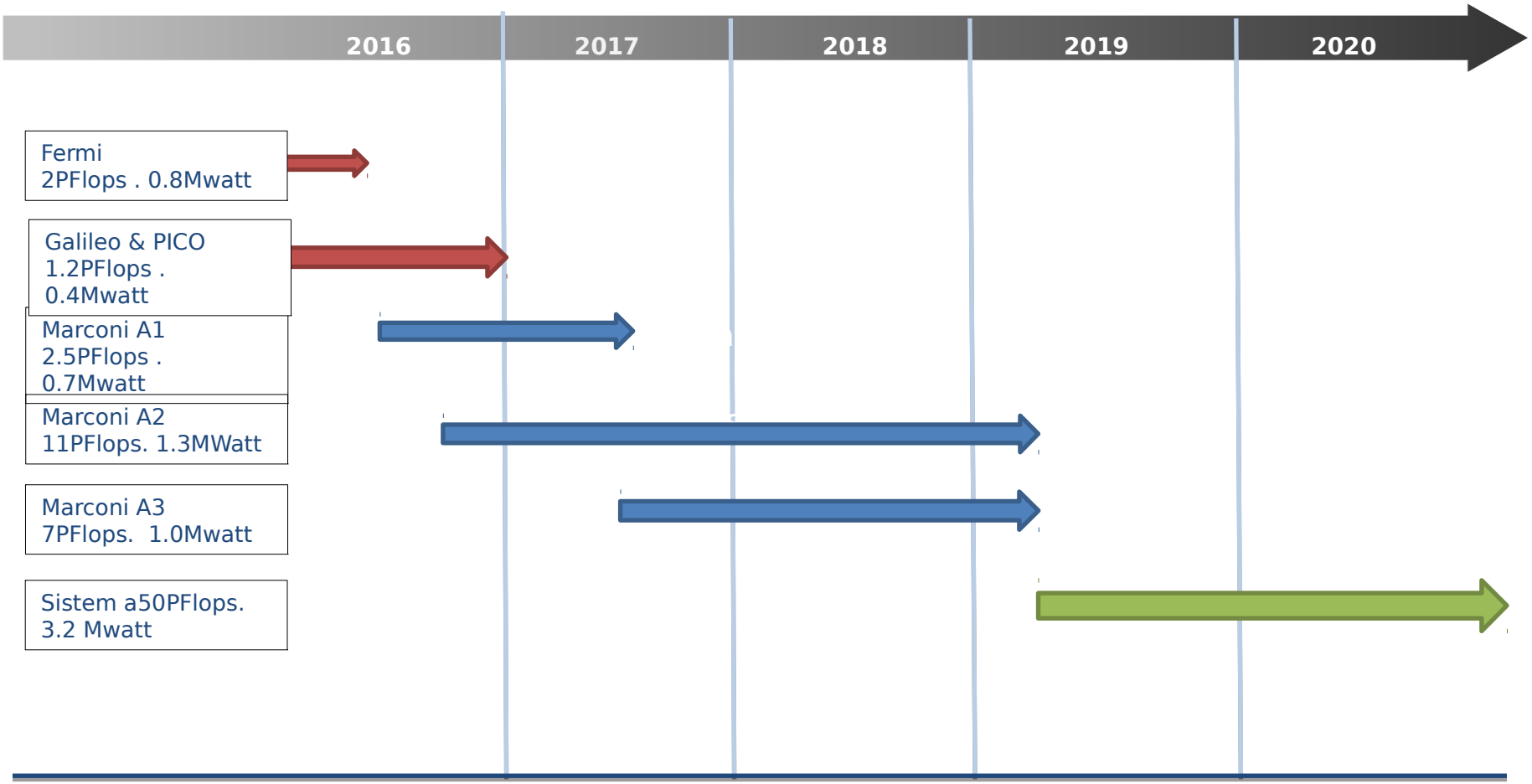
CINECA SCAI
SuperComputing Applications and Innovation

Introduction to Marconi architecture

Carlo Cavazzoni
Nico Sanna

CINECA

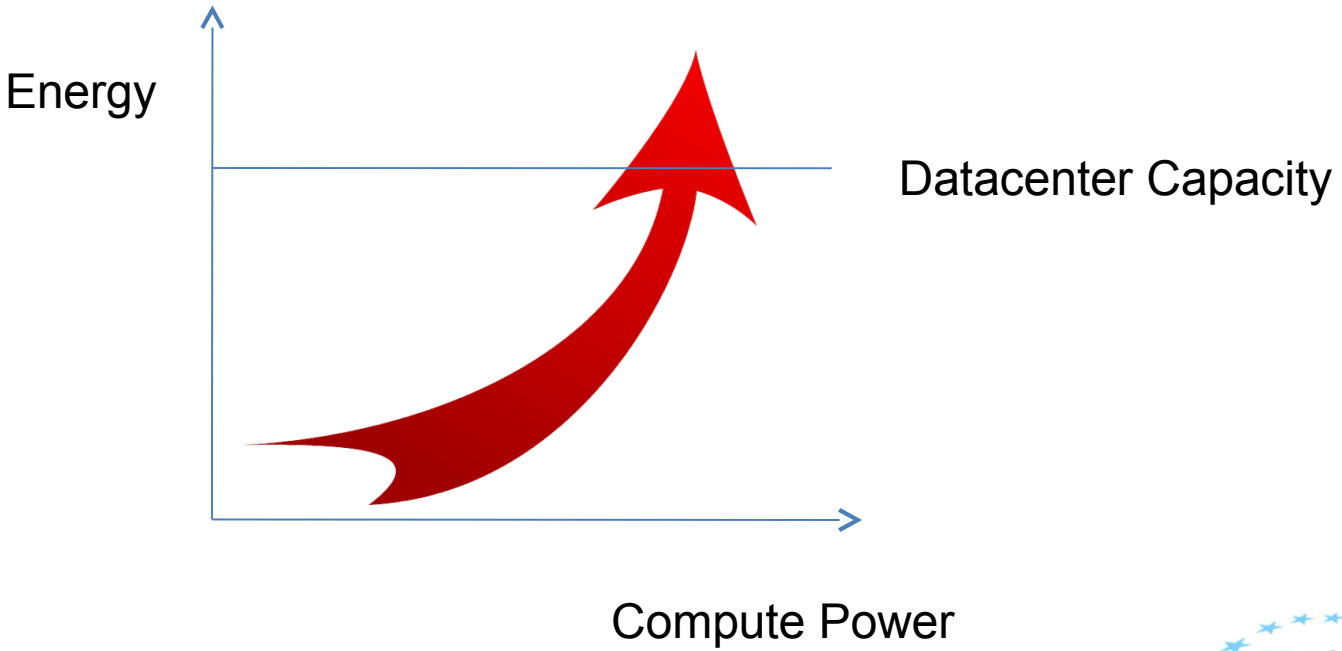


1.2Mwatt	2.4Mwatt	2.3Mwatt	2.3Mwatt	3.2Mwatt
50 rack	120 rack	120 rack	120 rack	150 rack
100mq	240mq	240mq	240mq	300mq

Energy trends

"traditional" RISK and CISC chips are designed for maximum performance for all possible workloads

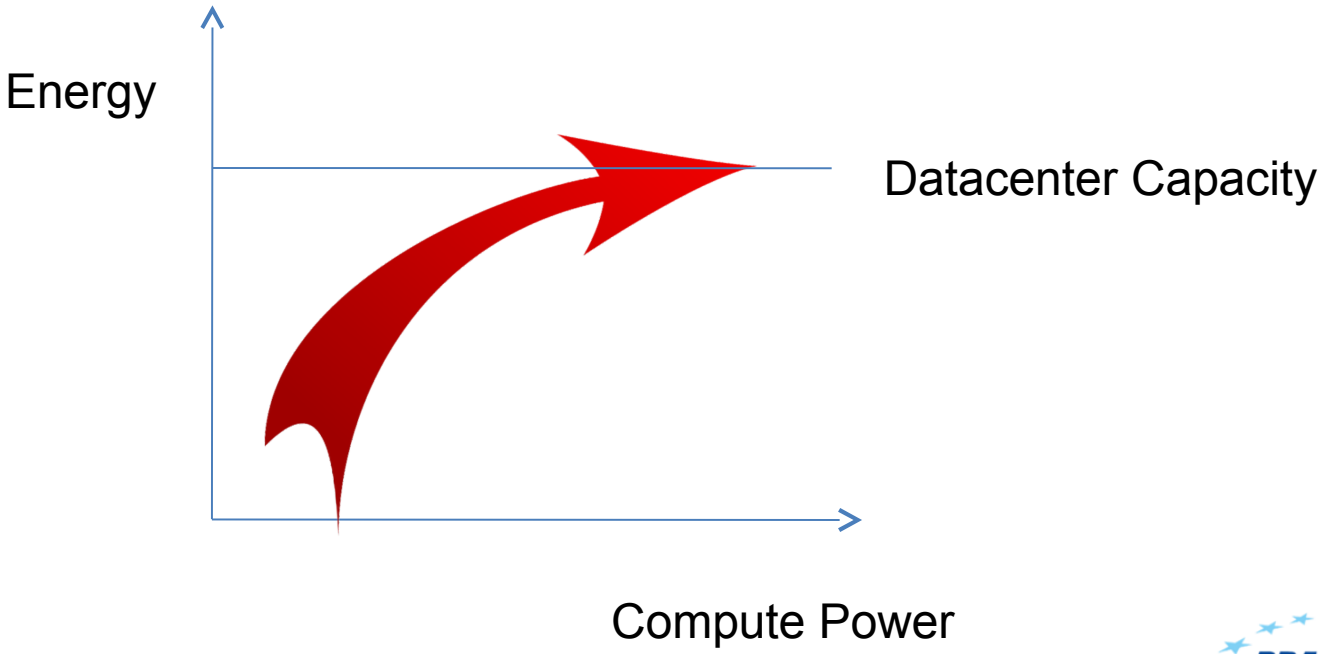
A lot of silicon to maximize single thread performace



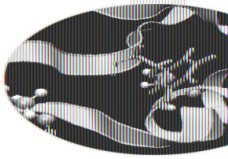
Change of paradigm

New chips designed for maximum performance in a small set of workloads

Simple functional units, poor single thread performance, but maximum throughput



High level system Characteristics



Tender proposal

Partition	Installation	CPU	# nodes	# of Racks	Power
A1 – Broadwell (2.1PFlops)	April 2016	E5-2697 v4	1512	25	700KW
A2 - Knight Landing (11 Pflops)	September 2016	KNL	3600	50	1300KW
A3 – Skylake (4.5PFlops)	June 2017	E5-2680 v5	1512	25	700KW

Network: Intel OmniPath

Marconi - Compute

Partizione A1

- 1512 Lenovo NeXtScale Server -> 2PFlops
- processore Intel E5-2697 v4 Broadwell
- 18 cores @ 2.3GHz.
- dual socket node: 36 core e 128GByte / nodo

Partizione A2

- 3600 server Intel AdamPass -> 11PFlops
- processore Intel PHI code name Knight Landing
- 68 cores @ 1.4GHz.
- single socket node: 96GByte DDR4 + 16GByte MCDRAM

Partizione A3

- 1512 Lenovo Stark Server -> 4.5PFlops
- processore Intel E5-2680v5 SkyLake
- 20 cores @ 2.??GHz
- dual socket node: 40 core e 196GByte /nodo

Marconi - Network

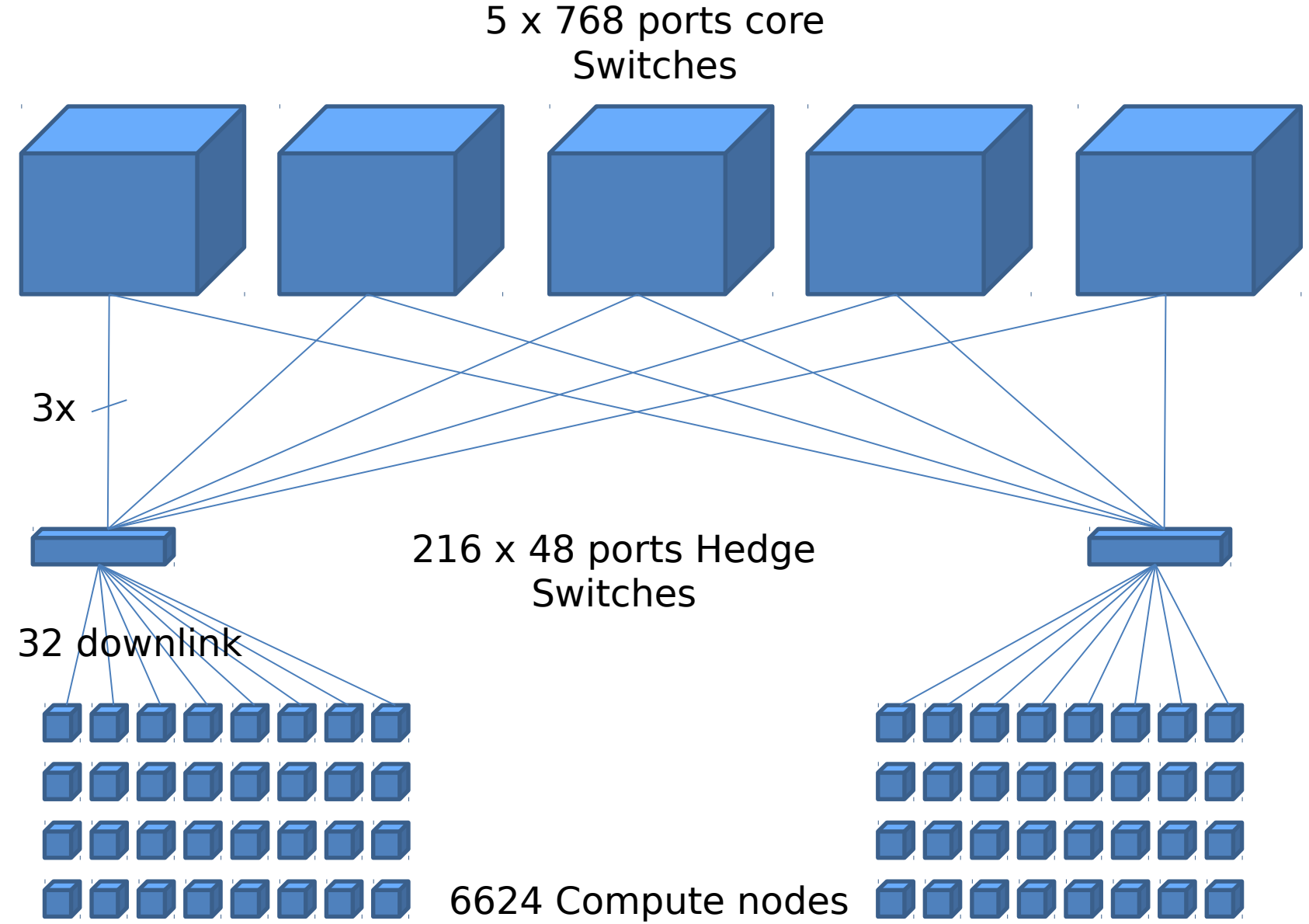
Network type: new Intel Omnipath
Largest Omnipath cluster of the world

Network topology: Fat-tree
2:1 oversubscription
tapering at the level of the core switches only

Core Switches: 5 x OPA Core Switch “Sawtooth Forest”
768 ports each

Hdge Switch: 216 OPA Edge Switch “Eldorado Forest”
48 ports each

Maximum system configuration:
5(opa) x 768(ports) x 2(tapering) -> 7680 servers



32 nodes
fully interconnected island

System layout

- CINECA Floor Plan

System A1:

- Mgmt (1x)
- Storage-Nodes (1x)
- GSS (4 x)
- OPA (5 x)
- BDW (21 x)

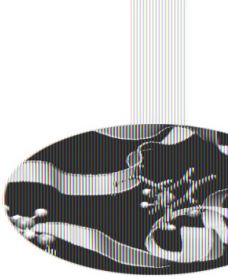
System A2:

- KNL (51 x)

System A3:

- SKL (21 x)

Using SPMD on Marconi – Phase I

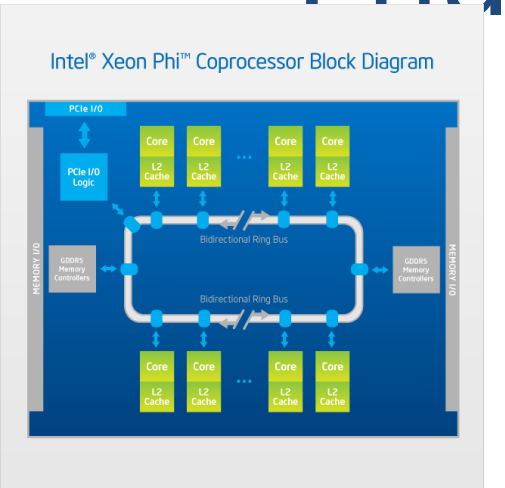


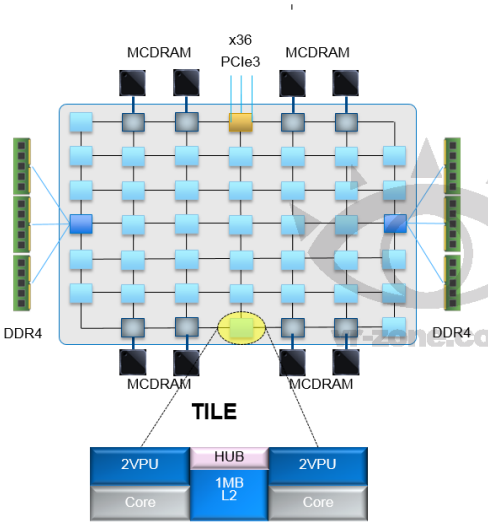
- **Phase 1: Broadwell nodes**
  - Similar to Haswell cores present on Galileo.
  - Expect only a small difference in single core performance wrt Galileo, but a big difference compared to Fermi.
  - More cores/node (36) should mean better OpenMP performance (e.g. for Gromacs) , but also MPI performance will improve (faster network).
  - Life much easier for SPMD programming models.

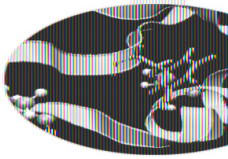
cores/node	36
Memory/node	128 GB

Using MD on Marconi – Phase II

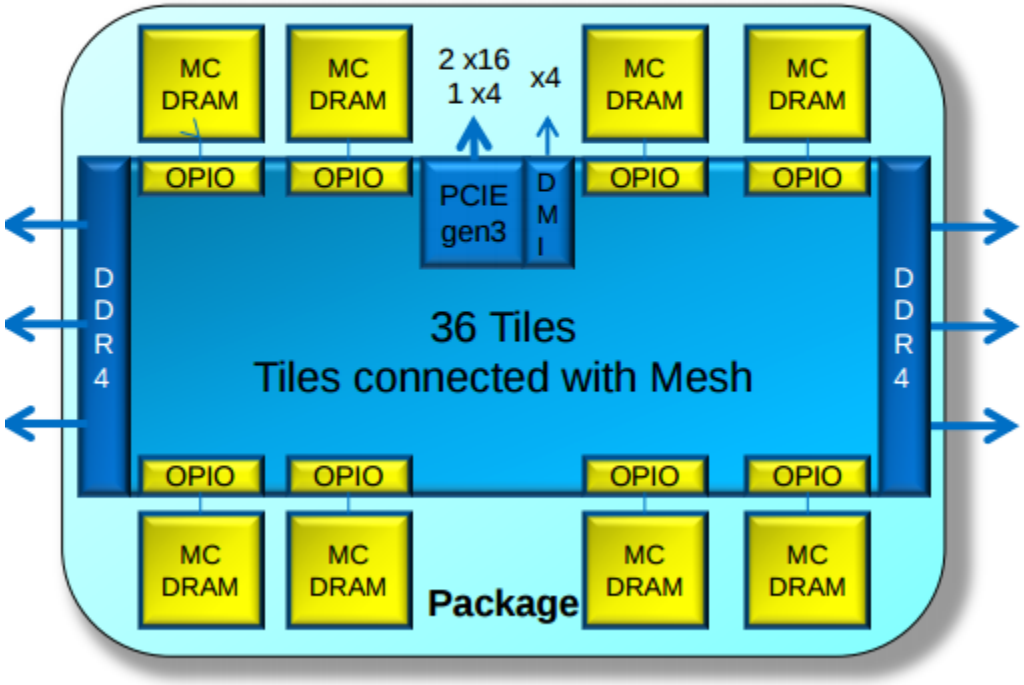
- **Phase 2: Knights Landing (KNL)**
  - A big unknown because very few people currently have access to KNL.
  - But we know the architecture of KNL and the differences and similarities with







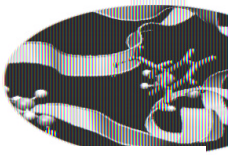
Knights Landing Overview



- Stand-alone, Self-boot CPU
- Up to 72 new Silvermont-based cores
- 4 Threads per core. 2 AVX 512 vector units
- Binary Compatible[1] with Intel® Xeon® processor
- 2-dimensional Mesh on-die interconnect

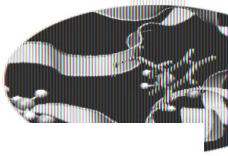
- MCDRAM: On-Package memory: 400+ GB/s of BW[2]
- DDR memory
- Intel® Omni-path Fabric

- 3+ TFLops (DP) peak per package
- ~3x ST performance over KNC



Many Trailblazing Improvements in KNL

Improvements	What/Why
Self Boot Processor	No PCIe bottleneck
Binary Compatibility with Xeon	Runs all legacy software. No recompilation.
New Core: SLM based	~3x higher ST performance over KNC
Improved Vector density	3+ TFLOPS (DP) peak per chip
AVX 512 ISA	New 512-bit Vector ISA with Masks
Scatter/Gather Engine	Hardware support for gather and scatter
New memory technology: MCDRAM + DDR	Large High Bandwidth Memory → MCDRAM Huge bulk memory → DDR
New on-die interconnect: Mesh	High BW connection between cores and memory



Intel® AVX Technology

AVX	AVX2
256-bit basic FP	Float16 (IVB 2012)
16 registers	256-bit FP FMA
NDS (and AVX128)	256-bit integer
Improved blend	PERMD
MASKMOV	Gather
Implicit unaligned	

SNB HSW

AVX-512

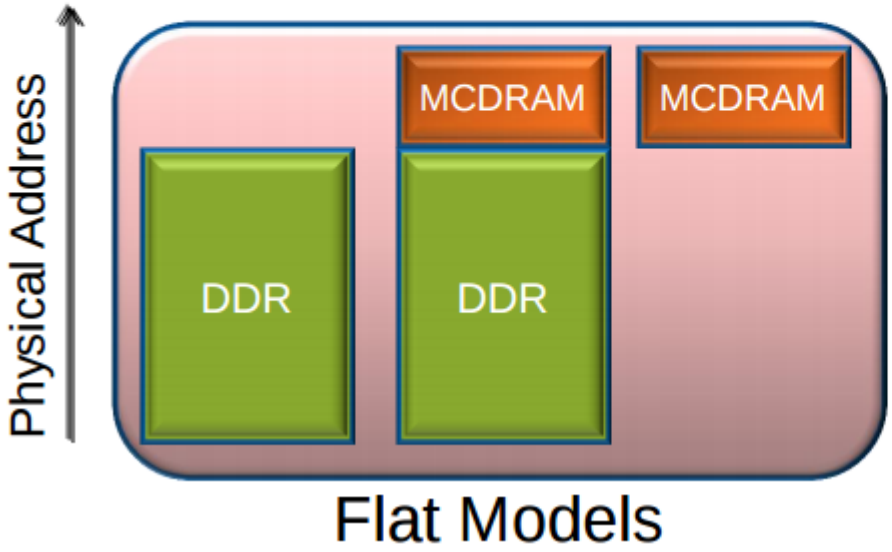
- 512-bit FP/Integer
- 32 registers
- 8 mask registers
- Embedded rounding
- Embedded broadcast
- Scalar/SSE/AVX "promotions"
- HPC additions
- Gather/Scatter

New!

KNL

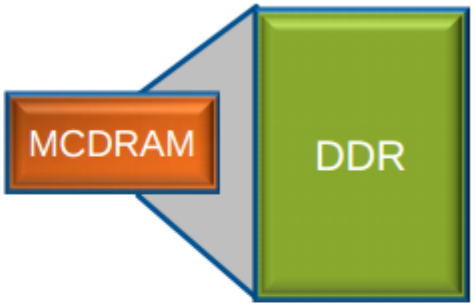
3 Memory Modes

- Mode selected at boot
- MCDRAM-Cache covers all DDR

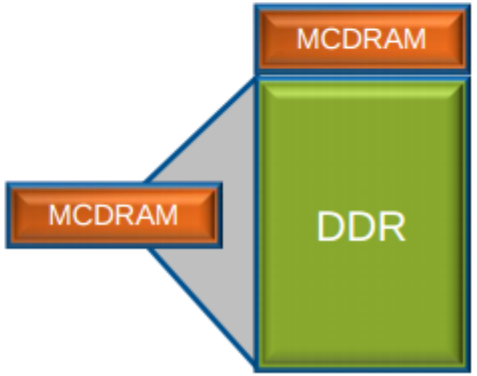


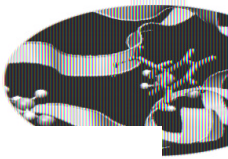
Flat Models

Cache Model



Hybrid Model

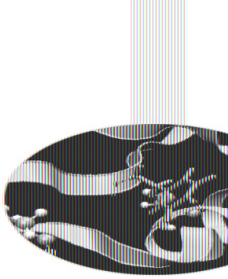




Summary

- Knights Landing (KNL) is the first self-boot Intel® Xeon Phi™ processor
- Many improvements for performance and programmability
  - Significant leap in scalar and vector performance
  - Significant increase in memory bandwidth and capacity
  - Binary compatible with Intel® Xeon® processor
- Common programming models between Intel® Xeon® processor and Intel® Xeon Phi™ processor
- KNL offers immense amount of parallelism (both data and thread)
  - Future trend is further increase in parallelism for both Intel® Xeon® processor and Intel® Xeon Phi™ processor
  - Developers need to prepare software to extract full benefits from this trend

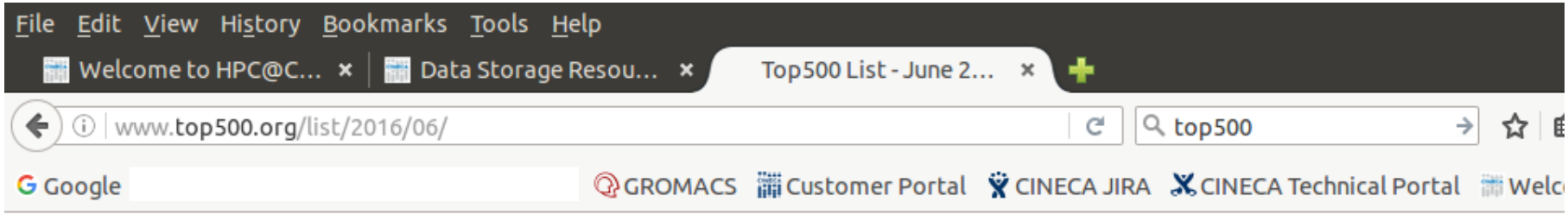
Xeon Phi KNC-KNL comparision



	KNC (Galileo)	KNL (Marconi)
#cores	61 (pentium)	68 (Atom)
Core frequency	1.238 GHz	1.4 Ghz
Memory	16GB GDDR5	96GB DDR4 +16Gb MCDRAM
Internal network	Bi-directional Ring	Mesh
Vectorisation	512 bit /core	2xAVX-512 /core
Usage	Co-processor	Standalone
Performance (Gflops)	1208 (dp)/2416 (sp)	~3000 (dp)
Power	~300W	~200W

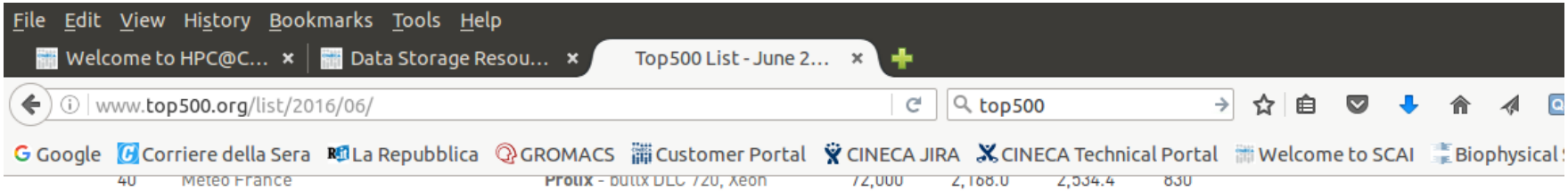
A KNC core can be 10x slower than a Haswell core. A KNL core is expected to be 2-3X slower. Big differences also in memory bandwidth.

Top500 List: Marconi – A1



RANK	SITE	SYSTEM	CORES	(TFLOP/S)	(TFLOP/S)	(KW)
1	National Supercomputing Center in Wuxi China	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway NRCPC	10,649,600	93,014.6	125,435.9	15,371
2	National Super Computer Center in Guangzhou China	**Tianhe-2 (MilkyWay-2)** - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31S1P NUDT	3,120,000	33,862.7	54,902.4	17,808
3	DOE/SC/Oak Ridge National Laboratory United States	**Titan** - Cray XK7 , Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x Cray Inc.	560,640	17,590.0	27,112.5	8,209
4	DOE/NNSA/LLNL United States	**Sequoia** - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom IBM	1,572,864	17,173.2	20,132.7	7,890
5	RIKEN Advanced Institute for Computational Science (AICS) Japan	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect Fujitsu	705,024	10,510.0	11,280.4	12,660
6	DOE/SC/Argonne National Laboratory United States	**Mira** - BlueGene/Q, Power BQC 16C 1.60GHz, Custom IBM	786,432	8,586.6	10,066.3	3,945

Top500 List: Marconi – A1



40	Meteo France France	**Prolix** - bullx DLC 720, Xeon E5-2698v4 20C 2.2GHz, Infiniband FDR Bull, Atos Group	72,000	2,168.0	2,534.4	830
41	Moscow State University - Research Computing Center Russia	**Lomonosov 2** - T-Platform A-Class Cluster, Xeon E5-2697v3 14C 2.6GHz, Infiniband FDR, Nvidia K40m T-Platforms	42,688	2,102.0	2,962.3	1,079
42	LvLiang Cloud Computing Center China	**Tianhe-2 LvLiang Solution** - Tianhe-2 LvLiang, Intel Xeon E5-2692v2 12C 2.2GHz, TH Express-2, Intel Xeon Phi 31S1P NUDT	174,720	2,071.4	3,074.5	997
43	Japan Atomic Energy Agency (JAEA) Japan	SGI ICE X, Xeon E5-2680v3 12C 2.5GHz, Infiniband FDR SGI	60,240	1,929.4	2,409.6	
44	Commissariat a l'Energie Atomique (CEA) France	**Tera-1000-1** - bullx DLC 720, Xeon E5-2698v3 16C 2.3GHz, Infiniband FDR Bull, Atos Group	70,272	1,871.0	2,586.0	1,042
45	CINECA Italy	**Fermi** - BlueGene/Q, Power BQC 16C 1.60GHz, Custom IBM	163,840	1,788.9	2,097.2	822
46	CINECA Italy	**Marconi** - Lenovo NeXtScale nx360M5, Xeon E5-2697v4 18C 2.3GHz, Omni-Path Lenovo	54,432	1,723.9	2,003.1	
47	Government United States	**SwiftLucy** - Cluster Platform 3000 BL460c Gen9, Xeon E5-2680v3 12C 2.5GHz, Infiniband FDR	57,600	1,703.3	2,304.0	

Marconi – A1 HPL

Single node Linpack:

1 MPI task, 36 threads
perf range: 1.19 - 1.3TFlops
N = 104832, NB = 192 (90GByte)

Full system Linpack:

1 MPI task per node
perf range: 1.6 – 1.7PFs. Max Perf: 1.72389PFs
with Turbo-OFF.

Turbo-ON -> throttling

```
================================================================================
T/V                N    NB     P     Q               Time                 Gflops
--------------------------------------------------------------------------------
WC06C2C4     4320000   192    30    50           31178.23             1.72389e+06
HPL_pdgesv() start time Mon May 30 16:43:07 2016

HPL_pdgesv() end time   Tue May 31 01:22:46 2016

--------------------------------------------------------------------------------
||Ax-b||_oo/(eps*(||A||_oo*||x||_oo+||b||_oo)*N)=        0.0007856 ...... PASSED
================================================================================

Finished      1 tests with the following results:
              1 tests completed and passed residual checks,
              0 tests completed and failed residual checks,
              0 tests skipped because of illegal input values.
--------------------------------------------------------------------------------

End of Tests.
================================================================================
```

Marconi: Intel E5-2697 v4 Broadwell, 18 cores @ 2.3GHz.

Marconi – Phase III

- **Phase 3. Intel Skylake processors (mid-2017)**
  - Successor to Haswell, and launched in 2015.
  - Expect increase in performance and power efficiency.

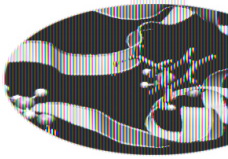
Marconi – Phase III

Tick-Tock Development Model:
Sustained Microprocessor Leadership

Intel® Microarchitecture Codename Nehalem		Intel® Microarchitecture Codename Sandy Bridge		Intel® Microarchitecture Codename Haswell		Intel® Microarchitecture Codename Skylake	
Nehalem 45nm New Micro-architecture	Westmere 32nm New Process Technology	Sandy Bridge 32nm New Micro-architecture	Ivy Bridge 22nm New Process Technology	Haswell 22nm New Micro-architecture	Broadwell 14nm New Process Technology	Skylake 14nm New Micro-architecture	Future Product
Tock	Tick	Tock	Tick	Tock	Tick	Tock	Tick

Innovation delivers new microarchitecture with Skylake

Marconi – Phase III



Purley Roadmap Positioning

Shipping | In Development | In Planning

2015 | 2016 | 2017

	2015	2016	2017
Expandable	Brickland Platform Intel® Xeon® processor E7-8800/4800/2800 v2 product families/Intel® C602J chipset, Intel® C104/C102 Scalable Memory Buffer	Brickland Platform Intel® Xeon® processor E7-8800/4800 v3 product families/Intel® C602J chipset, Intel® C114/C112 Scalable Memory Buffer	Brickland Platform Intel® Xeon® processor E7-8800/4800 v4 product families/Intel® C602J chipset, Intel® C114/C112 Scalable Memory Buffer
Efficient Performance 4S/Performance Comms	Romley-EP 4S Platform Intel® Xeon® processor E5-4600 v2 product family/Intel® C600 series chipset	Grantley-EP 4S Platform Intel® Xeon® processor E5-4600 v3 product family/Intel® C610 series chipset	Grantley-EP 4S Platform Intel® Xeon® processor E5-4600 v4 product family/Intel® C610 series chipset
Efficient Performance 2S/Performance & Mid-Range Storage/Performance Comms	Grantley-EP 2S Platform Intel® Xeon® processor E5-2600 v3 product family Intel® C610 series chipset Intel® Ethernet XL710 controller	Grantley-EP 2S Platform Intel® Xeon® processor E5-2600 v4 product family/Intel® C610 series chipset Intel® Ethernet XL710 controller Intel® Ethernet X550 controller (Sageville)	Scalable Performance 2S, 4S, 8S+ Purley Platform Intel® Xeon® processor product family Skylake Server CPU Intel® Omni-Path Architecture (Storm Lake Gen 1) Lewisburg PCH Apache Pass Intel® Ethernet XL710 controller Intel® Ethernet X550 controller (Sageville)
Optimized for Comms Performance	Intel® 8900 Chipset option w/ QuickAssist Acceleration (River Forest)	Intel® 8900 Chipset option w/ QuickAssist Acceleration (River Forest)	
Intel® Xeon Phi™	Intel® Xeon Phi™ x100 Product Family Intel® Xeon Phi™ coprocessor 7100 / 5100 / 3100 family	Intel® Xeon Phi™ x200 Product Family Intel® Xeon Phi™ coprocessor (codenamed Knights Landing)	
		Groveport Platform Intel® Xeon Phi™ x200 processor with (optional) integrated Fabric	
1S Workstation	Grantley-EP 1S Workstation Platform Intel® Xeon® processor E5-1600 v3 product family Intel® C610 series chipset Intel® Ethernet XL710 controller	Grantley-EP 1S Workstation Platform Intel® Xeon® processor E5-1600 v4 product family/Intel® C610 series chipset Intel® Ethernet XL710 controller	Basin Falls 1S Workstation Platform Intel® Xeon® processor product family (Skylake Server socket R) Kaby Lake PCH

Product placement not representative of final launch date within the specified quarter. For more details refer to ILU and 5 Quarter Roadmap.

Marconi – Phase III

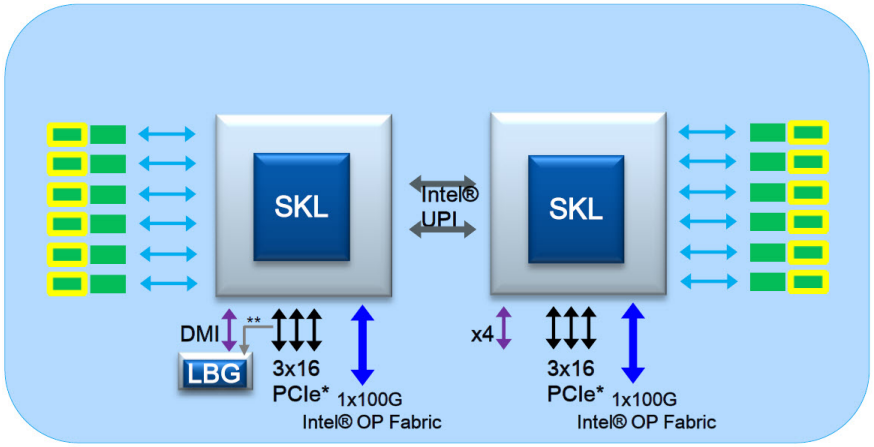
Platform Comparisons

Spec	Grantley with Broadwell-EP CPU	Brickland with Broadwell-EX CPU	Purley with Skylake CPU
CPU TDP (with IVR)	55-145W, 160W WS only	115-165W	45-165W
Socket	Socket R3	Socket R1	Socket P
Scalability	2S	2S, 4S, 8S	2S, 4S, 8S
Cores	Up to 22C with Intel® HT Technology	Up to 24C with Intel® HT Technology	Up to 28C with Intel® HT Technology
Memory	4 channels DDR4 per CPU RDIMM, LRDIMM	4 channels DDR4 per CPU RDIMM, LRDIMM	6 channels DDR4 per CPU RDIMM, LRDIMM
	1DPC=up to 2400, 2DPC= up to 2133, 3DPC=up to 1600	DDR3/4 Performance Mode 1333, 1600 DDR3/4 Lockstep mode 1333, 1600, 1866	2133, 2400 2DPC, 2666 1DPC No 3 DPC support
UPI	QPI: 2 v1.1 channels per CPU 9.6 GT/s max	QPI: 3 v1.1 channels per CPU 9.6 GT/s max	UPI: 2-3 channels per CPU (9.6, 10.4 GT/s)
PCIe*	PCIe* 3.0 (2.5, 5.0, 8.0 GT/s)	PCIe* 3.0 (2.5, 5.0, 8.0 GT/s)	PCIe* 3.0 (2.5, 5.0, 8.0 GT/s)
	40 lanes per CPU	32 lanes per CPU	48 lanes per CPU Bifurcation support: x16, x8, x4
PCH	Wellsburg: DMI2 – 4 lanes; Up to 6xUSB3, 8x USB2 ports, 10xSATA3 ports; GbE MAC (+ External PHY)	Patsburg: 14 USB2 ports, 4 SATA2 ports, 2 SATA3 ports	Lewisburg: DMI3 – 4 lanes; 14xUSB2 ports Up to: 10xUSB3; 14xSATA3, 20xPCIe*3 New: Innovation Engine, 4x10GbE ports, Intel® QuickAssist Technology
External Node Controller Support	None	3rd Party Node Controller	3rd Party Node Controller supported on select skus

Marconi – Phase III

2S Purley Platform Configuration Example

Typical 2S Configuration



DDR4 DIMMs
DDR4/Apache Pass
** Optional PCIe* uplink connection for Intel® QuickAssist Technology and Intel® Ethernet
Example DIMM population shown; please look up Apache Pass customer collateral for specific rules on DDR4/Apache Pass DIMM populations

Platform Ingredient Options

Fabric: Storm Lake Integrated, Storm Lake PCIe* card

Storage: SATA: Downieville, SATA: Youngsville, PCIe*: Coldstream, PCIe* Cliffdale

Software: Analytics, Efficiency, Performance, Secure Access, Tools

Networking: Lewisburg as 4x10GbE Integrated Network Solution with PHY, Fortville 4x10/2x40GbE (Controller), Sageville 2x10GBASE-T (Controller), Springville (1x1GBASE-T), Powerville (4x1GBASE-T)

Core Ingredients: Coppervale 10GBASE-T, Jacksonville GbE PHY

Accelerators: Intel® QuickAssist Accelerator Technology with Compression and Encryption, Lewisburg can also be used as PCIe* add-in card in end point mode

Intel® Xeon Phi™ Product Family: Knights Corner/Landing Coprocessors and Processors

FW/Bios: Manageability, Node Manager, Intel® RSTe, ME11, Intel® Trusted Execution Technology and Intel® Platform Protection Technology with Boot Guard