



Italian National Agency for New Technologies,
Energy and Sustainable Economic Development



EUROfusion



SuperComputing Applications and Innovation



MARCONI FUSION: The new High Performance Computing facility for European nuclear fusion modelling

11th IAEA Technical Meeting on Control, Data Acquisition and Remote Participation for Fusion Research 8-12 May 2017, Greifswald, Germany

F.Iannone, G.Bracco, C.Cavazzoni, R.Coelho, D. Coster, O. Hoenen, A.Maslennikov, S.Migliori, M.Owsiak, A.Quintiliani, B.Palak, V.Pais, F.Robin, E.Rossi, I.Voitsekhovitch



1101 0110 1100
0101 0010 1101
0001 0110 1110
1101 0010 1101
1111 1010 0000



OUTLINE

HPC for FUSION: a brief history

MARCONI: the Tier 0 HPC system at CINECA

Technical description

Benchmarks

Intel Xeon Phi: Knights Landing

Operations & Production environments

Gateway

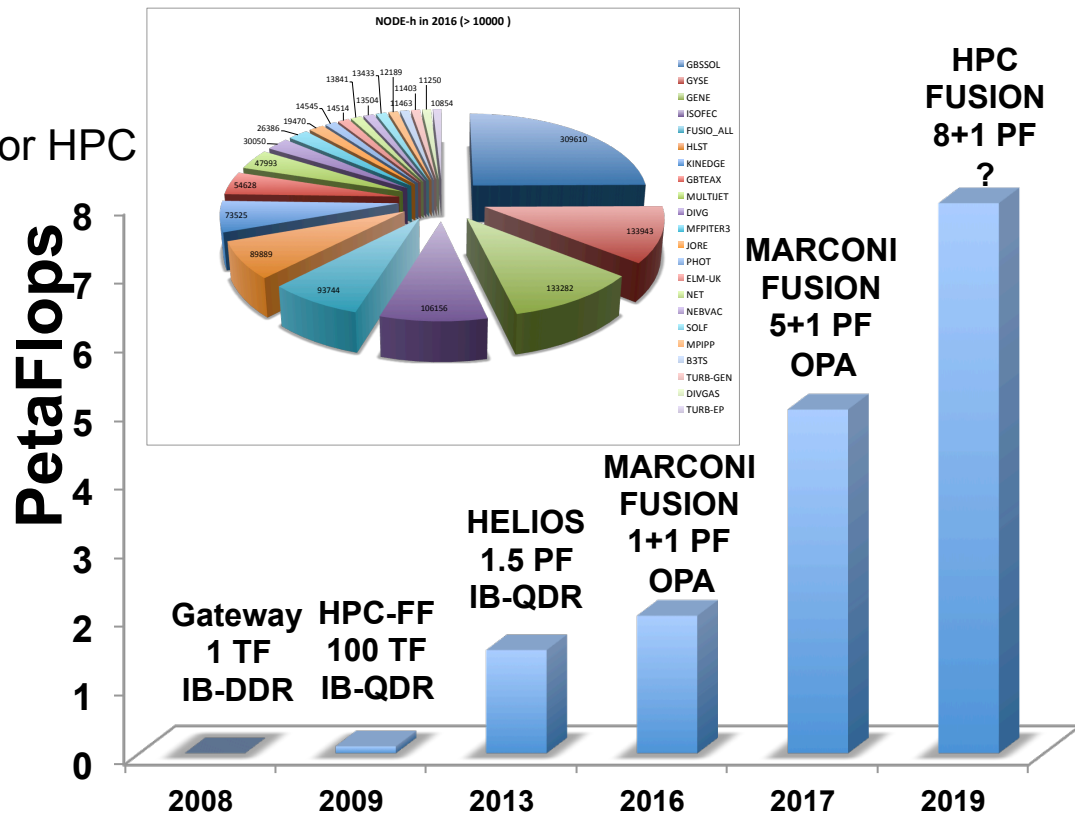
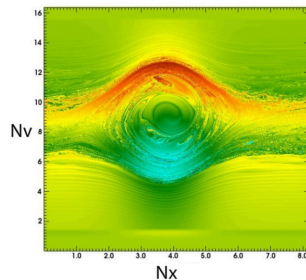
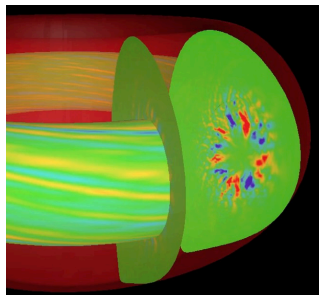
Parallel Programming Model : a case study

Conclusions

HPC FOR European Nuclear Fusion Research

History, Current, Roadmap

- EUROFusion Roadmap: resources for HPC
- Plasma Physic simulations:
 - Turbulence & Transport processes
 - Fast particle, MHD, edge physics
 - Heating & Current Drive
 - Integrated Tokamak Modelling
- Reactor materials & technologies



HPC system dedicated to fusion research in EU (2015)

The EUROfusion requests

1. STEP 1: 4 Pflops Jan.2017 to Dec.2018
2. STEP 2: 8 Pflops Jan.2018 to Dec.2019 (To be Defined)

ENEA & CINECA awarded the STEP 1 From: Jul.2016 to Dec. 2018

The Stackholders

ENEA is member of EUROFusion Consortium and national coordinator of Nuclear Fusion research

CINECA is the national TIER 0 for HPC, PRACE member, Linked Third Party of EUROFusion



F.Iannone et.al.

MARCONI FUSION: The new High Performance Computing facility for European nuclear fusion modelling

MARCONI: THE CINECA'S TIER 0 HPC SYSTEM

General Features

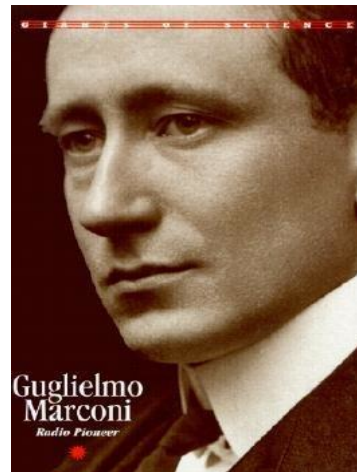
Conventional CPU: Intel Xeon: Broadwell (BDW) & Skylake (SKY)

Accelerator CPU: Intel Xeon Phi: Knights Landing (KNL)

Fabric: Intel Omni-Path / Architecture: LENOVO

TIER 0 HPC system : 18 Pflops , 3 MW power consumption

Installation Roadmap



BDW

A1: 2PFlops

1500 nodes

FUSION: 1PFlops

806 nodes

KNL

A2: 11PFlops

3600 nodes

FUSION: 1PFlops

392 nodes

SKY

A3: 7PFlops

2300 nodes

FUSION: 5PFlops

1550 nodes

Jul.2016

Sep.2016

Jul.2017

Dec.2018

MARCONI A1 : 2 Pflops of Broadwell

	MARCONI A1	MARCONI-FUSION
Model:	Lenovo NextScale Intel OmniPath Cluster	
Architecture:		
Nodes:	1512	806
Processors:	2x18 cores Intel Xeon E5-2697 (BDW) at 2.3 GHz	
Cores:	36 cores/node – Total=54432	Total= 29016
RAM:	128 GB/node , 3.5 GB/core	
Racks:	21 + 1 (mgmt.shared)	~11+1 (mgmt. shared)
Peak Performance:	= 2 cpu × 18 cores × 2.3Hz × 16 fpo × 1512nodes= 2.003 PFlops	= 2 cpu × 18 cores × 2.3Hz × 16 fpo × 806nodes= 1.068 PFlops
Disk Space:	17 PB (raw)	5 PB (raw)
Power:	700 kW	350 kW



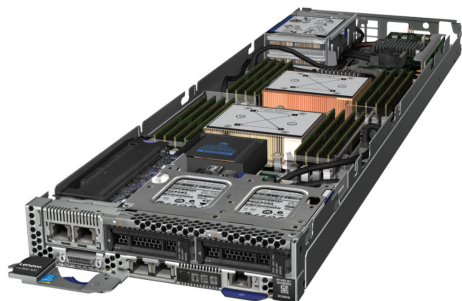
**CINECA CED
MARCONI A1**



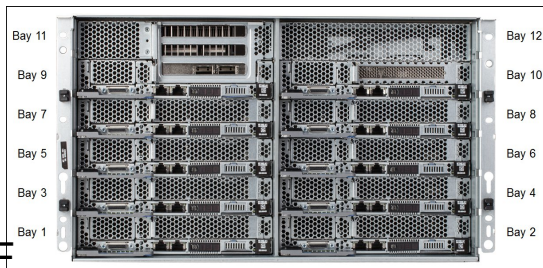
MARCONI A1 : 2 Pflops of Broadwell

Technical Details

- 2 x CPU Intel Xeon E5-2697 v4
18 core @ 2.3 GHz
- 8 x 16GB DIMM DDR4 2400MHz
- 1x120 GB SATA MLC S3500
- 1 x link OPA 100Gb/s, 2 links 1GbE

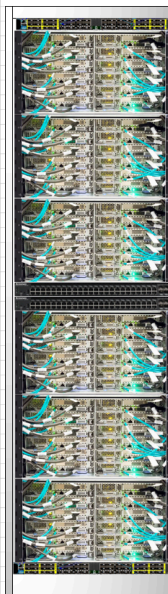


NeXtScale n1200 chassis



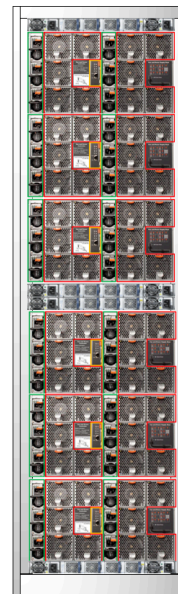
Front View

42 U



Rear View

42 U



1x Omni-path Eldorado Forest
edge switch 100 Gbps 48 ports

LENOVO

6x NextScale N1200
12x NX360-M5 nodes
2x CPU Xeon E5-2697 V4
BDW 18 core 2.3 GHz
72 nodes/rack

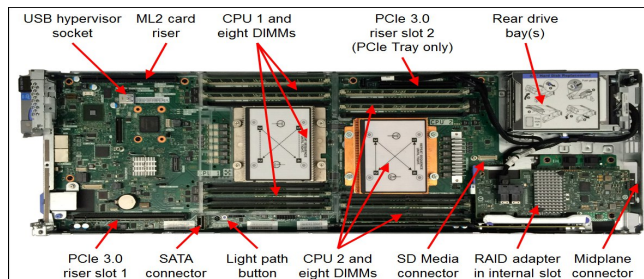
2x LENOVO G8052
GEthernet switch: 1 Gbps 48 ports

6x Power Supplies

10x Fan Modules

Fan & Power Controller

1x Omni-path Eldorado Forest
edge switch 100 Gbps 48 ports



NeXtScale nx360M5

rack

MARCONI : The low latency network

INTEL Omni-Path: a 100 Gbps for Exascale HPC systems

2 Levels, Tapered Fat-Tree (2:1 oversubscription)

5 Director *Sawtooth Forest* Core Switches for the whole MARCONI

MARCONI

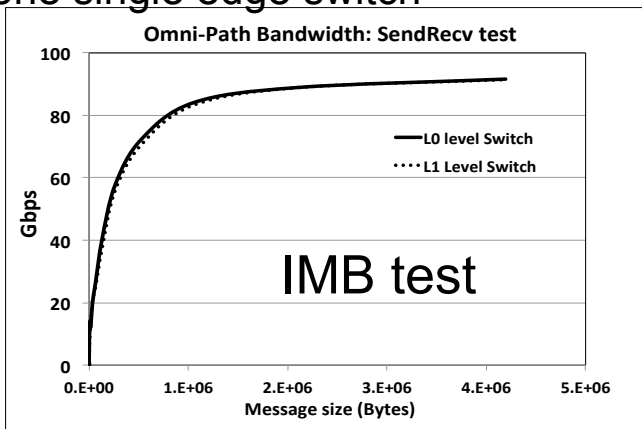
A1:42 A2:216 Edge *Eldorado Forest* switches (48p)

32 nodes on one single edge switch

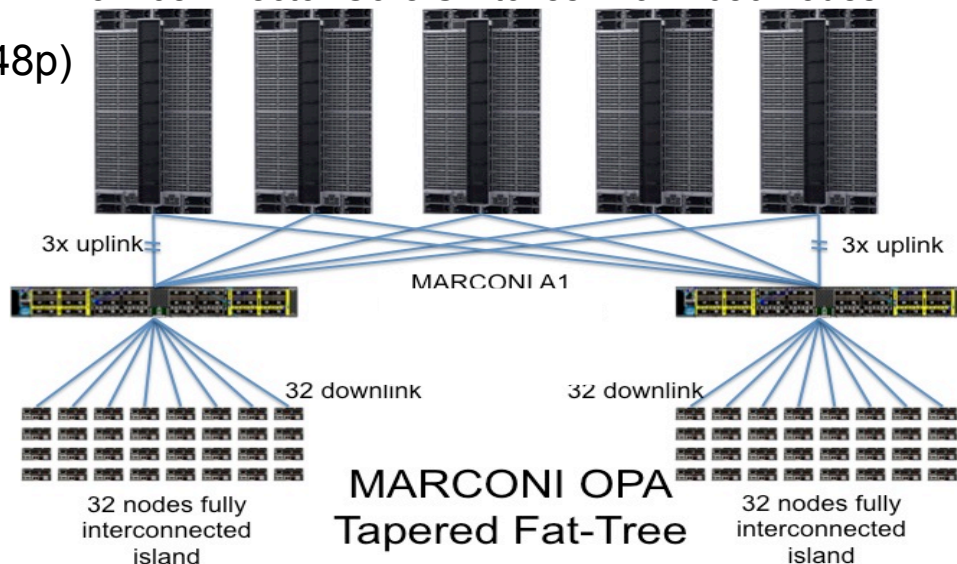
Latency

L0: 1.45 μ s

L1: 2.35 μ s



5x768 Director Core Switches/ max 7680 nodes

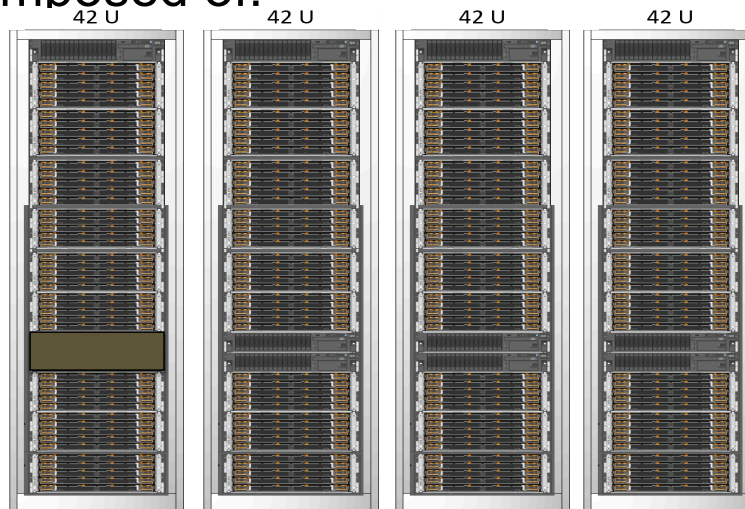


MARCONI: The High Performance Storage

LENOVO GSS (GPFS Storage System)

6 x GSS mod.26 in 4 racks each composed of:

- 2x3650 M4 servers with OPA
- 6 disk enclosures
- 8TB x 348 disk SATA



MARCONI
STORAGE
AREA
17 PB
(raw)

- 6 LENOVO GSS-26
- 2x3650 M4 server
 - 6 disk enclosure
 - 8 TB x 384 disk drives

Raw capacity: **17 PB**

Usable capacity: **12 PB**

Raw capacity for Fusion: **5 PB**

Spectrum Scale GPFS v. 4.2.1

IOR configuration	read max [MB/S]	write max.[MB/S]
124 compute node of 2 different racks	101477.89	97303.87
72 compute node of the same racks	101663.25	97658.31

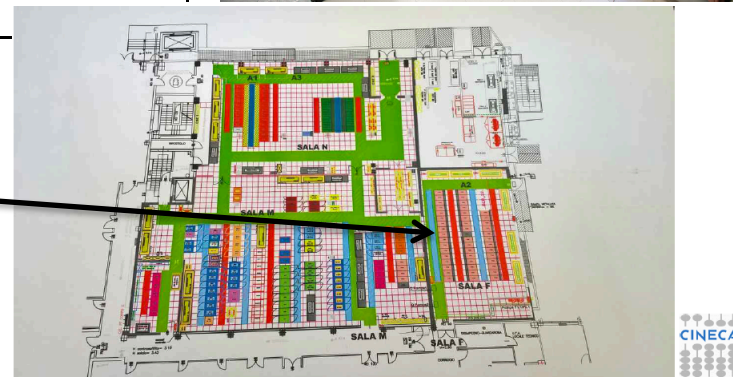
100
GB/s

MARCONI A2: 11 Pflops of Knights Landing

	MARCONI A2	MARCONI-FUSION
Model:	INTEL SERVER BOARD Intel OmniPath Cluster	
Architecture:		
Nodes:	3600	340
Processors:	Intel Knights Landing (KNL) 1 x Xeon Phi 7250 (KNL) at 1.4 GHz	
Cores:	68 cores/node – Total=244800	Total=27200
RAM:	96 GB/node	
Racks:	50 + 1 (mgmt.shared)	~6+1 (mgmt. shared)
Peak Performance:	= 3 TF x 3600 nodes= 10.8 PFlops	= 3TF x 340 nodes= 1.02 PFlops
Disk Space:	17 PB (raw)	5 PB (raw) of MARCONI A1
Power:	1300 kW	120 kW



**CINECA CED
MARCONI A2**



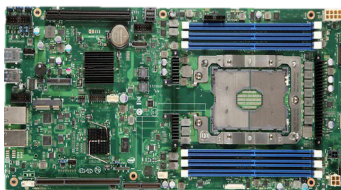
MARCONI A2: 11 Pflops of Knights Landing

Technical details

- 1 x CPU Intel Xeon Phi Knights Landing-D 7250 68 cores @ 1.4 GHz, 272 cores with Hyper-threading
- 6 x 16GB DIMM DDR4 2400MHz
- 1x120 GB SSD DCS3510
- 1 x link OPA 100Gb/s, 2 links 1GbE



Intel Server Board
SF7200AP

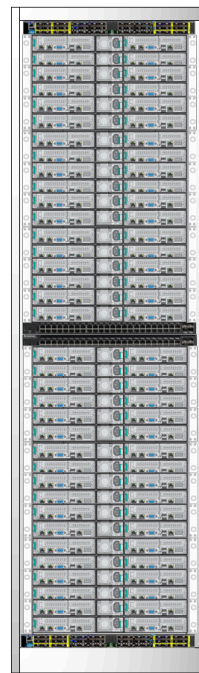


Intel Server Chassis
H2312XXLR2



Rear View

42 U



1x Omni-path Eldorado Forest
edge switch 100 Gbps 48 ports

LENOVO

18 x Intel Server

Chassis H2312XXLR2

4 x Intel Server Board

S7200AP

1 x CPU KNL BIN1

6 x16 GB RDIMM, DDR4,
2400 MHz

72 nodes/rack

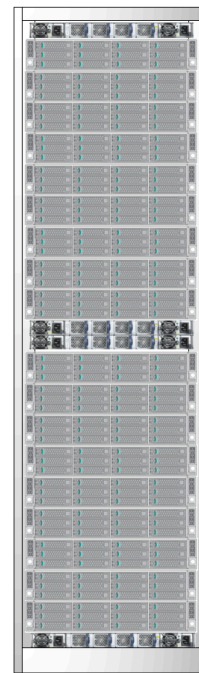
2x LENOVO G8052

GEthernet switch: 1 Gbps 48 ports

rack

Front View

42 U



1x Omni-path Eldorado Forest
edge switch 100 Gbps 48 ports

MARCONI: PERFORMANCE BENCHMARKS

MARCONI A1: High Performance Linpack

HPL	#MPI /# Thread	N / NB	Perf range
Single Node	1 / 36	104832 / 192 (90GB)	1.29 – 1.30 TFlops
Full System	1 / 36 for node	4320000/ 192 (90 GB)	1.6 – 1.72389 PFlops

Efficiency: **87 %** ←

TOP500 list Jun.2016 ranking: **47th**

MARCONI A2: High Performance Linpack

HPL	#MPI /# Thread	N / NB	Perf range
Single Node	1 / 64	40000 / 336 (12GB)	1.9 TFlops
Full System	1 / 64 for node	6287568 / 226 (83 GB)	6.22 PFlops

Efficiency: **90 %** ←

TOP500 list Nov.2016 ranking: **12th**

```
=====
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.
=====
```

```
=====
T/V          N  NB  P  Q          Time          Gflops
-----
W[0] R00L2L4 6287568 336 28 127          26628.96          6.22304e+06
HPL_pdgesv() start time Fri Nov 4 23:10:08 2016]

HPL_pdgesv() end time Sat Nov 5 06:33:57 2016

HPL Efficiency by CPU Cycle 2505.196%
HPL Efficiency by BUS Cycle 2439.395%

||Ax-b||_oo/(eps*(||A||_oo*||x||_oo+||b||_oo)*N)= 0.0006293 ..... 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.
=====
```

Knights Landing

INTEL Xeon Phi architecture family

Intel® Xeon Phi™ coprocessor product family “Knights Corner”	Intel® Xeon Phi™ coprocessor product family “Knights Landing”	Upcoming generation of the Intel® MIC architecture “Knights Hill”
2013	2H'2016	2017?
22 nm process	14 nm process	10 nm process
1 TeraFLOP DP peak	3+ TeraFLOP DP peak	?
57-61 cores In-order core architecture 1 Vector Unit per core	72 cores (36 tiles) Out-of-order architecture based on Intel® Atom™ core 2 Vector Units per core Up to 3x single thread performance w.r.t. Knights Corner	?
6-16 GB GDDR5 memory	On package, 8-16 GB high bandwidth memory (HBM) with flexible models: cache, flat, hybrid Up to 384 GB DDR4 main memory	?
Intel® Initial Many Core Instructions (IMIC)	Intel® Advanced Vector Extensions (AVX-512) Binary compatible with AVX2	?
PCIe coprocessor	Stand alone processor and PCIe coprocessor versions	?
Intel® True Scale fabric	Intel® Omni-Path™ fabric (integrated in some models)	2nd generation Intel® Omni-Path™ fabric

Knights Mill

Knights Landing (cont.. II)

INTEL Xeon Phi Knights Landing products

The First Many-Integrated-Core (MIC)

Self-boot (standard OS)

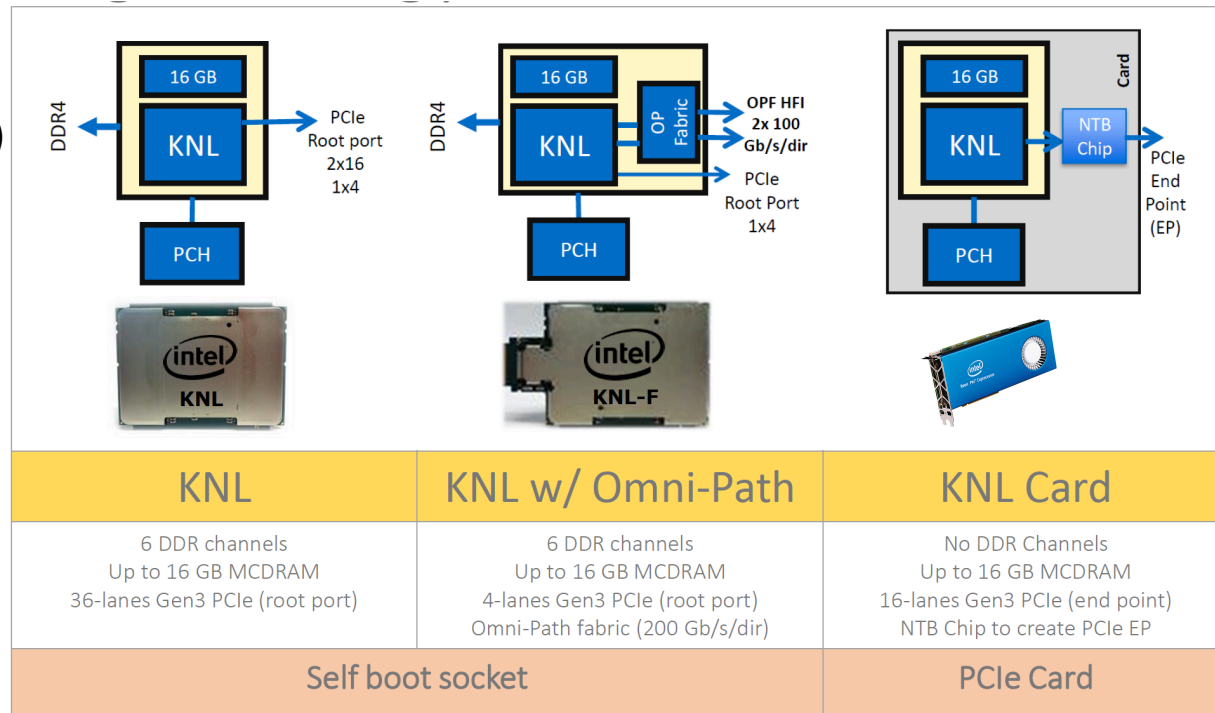
Binary compatible

Scalar & Vector improvement

High Bandwidth Memory

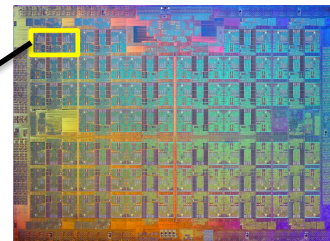
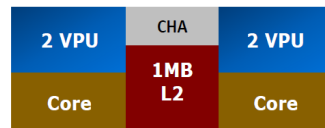
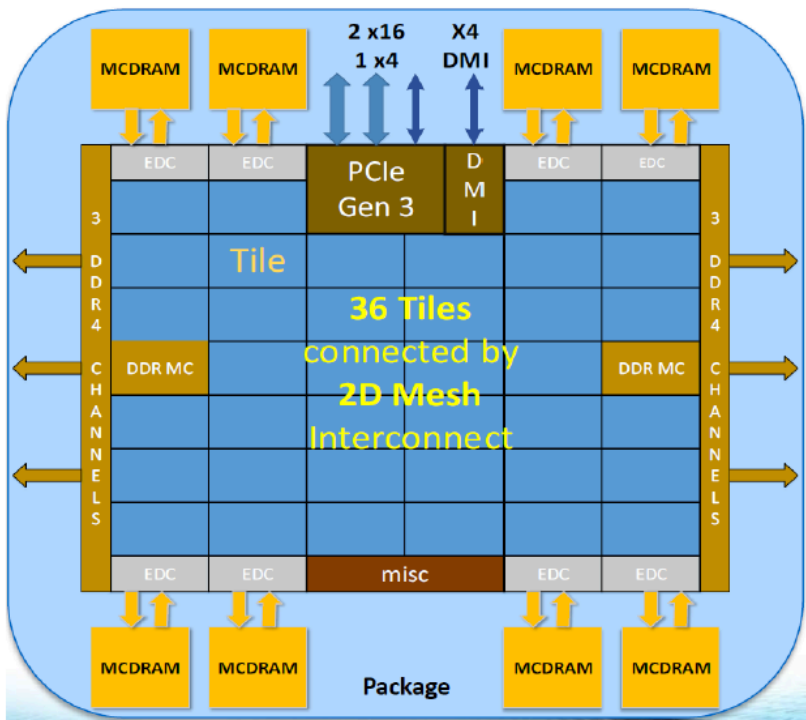
High Memory capacity

Integration of Fabric



Knights Landing (cont.. III)

INTEL Xeon Phi Knights Landing die



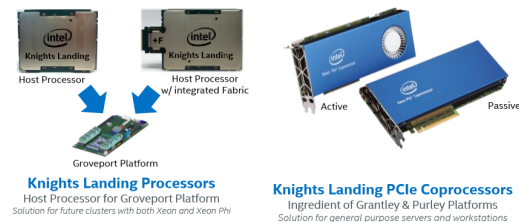
- Single Socket node
- **36** tiles connected by coherent 2D mesh
- Every tile 2 Silvermont cores + 2 512 bit VPU/core
- 1MB L2 cache
- Memory
- **16 GB** of High Bandwidth Memory (MCDRAM)
- DDR4, 6 channels @2.4 GHz , up to 384 GB
- IO & Fabric
- 36 Lanes PCIe Gen3
- 4 lanes of DMI for chipset
- On-package Omni-Path fabric

Knights Landing (cont.. IV)

INTEL Xeon Phi Knights Corner vs Knights Landing

Feature	KNC	KNL
Cores	<=61 cores Pentium, 1.1 GHz	<=72 Silvermont, 1.4GHz (KNL 7250)
Boot-up	Co-processor so needs host CPU	Standalone, self boot
Internal Network	Bi-directional ring	2D Mesh
Connections	PCIe	PCIe, Intel OmniPath or other vendor.
Memory	8-16GB on board	16 GB MCDRAM (High Bandwidth Memory) on board Supports upto 384Gb DDR
Vectorisation	512 bit SIMD/core	2x AVX 512 units/core
Xeon Compatibility	For Native mode recompile with -mic flag.	Binary compatible, although recompilation recommended (for vectorisation)
Peak Performance	~1 Tflops (DP)	~3 Tflops (DP)
Power consumption	300W	215W (KNL 7250)*

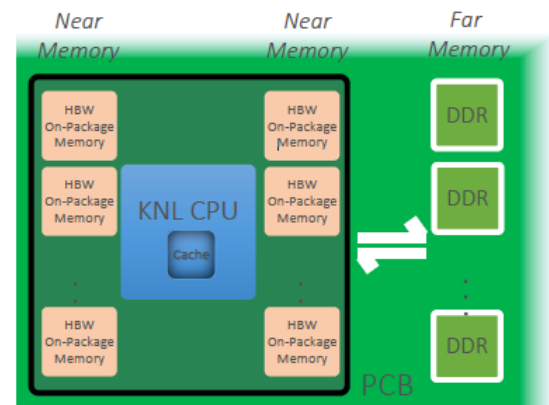
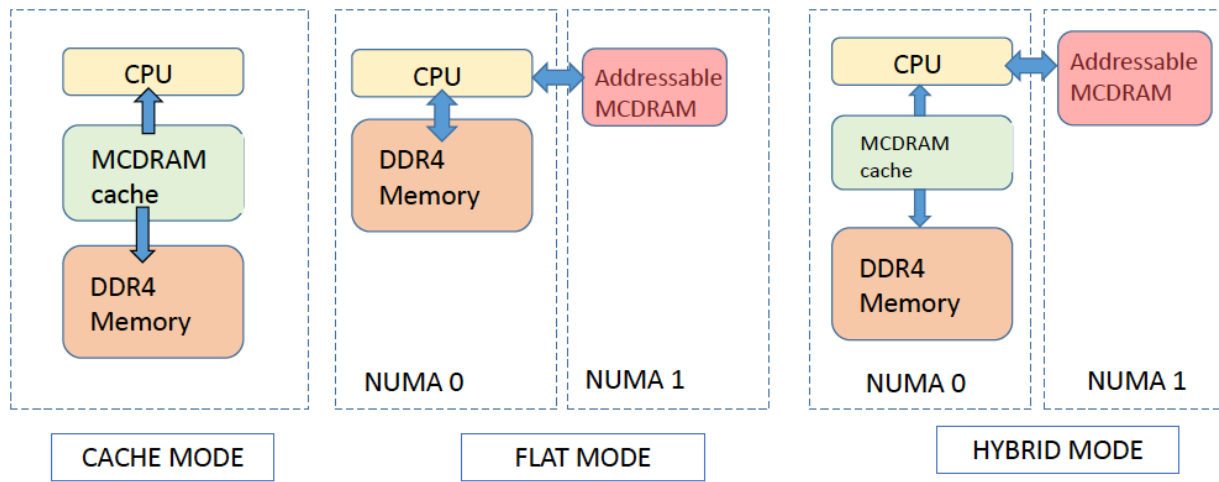
Knights Landing (Host or PCIe)



Knights Landing (cont.. V)

Using MCDRAM (High Bandwidth Memory) in KNL

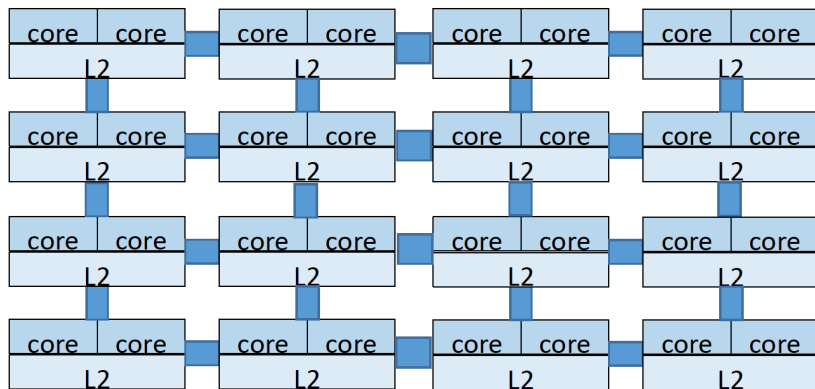
- Multi-Channel Dynamic RAM (MCDRAM) is High Bandwidth Memory (HBM)
 - > 400 GB/s – 5x faster than DDR4 RAM ($\approx 90\text{GB/s}$)
 - THREE MODEs configurable on BOOT in BIOS: **CACHE** / **FLAT** / **HYBRID**



Knights Landing (cont.. VI)

2D MESH Internal Network in KNL

- Cache organization
 - MESIF protocol for cohereting all the caches in the mesh
 - Each tile has a directory (*tag directory*) organized in the DTD (*Distributed Tag Directory*)
 - DTD identifies on the tile the location of any cache line
 - If a tile cannot find some data from its local cache then it must query the DTD to find the data



Each core has
its own L1 cache
(32K)

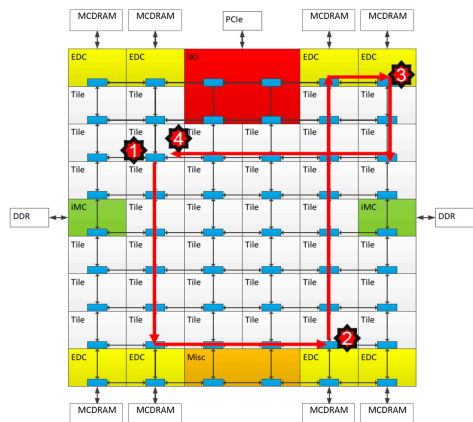
See, for example, <https://colfaxresearch.com/knl-numa/>

Knights Landing (cont.. VII)

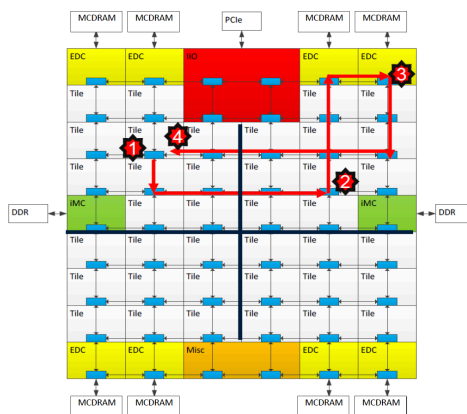
2D MESH Internal Network in KNL

- Cache Clustering Modes: reduce latency and maximise bandwidth
 - THREE MODEs configurable on BOOT in BIOS:
 - All-to-All** : Uniform Address for all, No affinity, Lower performance
 - Quadrant/Hemisphere**: 4 Quadrant / 2 Hemisphere, Uniform Address and Affinity in the same quadrant / hemisphere
 - SNC-4 / SNC-2** : Quadrant as NUMA domain, Affinity lowest latency , handled by software

Cluster Mode: All-to-All



Cluster Mode: Quadrant



Cluster Mode : Sub-NUMA



Knights Landing (cont.. VIII)

Software visible memory KNL configuration

1. Cache mode / Quadrant

```
[andrey@kn13 ~]$ numactl --hardware
available: 1 nodes (0)
node 0 cpus: 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99
node 0 size: 98200 MB
node 0 free: 92552 MB
node distances:
node 0
0: 10
[andrey@kn13 ~]$
```

2. Flat mode / Quadrant

```
[andrey@kn14 ~]$ numactl --hardware
available: 2 nodes (0-1)
node 0 cpus: 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99
node 0 size: 98200 MB
node 0 free: 94632 MB
node 1 cpus:
node 1 size: 16384 MB
node 1 free: 15929 MB
node distances:
node 0 1
0: 10 31
1: 31 10
[andrey@kn14 ~]$
```

3. Cache mode / SNC-4

```
[andrey@kn13 ~]$ numactl --hardware
available: 4 nodes (0-3)
node 0 cpus: 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99
node 0 size: 24472 MB
node 0 free: 22989 MB
node 1 cpus: 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99
node 1 size: 24576 MB
node 1 free: 23807 MB
node 2 cpus: 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99
node 2 size: 24576 MB
node 2 free: 23835 MB
node 3 cpus: 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99
node 3 size: 24576 MB
node 3 free: 23748 MB
node distances:
node 0 1 2 3
0: 10 21 21 21
1: 21 10 21 21
2: 21 21 10 21
3: 21 21 21 10
[andrey@kn13 ~]$
```

4. Flat mode with sub-NUMA clustering (SNC-4)

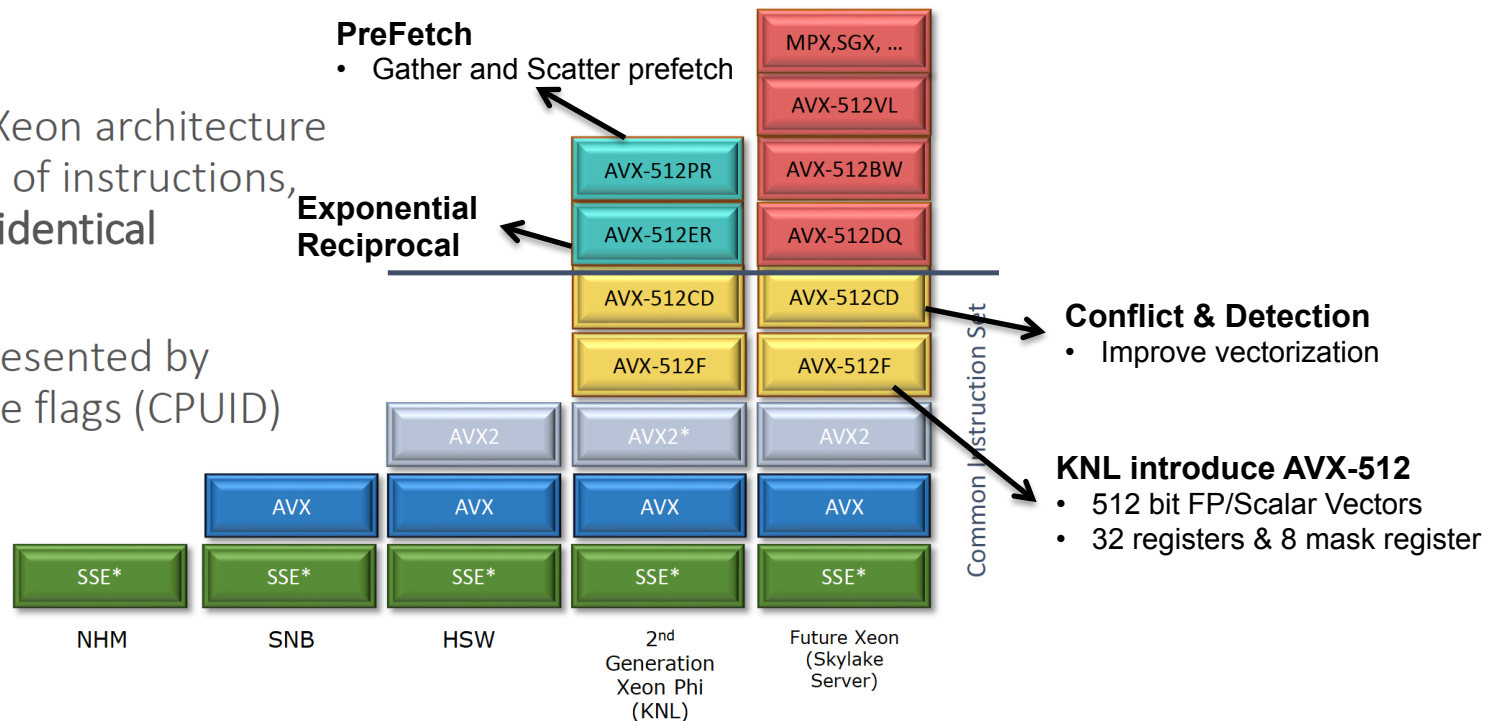
```
[andrey@kn13 ~]$ numactl --hardware
available: 8 nodes (0-7)
node 0 cpus: 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99
node 0 size: 24472 MB
node 0 free: 23186 MB
node 1 cpus: 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99
node 1 size: 24576 MB
node 1 free: 23807 MB
node 2 cpus: 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99
node 2 size: 24576 MB
node 2 free: 23741 MB
node 3 cpus: 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99
node 3 size: 24576 MB
node 3 free: 23756 MB
node 4 cpus:
node 4 size: 4096 MB
node 4 free: 3982 MB
node 5 cpus:
node 5 size: 4096 MB
node 5 free: 3981 MB
node 6 cpus:
node 6 size: 4096 MB
node 6 free: 3982 MB
node 7 cpus:
node 7 size: 4096 MB
node 7 free: 3979 MB
node distances:
node 0 1 2 3 4 5 6 7
0: 10 21 21 21 21 21 21 21
1: 21 10 21 21 21 21 21 21
2: 21 21 10 21 21 21 21 21
3: 21 21 21 10 21 21 21 21
4: 31 41 41 41 10 41 41 41
5: 41 31 41 41 41 10 41 41
6: 41 41 41 31 41 41 10 41
7: 41 41 31 41 41 41 41 10
[andrey@kn13 ~]$
```

Kinghts Landing (cont.. IX)

Intel Xeon PHI Kinghts Landing and future Intel Xeon Skylake

KNL and future Xeon architecture share a large set of instructions, but sets are not identical

Subsets are represented by individual feature flags (CPUID)



Knights Landing (cont.. X)

Intel Target architecture compiler options

Option	Description
<u>-mmic</u>	Builds an application that runs natively on Intel® MIC 1 st generation Architecture.
<u>-xfeature</u> <u>-xHost</u>	Tells the compiler which processor features it may target, referring to which instruction sets and optimizations it may generate (not available for Intel® Xeon Phi™ 1st gen architecture). Values for <i>feature</i> are: • COMMON-AVX512 (includes AVX512 FI and CDI instructions) • MIC-AVX512 (includes AVX512 FI, CDI, PFI, and ERI instructions) • CORE-AVX512 (includes AVX512 FI, CDI, BWI, DQI, and VLE instructions) • CORE-AVX2 • CORE-AVX-I (including RDRND instruction) • AVX • SSE4.2, SSE4.1 • ATOM_SSE4.2, ATOM_SSSE3 (including MOVBE instruction) • SSSE3, SSE3, SSE2 When using -xHost, the compiler will generate instructions for the highest instruction set available on the compilation host processor.
<u>-axfeature</u>	Tells the compiler to generate multiple, feature-specific auto-dispatch code paths for Intel® processors if there is a performance benefit. Values for <i>feature</i> are the same described for -xfeature option. Multiple features/paths possible, e.g.: -axSSE2, AVX . It also generates a baseline code path for the default case.
<u>-m, -march</u> <u>/arch</u>	Generate code exclusively for a given CPU

Vectorized code will be different depending on the chosen target architecture

MARCONI FUSION Operations

Project Implementing Agreement: Key Performance Indicators

- PIA KPIs:
 - HPC system: Availability (>97 %) , Usage (>85 %)
 - ICT infrastructure: Network (WAN:10Gbps) – Long Term Storage
 - Users Support
- Operation Committee monthly meeting – Quarterly report
- Allocation Committee: call for HPC projects (annual issue)
- Project Committee (Decisions)

MARCONI FUSION

PRODUCTION ENVIRONMENTS

- Login nodes MARCONI A1 (BDW): *login.marconi.cineca.it*
- MARCONI A1 as default. First rule:

```
franco-737:~ franco$ ssh fiannone@login.marconi.cineca.it
Last login: Mon May 8 11:54:48 2017 from nat52-2.frascati.enea.it
*****
* Welcome to MARCONI /
* MARCONI-fusion @ CINECA - NextScale cluster - CentOS 7.2!
*
* Broadwell partition - 1512 Compute nodes:
* - 2 18-core Intel(R) Xeon(R) E5-2697 v4 @ 2.30GHz per Compute node
* - 128 GB RAM per Compute node
* * KNL partition - 3680 Compute nodes:
* - 1 68-core Intel(R) Knights Landing @ 1.40GHz per Compute node
* - 16 GB MCDRAM per Compute node
* - 93 GB RAM per Compute node
* Intel OmniPath (100Gb/s) high-performance network
* PBSpro 13 batch scheduler
*
* wiki.u-gov.it/confluence/display/SCAIUS/UG3.1%3A+MARCONI+UserGuide
* for a guide on Marconi
*
* mailto: superc@cineca.it for support
*****
* This system is in full-production. *
*****
IMPORTANT:
- A new version of "module" is installed. Applications are available through
domain-based profiles to be loaded before the module itself. Use the "modmap"
command to identify the correct profile ("modmap -h" for help).
- Marconi is little-endian (like GALILEO and PICO), in contrast with FERMI
which was big-endian.
- Load the module "env-kenl" to switch from Broadwell to KNL environment.
Unload it or load "env-bdw" to revert to Broadwell environment.
*****
Since March 15th accounting is enabled on KNL partition. The computing hours
spent with your jobs will be now detracted from your budget, as it already
happens for Broadwell and other CINECA HPC clusters.
*****
IMPORTANT: Starting from April 27th, 2017, an automatic cleaning procedure
for the SCINECA_SCRATCH is active.
Each day all files older than 50 days will be cancelled.
*****

Executing .bashrc
End of .bashrc
[fiannone@r000u06l01 ~]$
```

To switch on MARCONI A2 (KNL):

[fiannone@r000u06l01 ~]\$ module load env-kenl
(KNL) [fiannone@r000u06l01 ~]\$

Back to MARCONI A1(BDW):

(KNL) [fiannone@r000u06l01 ~]\$ module load env-bdw
(BDW) [fiannone@r000u06l01 ~]\$

MARCONI FUSION

PRODUCTION ENVIRONMENTS

- Resources Management System: PBSpro v.13.1
- 2 separate PBS clusters for BDW and KNL

MARCONI A1(BDW):

```
(BDW) [fiannone@r000u06l01 ~]$ qstat -Q
Queue           Max    Tot  Ena Str  Que  Run  Hld  Wat  Trn  Ext Type
-----
bigprod          0    21 yes yes  12    1    8    0    0    0 Exec
xfuagw           0    63 yes yes   0   57    6    0    0    0 Exec
xfualongprod     0    38 yes yes   0   38    0    0    0    0 Exec
xfuaprod         0   145 yes yes  119   10   16    0    0    0 Exec
xfuadebug        0    1 yes yes   0    1    0    0    0    0 Exec
xfuabigprod      0    7 yes yes   1    5    1    0    0    0 Exec
xfualowprio      0    0 yes yes   0    0    0    0    0    0 Exec
test             0    0 yes yes   0    0    0    0    0    0 Exec
serial           0    0 yes yes   0    0    0    0    0    0 Exec
system           0    1 yes yes   0    0    1    0    0    0 Exec
route            0    0 yes yes   0    0    0    0    0    0 Rout
meteopar         0    3 yes yes   0    3    0    0    0    0 Exec
meteoser         0    1 yes yes   0    1    0    0    0    0 Exec
special          0    0 no yes   0    0    0    0    0    0 Exec
debug            0    2 yes yes   0    2    0    0    0    0 Exec
xfuagwdebug      0    0 yes yes   0    0    0    0    0    0 Exec
quarantine       0    0 yes yes   0    0    0    0    0    0 Exec
admin            0    0 yes yes   0    0    0    0    0    0 Exec
xfuaspecial      0    2 yes yes   0    2    0    0    0    0 Exec
gwg2             0    0 yes yes   0    0    0    0    0    0 Exec
prod             0  4358 yes yes 3093  268  993    0    0    0 Exec
R641493          0    0 yes yes   0    0    0    0    0    0 Exec
R645178          0    6 yes yes   5    1    0    0    0    0 Exec
```

MARCONI A2 (KNL):

```
(KNL) [fiannone@r000u06l01 ~]$ qstat -Q
Queue           Max    Tot  Ena Str  Que  Run  Hld  Wat  Trn  Ext Type
-----
knlsystem        0    0 yes yes   0    0    0    0    0    0 Exec
knlroute         0    0 yes yes   0    0    0    0    0    0 Rout
knlquarantine    0    0 yes yes   0    0    0    0    0    0 Exec
knldebug         0    3 yes yes   0    0    3    0    0    0 Exec
knladmin         0    0 yes yes   0    0    0    0    0    0 Exec
xfuaknldebug     0    0 yes yes   0    0    0    0    0    0 Exec
xfuaknlprod      0   29 yes yes   2   27    0    0    0    0 Exec
knltest          0    0 yes yes   0    0    0    0    0    0 Exec
knlprod          0   694 yes yes  337  256  99    0    0    0 Exec
R168296          0    0 yes yes   0    0    0    0    0    0 Exec
R169245          0    0 yes no    0    0    0    0    0    0 Exec
```

MARCONI FUSION

PRODUCTION ENVIRONMENTS

- INTEL COMPILER for KNL
mpicc -xMIC-AVX512 -o myexe mycode.c
- SUBMITTING a JOB with PBS script on KNL

```
#!/bin/bash
#PBS -q <queue>           # xfuaknlprod, xfuaknldebug
#PBS -n jobname
#PBS -e job.err
#PBS -o job.out
#PBS -l walltime=24:00:00  #maximum walltime requirable

# Total reserved nodes = 340
#PBS -l select=196:ncpus=68:mpiprocs=68:mem=93GB:mcdram=cache
or
#PBS -l select=144:ncpus=68:mpiprocs=68:mem=108GB:mcdram=flat

#PBS -A <account_no>

cd $PBS_O_WORKDIR
module load autoloader intelmpi/2017--binary
mpirun -n xx ./myexe
```

MARCONI FUSION

PRODUCTION ENVIRONMENTS

- **MODULE Tool**

Modulefiles grouped in profiles:

Profiles currently defined are the following:

- **profile/base** (default): all serial and parallel compilers, most common libraries (Intel or Python compiled), most common tools for debugging and profiling
- **profile/advanced**: libraries compiled with Gnu, Openmpi-intel, mvapich2 and other “less recommended” modules
- **profile/archive**: older versions of updated modules, for retrocompatibility
- **domain profiles**: here you can find all the application softwares. They are divided in profiles based on scientific domains.

Current available domain profiles are:

- **profile/astro**
- **profile/lifesc**
- **profile/bioinf**
- **profile/phys**
- **profile/chem**

modmap tool to navigate

Usage examples:

modmap -m <modulename>

to know where to find a specific module

modmap -p <profilename>

to get the list of modules contained in a specific profile

modmap -c <categoryname>

to get the list of all the modules in a specific category (tools, libraries,...), divided by profile

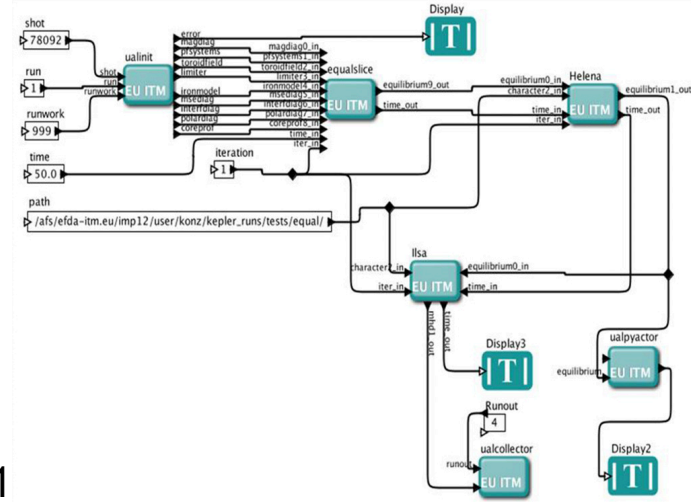
modmap -h

to get an help about the command usage

```
(BDW) [fiannone@r000u06l01 ~]$ modmap -m namd
Profile: advanced
Profile: archive
      namd
      2.11
Profile: astro
Profile: base
Profile: bioinf
Profile: chem
      namd
      2.12
Profile: deeplrn
Profile: knl
      namd
      2.12_knl
Profile: lifesc
      namd
      2.12
      2.12_knl
Profile: phys
      namd
      2.12
Profile: orofile
```

The HPC infrastructure

- Development platform in operation since 2009
- A community of +250 users
- Integrated in MARCONI Fusion
 - ✓ 2 dedicated login nodes provided in kind by CINECA
 - ✓ 24 dedicated compute nodes belongs to MARCONI A1 compute nodes able to swap on Gateway users interactive sessions + 14 compute nodes.
 - ✓ Access in kind to 14 compute nodes of MARCONI A1 CINECA partition with a batch queue for Gateway users to submit jobs on max 14 compute nodes.
 - ✓ 600 TB of storage area.

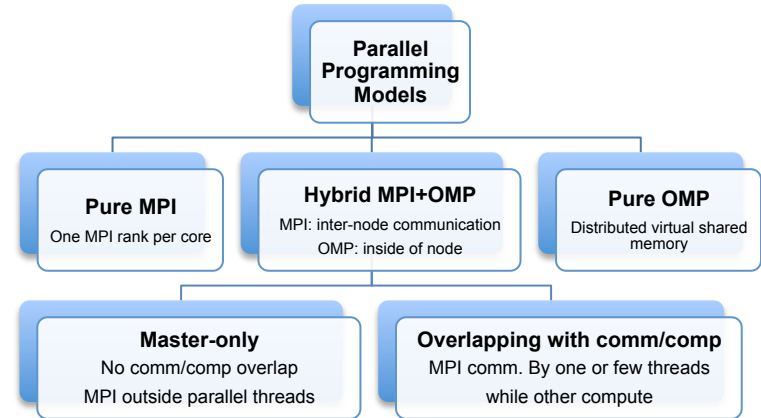
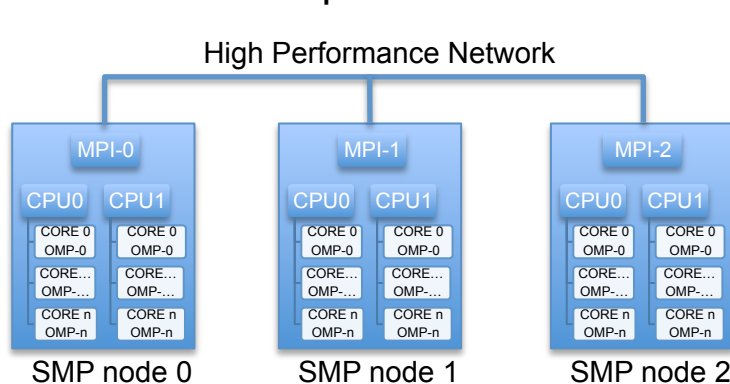


Parallel Hybrid Programming Model

SMP cluster: Multi-socket/core nodes with a High Performance Network

Each node is mapped (at least) one MPI rank and OpenMP threads more

- OpenMP within node allows a good use of the memory resource (ccNUMA)
- One MPI rank per node avoids additional communication overheads



Many-In-Core (MIC) clusters boost towards hybrid model

Parallel Programming Model

Message Passing Interface (MPI)

Compute nodes communication over low latency network

- Scalability with number of nodes
- Overhead communication limits the number of MPI ranks
- MPI affinity: pinning the rank to the processor

Multi-Threading (OpenMP)

Symmetric Multi-Processing with Shared Memory

- Parallel Threads on single multi-socket/core node
- Easy to deploy with *#pragma omp parallel* directive
- Cache-Coherent Non Uniform Memory Access (ccNUMA: faster access of the core)
- OMP affinity: pinning the thread to the core

OpenMP thread affinity	
Variable	Description
KMP_AFFINITY=[scatter,compact,proclist={...}]	Change the affinity
KMP_PLACE_THREADS, KMP_HW_SUBSET=<T>T	Threads per core
OMP_NUM_THREADS=n	Number of threads per rank
MPI rank affinity	
Variable	Description
I_MPI_PIN	Turn on/off pinning
I_MPI_PIN_PROCESSOR_LIST=<value>	Define a processor subset and the mapping rules for MPI processes within this subset
I_MPI_PIN_DOMAIN=[node,core,numa]	Define a number of domains of logical processor

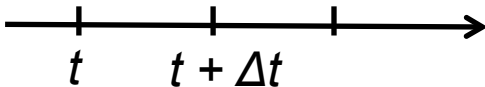
Hybrid case study

N-body problem

Simulation of a dynamic system of N bodies interacting by means of a physical force such as : gravity

Newton's law $m_i \ddot{\vec{r}}_i = \vec{F}_i$ *Gravitational physical force* $\vec{F}_i = \sum_{j \neq i} \frac{Gm_i m_j (\vec{r}_j - \vec{r}_i)}{|\vec{r}_j - \vec{r}_i|^3}$

Numerical integration


$$\dot{\vec{r}}_i(t + \Delta t) = \dot{\vec{r}}_i(t) + \Delta t \cdot \frac{\vec{F}_i}{m_i}$$
$$\vec{r}_i(t + \Delta t) = \vec{r}_i(t) + \Delta t \cdot \dot{\vec{r}}_i(t)$$

A Particle-Particle brute force approach with $O(N^2)$ complexity

N-body problem

Hybrid MPI+OMP Master-only model

MPI rank per node and OpenMP threads scattered on the cores of the node, with no MPI calls inside MP parallel regions

```
set initial positions and velocities
set subset particles number for each rank
for each time-step  $\Delta t$  do
  #pragma omp parallel for
  for each particle  $i$ th do
    for each particle  $j$ th do
      evaluate the force on  $i$ th particle
    endfor  $j$ th particle
    integrate to calculate  $\Delta x$  for  $i$ th particle
  endfor  $i$ th particle
  propagate results of time-step on all ranks
endfor time-step
```

OMP Parallel region

MPI comm. function

N-body simulation tests

BDW vs KNL at MARCONI FUSION

Processor	Name	Frequency	Threads	Vector width
	BDW E5-2697 v4	2.3	$2 \times 18 = 36$	256 bits (8 floats)
	KNL-D 7250	1.4	68	512 bits (16 floats)

One time-step, N bodies = 10^6 , 10^7 , 10^8

PBS queue with walltime of 24 h

KNL cache and flat MCDRAM configuration

N-body simulation results

BDW vs KNL at MARCONI

Execution time for Pure OpenMP on single node

	BDW		KNL FLAT		KNL CACHE	
N-body	1 thread (sec.)	36 threads (sec.)	1 thread (sec.)	68 threads (sec.)	1 thread (sec.)	68 threads (sec.)
10^6	3037.9	89.5	3164.4	58.1	3175.8	58.2
10^7		10468.2		8804.6		8942.8

Execution time for Hybrid MPI+OpenMP on multi-nodes

	BDW			KNL FLAT			KNL CACHE		
N-body	2 nodes (sec.)	16 nodes (sec.)	32 nodes (sec.)	2 nodes (sec.)	16 nodes (sec.)	32 nodes (sec.)	2 nodes (sec.)	16 nodes (sec.)	32 nodes (sec.)
10^6	45	5.65	2.85	29.4	4.12	2.29	29.5	4.23	2.31
10^7	5275.3	663.6	332.1	4247.4	584	293.1	4243.2	581	301.5
10^8		66317	33086.5		77882	39944		79125	39909

Conclusions

HPC for FUSION in operation until Dec.2018 (Gateway too)

INTEL trend : to converge on Many Integrated Core

INTEL Omni-Path: 2nd Generation for Exascale

Hybrid Parallel Programming Model