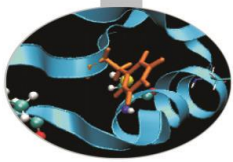
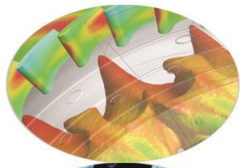
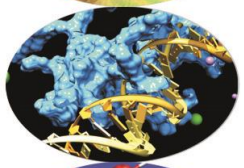
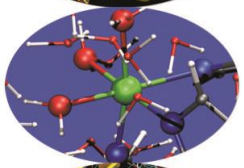
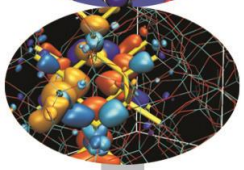
CINECA SCAI
SuperComputing Applications and Innovation



Oveview on HPC cluster
MARCONI

Installation roadmap and technical informations

Alessandro Marani

CINECA

A new Supercomputer (codename: MARCONI) has been installed at CINECA, available for Italian and European research community.

It is a Lenovo NeXtScale system based on Intel technology, with a final peak performance of around 20PFlop/s.

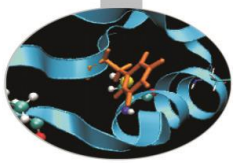
Deployment of Marconi is started July 2016, complete delivery reached at August 2017.

The first part (2 PFlop/s conventional "Broadwell" based) is in full production as July 2016.

The second part (11 Pflop/s "Knigths Landing" based) is in full production as January 2017.

The third part (5+ Pflop/s " SkyLakes" based) is in full production as August 2017.

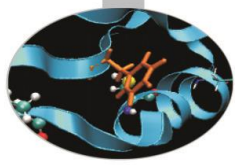
A short history



- In 2015 the computing resources in Cineca were:
  - Tier-0: **FERMI** (acquired in summer 2012)
  - Tier-1: **GALILEO** (acquired in Jan 2015)
  - Front-end, Viz, BigData: **PICO** (acquired in Nov 2014)
- FERMI arrived at the expected end of its activity.
- The Cineca governing bodies, aiming at supporting scientific research, approved a development plan, with an investment of Euro 50 million in two phases, from 2016 to 2020:
  - 2 x 5 → 10 Pflops in 2016-2017
  - 10 x 5 → 50 Pflops in 2019-2020

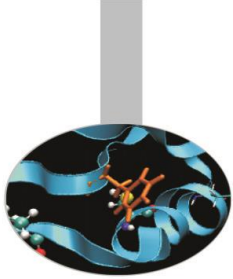
Marconi

MARCONI: the new Tier-0 system



- A tender was issued in 2015 and assigned Jan 2016 to lenovo
- The system has been delivered in three phases:
  - A1: April 2016 (BDW 2 PFs)
  - A2: Sept 2016 (KNL 11 PFs)
  - A3: July 2017 (SKL 4.5 PFs)
- In total:
  - 17.5 PFs peak performance
  - 17PB row storage
  - 3MWatt power consumption

MARCONI:
New Tier-0 system



Technical Features:

- Intel based
- Architecture: Lenovo NeXtScale
- Fabric: Intel OmniPath

A2

KNL 68cores, 1.4 GHz;
3600 nodes, 11 PFs

A1

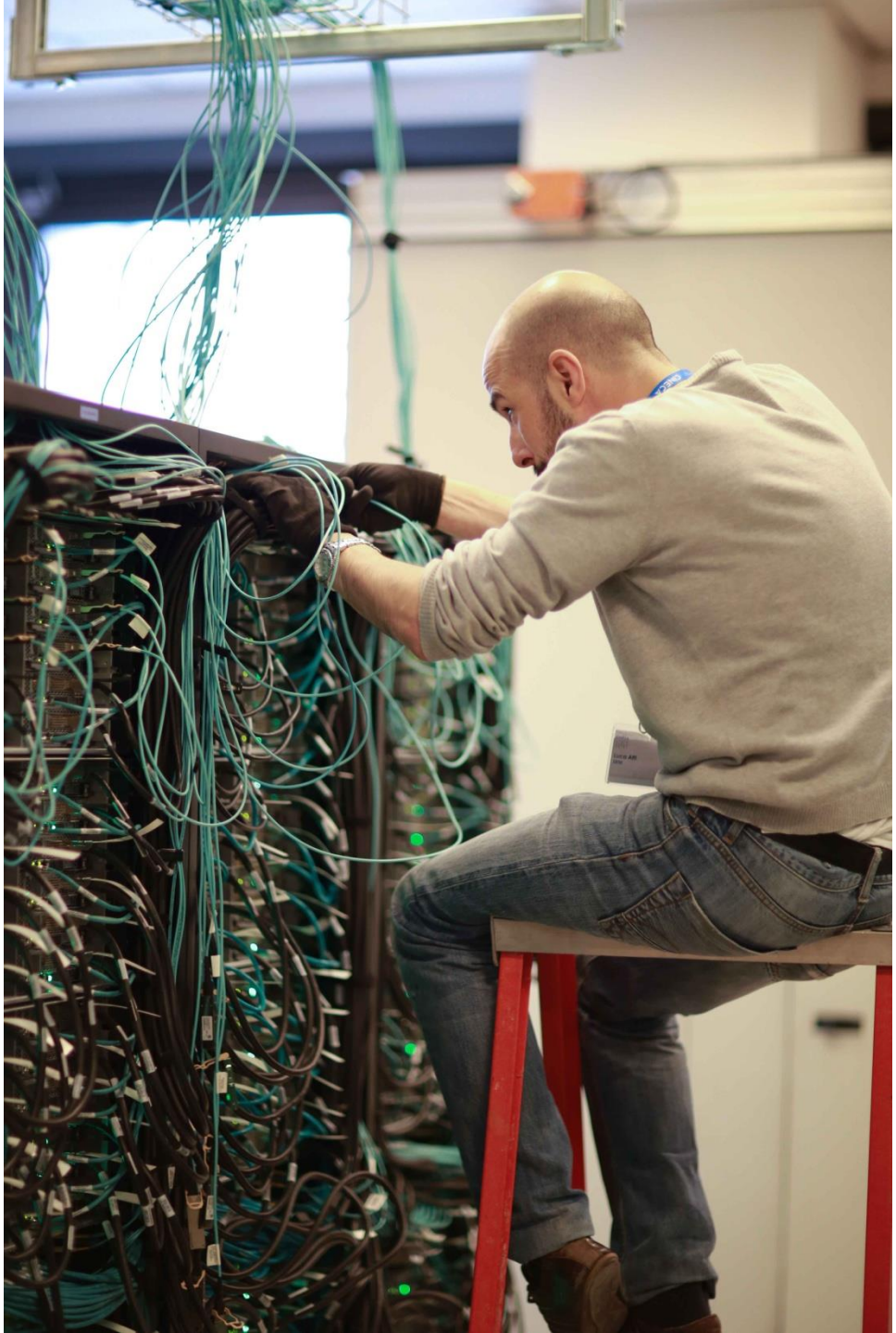
BRD 2x18 cores, 2.3GHz
1512 nodes, 2PFs

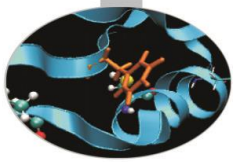
A3

SKL 2x24 cores, 2.1 GHz;
1512 nodes, 4.5 PFs

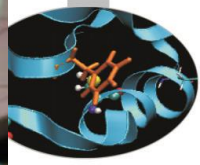
CINECA SCAI

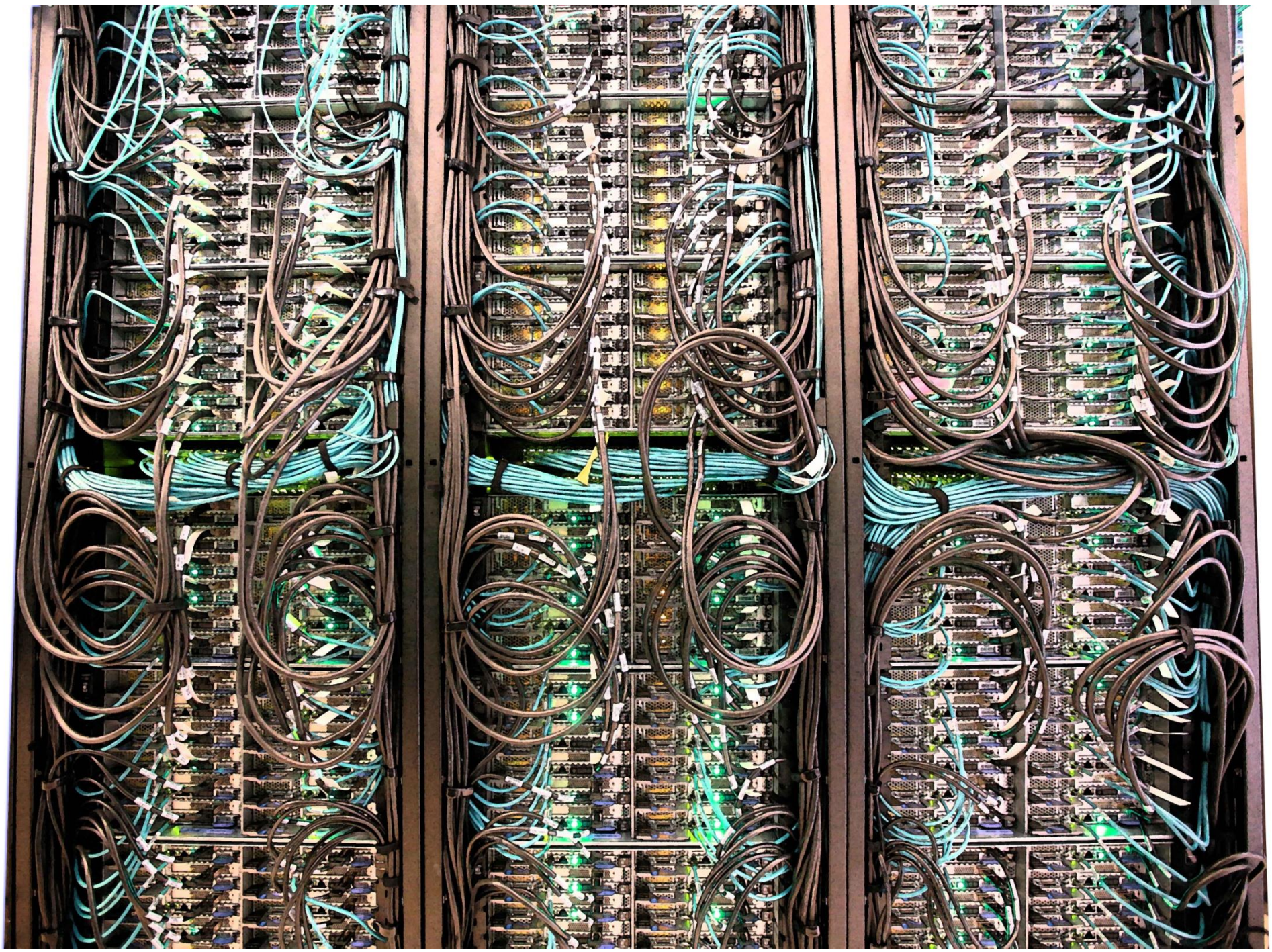
SuperComputing Applications and Innovation





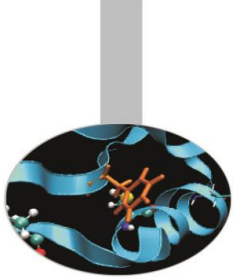
CINECA





CINECA SCAI
SuperComputing Applications and Innovation

Technical details



A1

Peak Perf.	Comp. Nodes	Socket	RAM/CN	Interconnect	Rack #	Service & Mgmt nodes
2PFs	1512	2x Intel Broadwell 18cores @2.3GHz	128 GB	Intel OmniPath 2:1 100Gb/s	21	8 Front End Nodes (2xBDW 18c +128GB RAM)+ xx MGMT nodes

Core tot: 54.432
Core-h/year=476.824.320

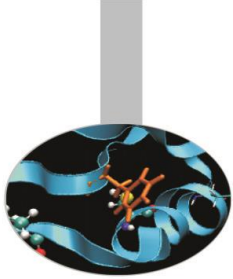
A2

Peak Perf.	Comp. Nodes	Socket	RAM/CN	Interconnect	Rack #	Service & Mgmt nodes
11 PFs	3600	Intel KnightsLanding 68cores @1.4 GHz	96+16 GB	Intel OmniPath 2:1 100Gb/s	50	Share login nodes with A1

Core tot: 244.800
Core-h/year=2.144.448.000

CINECA

Technical details



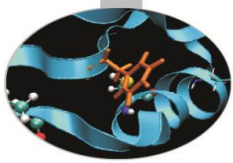
A3

Peak Perf.	Comp. Nodes	Socket	RAM/CN	Interconnect	Rack #	Service & Mgmt nodes
5PFs	1512 (+ 1.000)	2x Intel SkyLake 24cores @2.1GHz	192 GB	Intel OmniPath 2:1 100Gb/s	21	Share login nodes with A1 and A2

Core tot: 60.480

Core-h/year=529.804.800

MARCONI IN TOP500



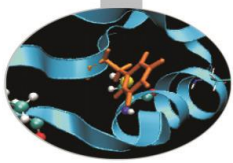
		Dell EMC				
13	GSIC Center, Tokyo Institute of Technology Japan	**TSUBAME3.0** - SGI ICE XA, IP139-SXM2, Xeon E5-2680v4 14C 2.4GHz, Intel Omni-Path, NVIDIA Tesla P100 SXM2 HPE	135,828	8,125.0	12,127.1	792
14	CINECA Italy	**Marconi Intel Xeon Phi** - CINECA Cluster, Lenovo SD530, Intel Xeon Phi 7250 68C 1.4GHz/Platinum 8160, Intel Omni-Path Lenovo	314,384	7,471.1	15,372.0	
15	United Kingdom Meteorological Office United Kingdom	Cray XC40, Xeon E5-2695v4 18C 2.1GHz, Aries interconnect Cray Inc.	241,920	7,038.9	8,128.5	
16	Barcelona Supercomputing Center Spain	**MareNostrum** - Lenovo SD530, Xeon Platinum 8160 24C 2.1GHz, Intel Omni-Path Lenovo	153,216	6,470.8	10,296.1	1,632

MARCONI A2+A3 ranks **#14** in the latest Top500, updated on November 2017, with a measured performance of **7.47 PFlop/s**

It is also the most powerful supercomputer in EU

MARCONI A1 ranks **#72** with a measured performance of **1.72 PFlop/s**

(some of) Our Users



EUROfusion

CINECA
ISCRA

LISA

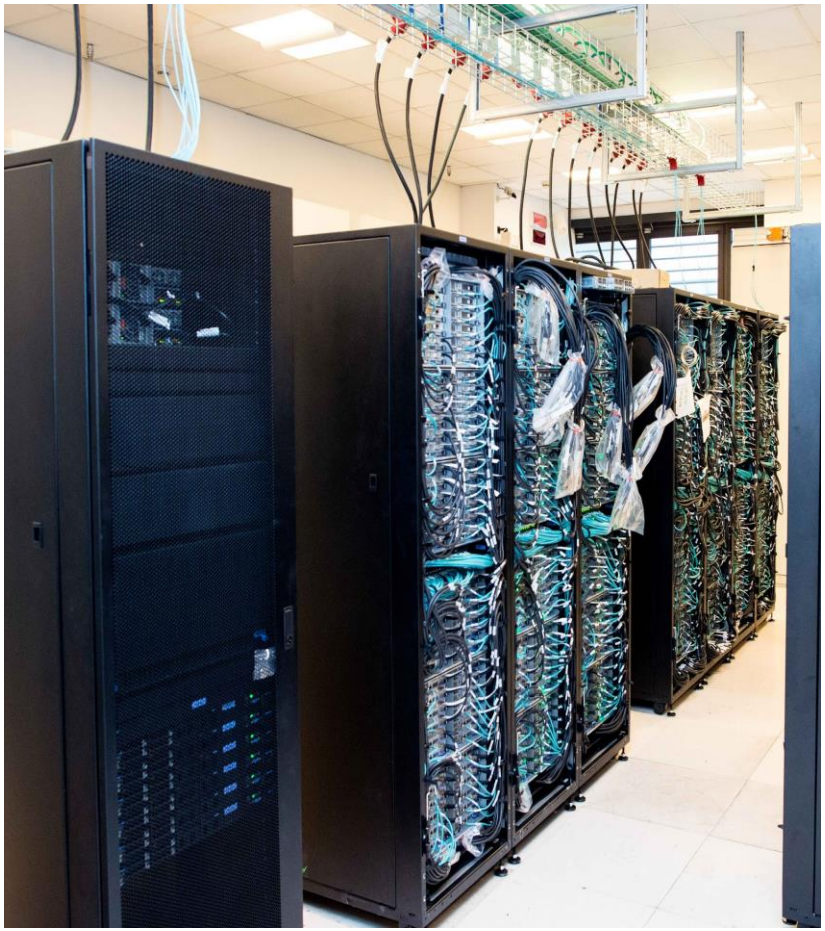
PRACE

INFN
Istituto Nazionale
di Fisica Nucleare

ma per seguir virtute e canoscenza
SISSA

ff
FORTISSIMO

MARCONI A1 (BDW)



Processor: Intel Xeon Broadwell, 2.3 GHz
Number of processors (cores): 54432
Number of nodes: 1512 (36 cores/node)

RAM: 190 TB (5 GB/core)
Interconnection network: Intel OmniPath

Peak performance: 2 PFlop/s

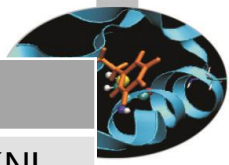
MARCONI A2 (KNL)

Processor: Intel Knights Landing, 1.4 GHz
Number of processors (cores): 244800
Number of nodes: 3600 (68 cores/node)

RAM: 337,5 TB (1,4 GB/core)
Interconnection network: Intel OmniPath

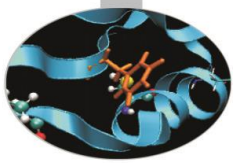
Peak performance: 11 PFlop/s

KNC →KNL key differences



Feature	KNC	KNL
Cores	<=61 cores Pentium, 1.1 GHz [in-order]	<=72 Silvermont, 1.4GHz (KNL 7250)[out-of-order]
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 AVX2 512 units/core
Xeon Compatibility	For Native mode recompile with -mic flag.	Binary compatible, although recompilation recommended (for vectorization)
Peak Performance	~1 Tflops (DP)	~3 Tflops (DP)
Power consumption	300W	215W (KNL 7250)

MCDRAM modes



Cache mode

- The MCDRAM is used as cache so it may give performance benefits if DDR memory accesses are reduced.
- Transparent to users so no modifications required.
- But increases latency if data is not found in cache (DDR → MCDRAM → L2).

Flat mode

- High bandwidth, low latency.
- More complicated to use – requires software or environmental changes.

Hybrid

- Benefits of both, but smaller sizes.

For production reasons, MCDRAM on MARCONI has been fixed on “cache” for all production nodes

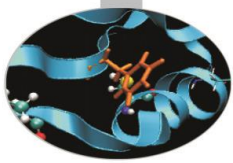
MARCONI A3 (SKL)

Processor: Intel SkyLake, 2.1 GHz
Number of processors (cores): 73576
Numeber of nodes: 1512 (48 cores/node)

RAM: 283,5 TB (4 GB/core)
Interconnection network: Intel OmniPath

Peak performance: 4.5 PFlop/s

MARCONI network



Network type: new Intel Omnipath, 100 Gb/s. MARCONI is the largest Omnipath cluster of the world.

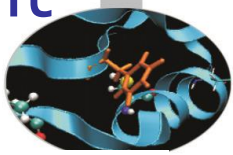
Network topology: Fat-tree 2:1 oversubscription tapering at the level of the core switches only.

Core Switches: 5 x OPA Core Switch "Sawtooth Forest", 768 ports each.

Edge Switch: 216 OPA Edge Switch "Eldorado Forest", 48 ports each.

Maximum system configuration: 5(opa) x 768 (ports) x 2 (tapering) → 7680 servers.

MARCONI front-end environment



Login nodes: 8 Login (3 available for regular users). Each one contains **2 x Intel Xeon Processor E5-2697 v4** with a clock of 2.30GHz and 128 GB of memory. Login nodes are shared between three partitions.

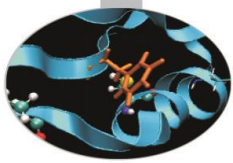
CentOS 7.2 is the OS of MARCONI. Since the architecture is Intel-based, Intel compilers family suite is the recommended one to use for compiling your codes on the cluster. Many of the most common tools, libraries and domain specific applications are installed on our environment and made available through a module system.

You can submit your batch or interactive job on a specific partition of MARCONI, that will be dispatched by **Pro Batch Scheduler** v. 13.1.0.

```
*******************************************************************************
* Welcome to MARCONI /                                                        *
*            MARCONI-fusion @ CINECA - NeXtScale cluster - CentOS 7.2!        *
*                                                                             *
* Broadwell partition - 1512 Compute nodes with:                              *
*    - 2*18-core Intel(R) Xeon(R) E5-2697 v4 @ 2.30GHz                        *
*    - 128 GB RAM                                                             *
* KNL partition - 3600 Compute nodes with:                                    *
*    - 1*68-core Intel(R) Knights Landing @ 1.40GHz                           *
*    - 16 GB MCDRAM + 93 GB RAM                                               *
* SKL partition - 1512+792 nodes with:                                        *
*    - 2*24-core Intel Xeon 8160 CPU @ 2.10GHz                                *
*    - 192 GB DDR4 RAM                                                        *
* Intel OmniPath (100Gb/s) high-performance network                           *
* PBSpro 13 batch scheduler                                                   *
*                                                                             *
* For a guide on Marconi:                                                     *
* wiki.u-gov.it/confluence/display/SCAIUS/UG3.1%3A+MARCONI+UserGuide           *
* For support:                                                                *
* mailto: superc@cineca.it                                                    *
*******************************************************************************
* This system is in full-production. *
==============================================================================
IMPORTANT:
- A new version of "module" is installed and based on profiles. Use the "modmap"
  command to identify the correct profile ("modmap -h" for help).
- "module load env-knl" to switch to KNL environment
  "module load env-skl" to switch to SKL environment
  "module load env-bdw" to switch to Broadwell environment (default)

==============================================================================
  IMPORTANT: Starting from April 27th, 2017, an automatic cleaning procedure
  for the $CINECA_SCRATCH is active.
  Each day all files older than 50 days will be cancelled.
==============================================================================
------------------------------------------------------------------------------
SKL partition (Marconi-A3, first 1512 nodes) is in production since Aug 7.
This first delivery is for use of EUROfusion users only
Accounting on Marconi-A3 is active since September 6th, 2017
------------------------------------------------------------------------------
[amarani0@r000u06l01 ~]$
```

MARCONI filesystem



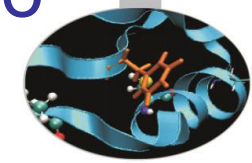
A high-performance Lenovo GSS storage subsystem, that integrates the IBM Spectrum Scale™ (GPFS) file system, is connected to the Intel Omni-Path Fabric and provides data storage capacity.

At login, you will find yourself into your “Home” space ($HOME). It is a permanent back-upped space with a 50 Gb quota, suited for compilation and small debugging.

For production, move into $CINECA_SCRATCH or $WORK filesystems. They are large, parallel filesystems suited for intensive I/O activity:

- $CINECA_SCRATCH: personal area, with no backup and no quota, but a periodic cleaning procedure.
- $WORK: area shared with all the collaborators of an account (i.e. project). It is not backupped and has a quota of 1 Tb (extendable upon request to User Support)

Benchmark: Quantum Espresso



Tests performed to verify strong scaling.

QE is a hybrid code, using MPI as well as OpenMP threads.

For all tests, KMP_AFFINITY=scatter (even thread distribution) has been used.

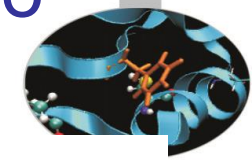
Next slides focus on computing performance vs:

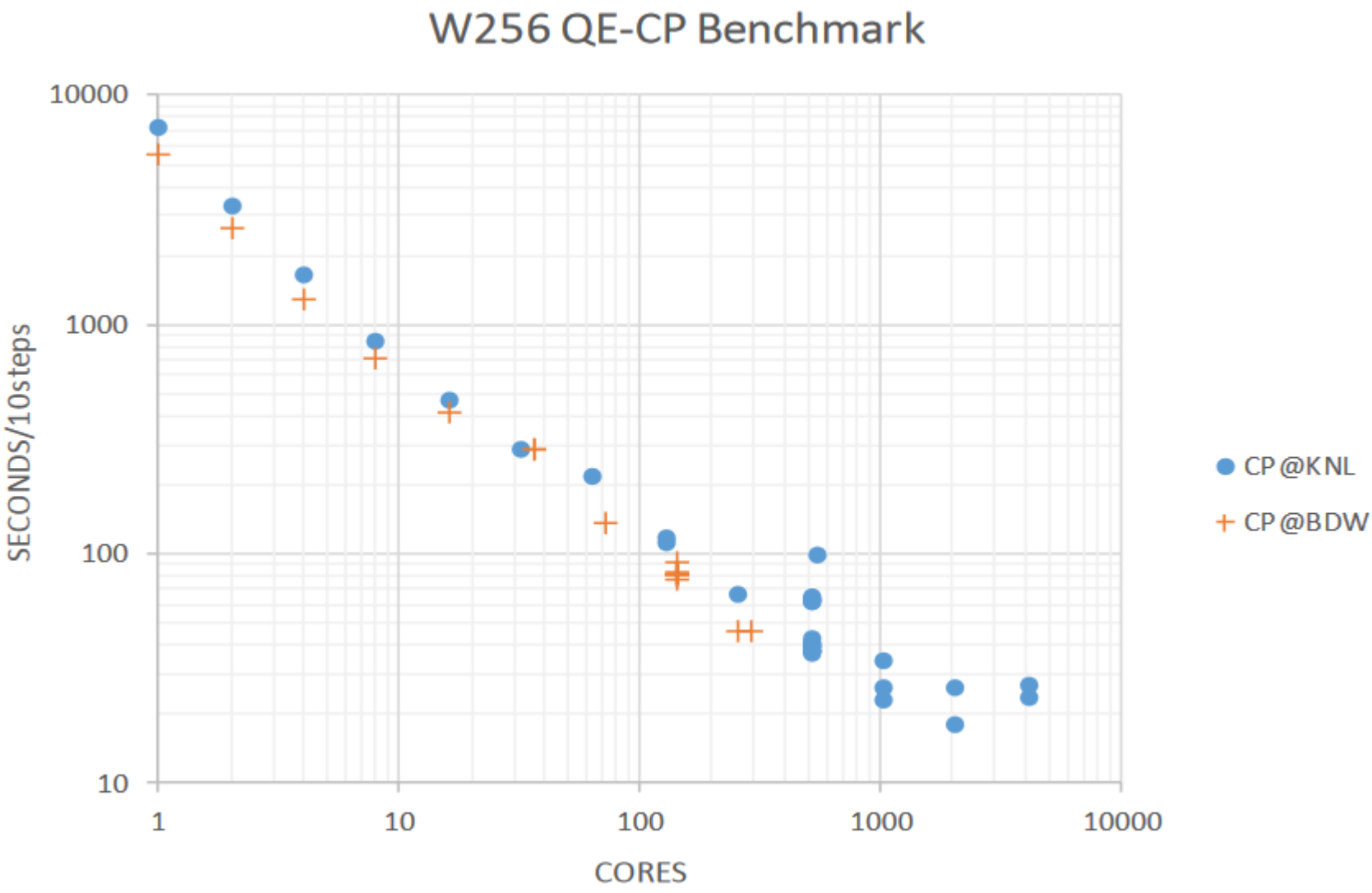
- Number of cores

-Power consumption

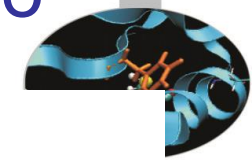
Results are compared between KNL and BDW

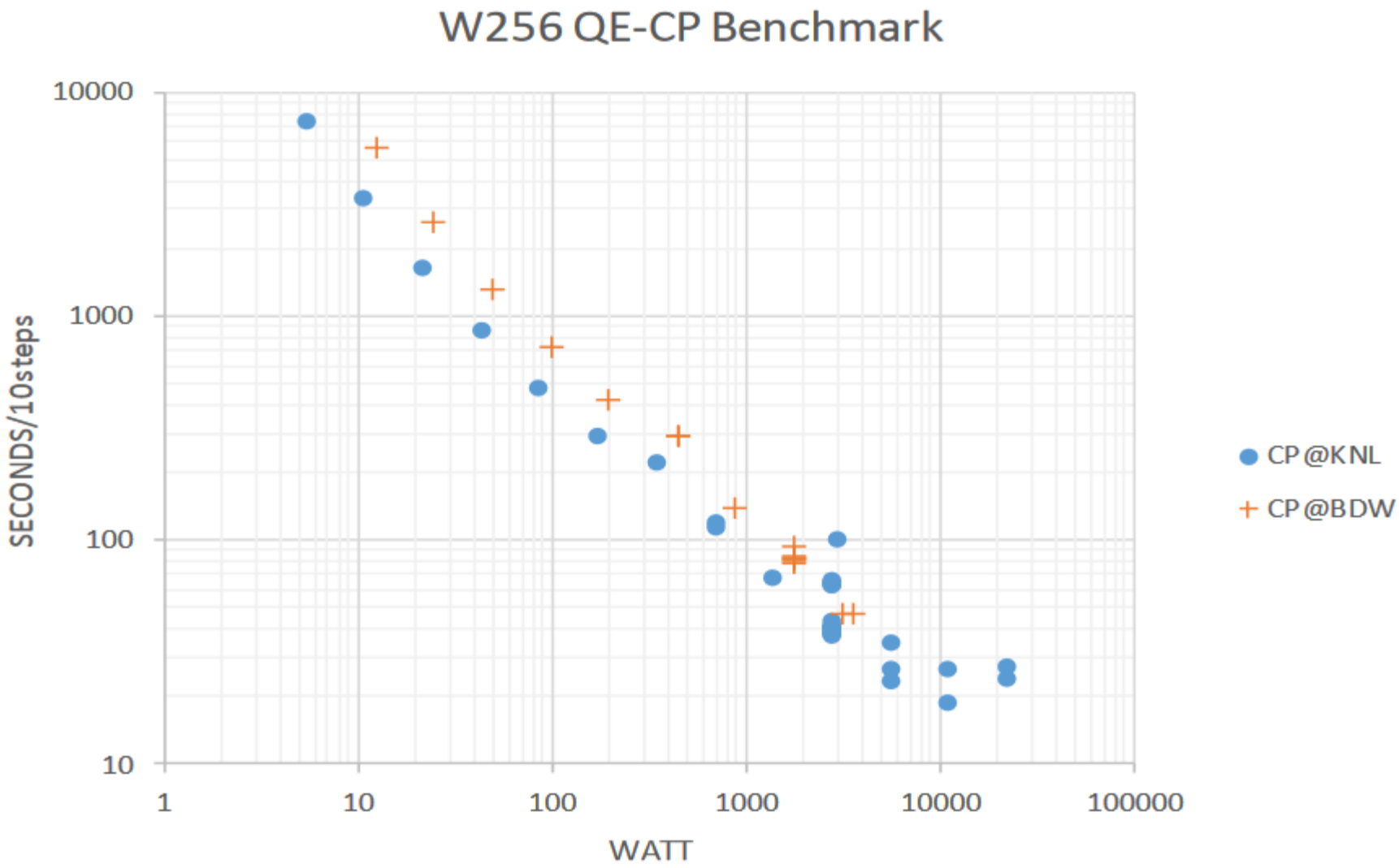
Benchmark: Quantum Espresso



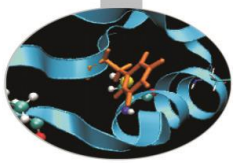


Benchmark: Quantum Espresso





Immediate future



- Currently, SkyLake partition is not available for all users. EUROFusion has reserved it and all the other users can work only on Broadwell and Knights Landing

- In the upcoming months, we will create another partition (A4?), based on another set of SkyLakes. This should make MARCONI able to reack theoretical performance of 20 PFlop/s overall, and will be made available to regular academic users.

- In order to make space for A4, a part of Broadwell partition will be removed. The nodes will be recycled on a new Tier-1 cluster, based on Infiniband rather than Omnipath interconnection network.