Fantastic HPC beasts and how to run on them

Quantum ESPRESSO Dev Meeting 2017

Fabio Affinito, Carlo Cavazzoni
Cineca

Summary

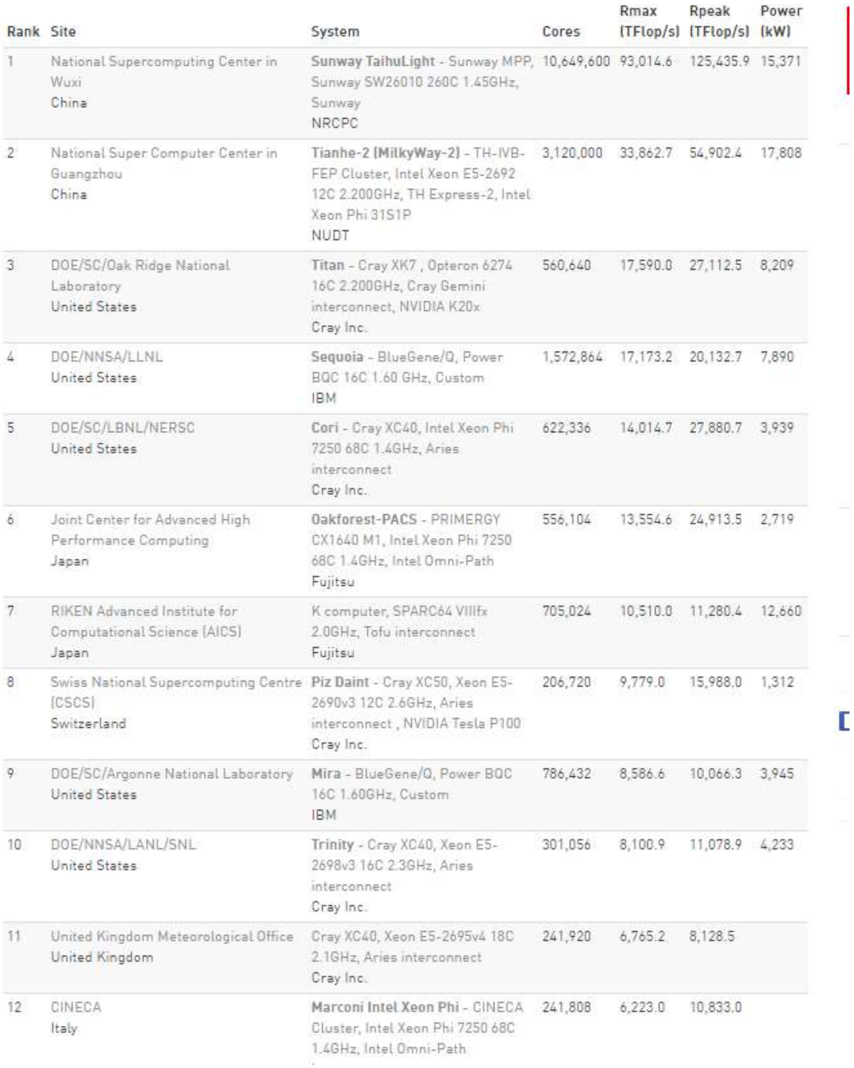
- What’s around in the HPC world
- Introducing MARCONI: the Cineca HPC infrastructure
- How to (happily) survive to MARCONI

Looking around us…

The Top500 list shows the status of the HPC facilities around the world:

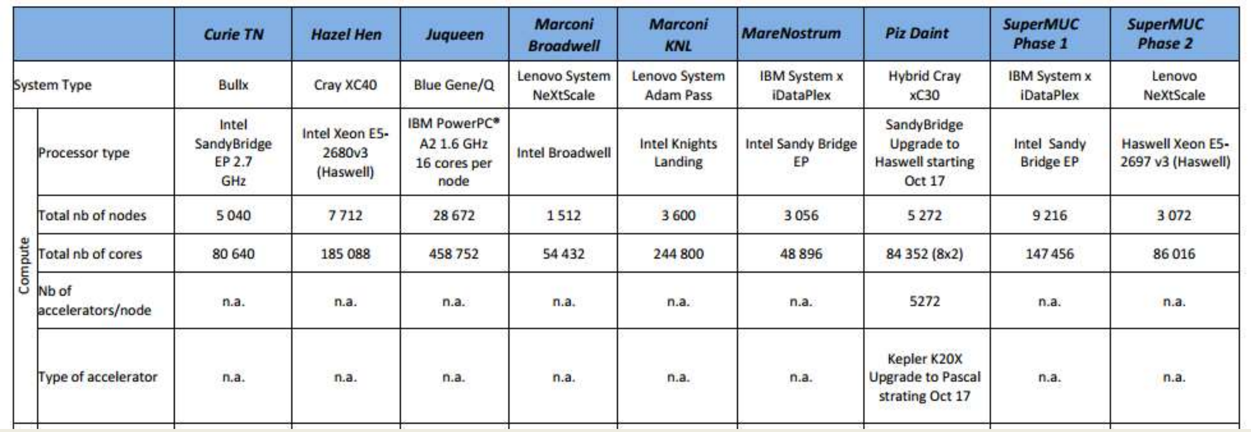
- Intel Xeon Phi appear at position 2, 5, 6.
- NVIDIA GPUs at position 3 and 8
- IBM BG/Q still occupy position 4 and 9

The IBM BG/Q architecture is at the end of its lifecycle and it is easy to see that the most important HPC facilities are based on many-cores architectures, namely Intel Xeon Phi and NVIDIA GPUs



Rank	Site	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (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	DOE/SC/LBNL/NERSC United States	**Cori** - Cray XC40, Intel Xeon Phi 7250 68C 1.4GHz, Aries interconnect Cray Inc.	622,336	14,014.7	27,880.7	3,939
6	Joint Center for Advanced High Performance Computing Japan	**Oakforest-PACS** - PRIMERGY CX1640 M1, Intel Xeon Phi 7250 68C 1.4GHz, Intel Omni-Path Fujitsu	556,104	13,554.6	24,913.5	2,719
7	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
8	Swiss National Supercomputing Centre (CSCS) Switzerland	**Piz Daint** - Cray XC50, Xeon E5-2690v3 12C 2.6GHz, Aries interconnect , NVIDIA Tesla P100 Cray Inc.	206,720	9,779.0	15,988.0	1,312
9	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
10	DOE/NNSA/LANL/SNL United States	**Trinity** - Cray XC40, Xeon E5-2698v3 16C 2.3GHz, Aries interconnect Cray Inc.	301,056	8,100.9	11,078.9	4,233
11	United Kingdom Meteorological Office United Kingdom	Cray XC40, Xeon E5-2695v4 18C 2.1GHz, Aries interconnect Cray Inc.	241,920	6,765.2	8,128.5	
12	CINECA Italy	**Marconi Intel Xeon Phi** - CINECA Cluster, Intel Xeon Phi 7250 68C 1.4GHz, Intel Omni-Path	241,808	6,223.0	10,833.0	

PRACE: the European infrastructure



		Curie TN	*Hazel Hen*	*Juqueen*	*Marconi Broadwell*	*Marconi KNL*	*MareNostrum*	*Piz Daint*	*SuperMUC Phase 1*	*SuperMUC Phase 2*
System Type		Bullx	Cray XC40	Blue Gene/Q	Lenovo System NeXtScale	Lenovo System Adam Pass	IBM System x iDataPlex	Hybrid Cray xC30	IBM System x iDataPlex	Lenovo NeXtScale
Compute	Processor type	Intel SandyBridge EP 2.7 GHz	Intel Xeon E5-2680v3 (Haswell)	IBM PowerPC® A2 1.6 GHz 16 cores per node	Intel Broadwell	Intel Knights Landing	Intel Sandy Bridge EP	SandyBridge Upgrade to Haswell starting Oct 17	Intel Sandy Bridge EP	Haswell Xeon E5-2697 v3 (Haswell)
	Total nb of nodes	5 040	7 712	28 672	1 512	3 600	3 056	5 272	9 216	3 072
	Total nb of cores	80 640	185 088	458 752	54 432	244 800	48 896	84 352 (8x2)	147 456	86 016
	Nb of accelerators/node	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	5272	n.a.	n.a.
	Type of accelerator	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	Kepler K20X Upgrade to Pascal strating Oct 17	n.a.	n.a.

Also Tier-0 machines are moving towards system equipped with Intel Xeon Phi (KNL) and NVIDIA GPUs

The italian infrastructure: MARCONI

Partition A1

1512 Lenovo NeXtScale Server > 2PFlops
Intel E5-2697 v4 Broadwell
18 cores @ 2.3GHz. 128GByte x node

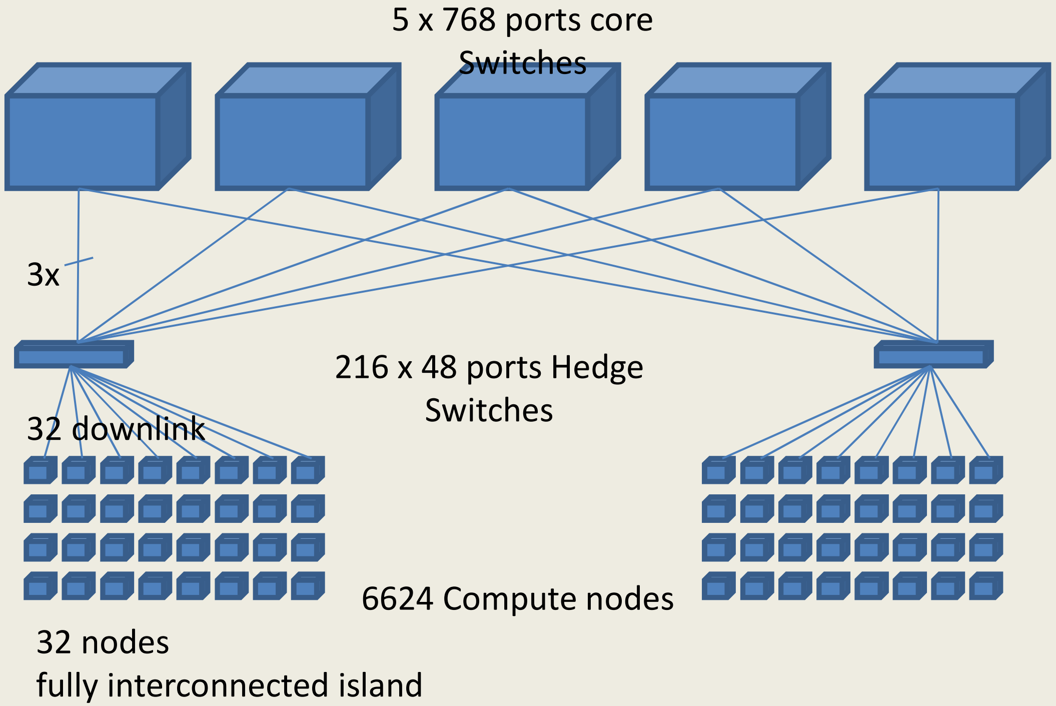
Partition A2

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

Partition A3

1512 Lenovo Stark Server > 4.5PFlops
Intel E5-26XXv5 SkyLake
2? cores @ 2.??GHz. 196GByte x node

Intel OmniPath interconnect



5 x 768 ports core Switches

3x

216 x 48 ports Hedge Switches

32 downlink

6624 Compute nodes

32 nodes
fully interconnected island

Marconi A1: Intel Broadwell

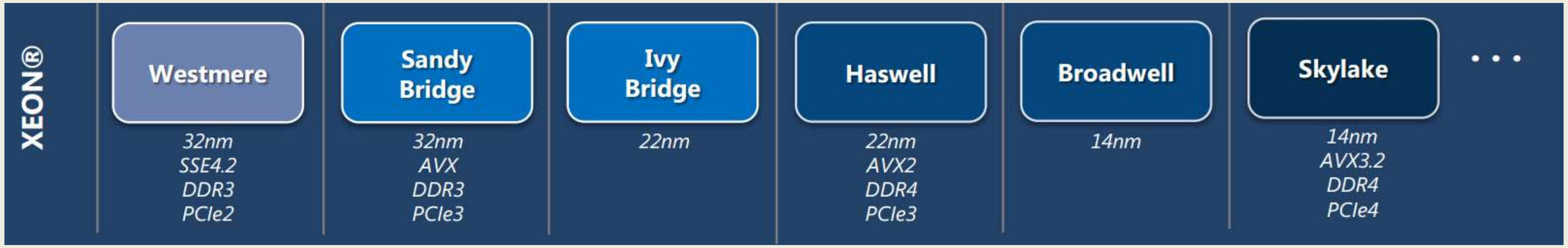
Not so much different from previous families of Intel Xeon CPUs

Single core is quite similar to Haswell (cfr GALILEO)

But… 36 cores per node: beware of the bandwith usage

Sometimes using OpenMP produces bad performances: use carefully

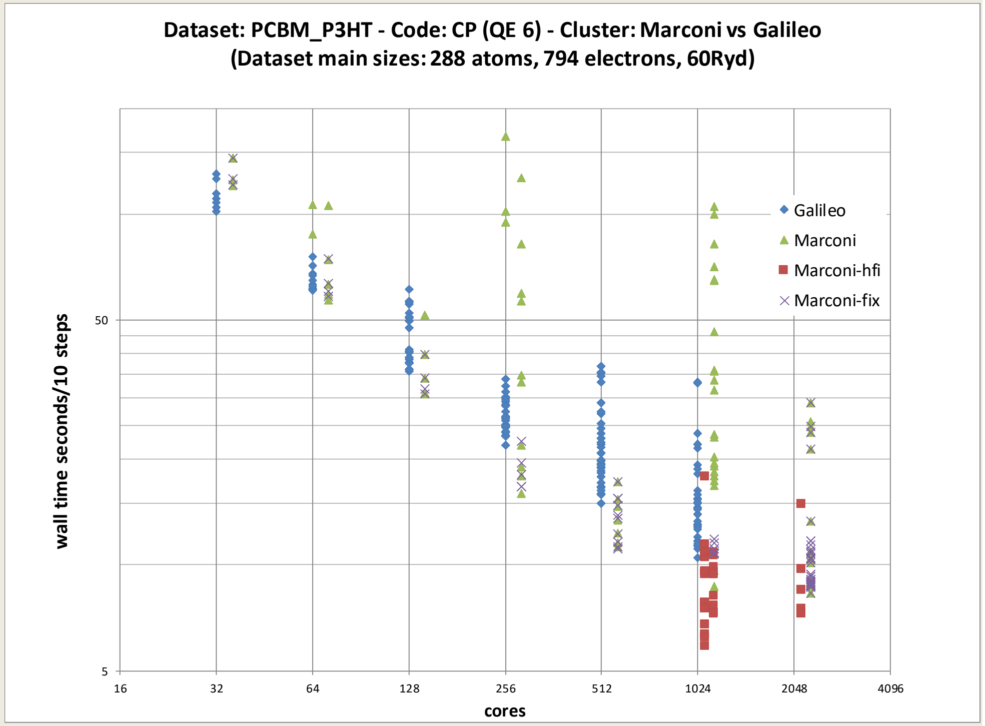
- Maybe binding cores using affinity can help



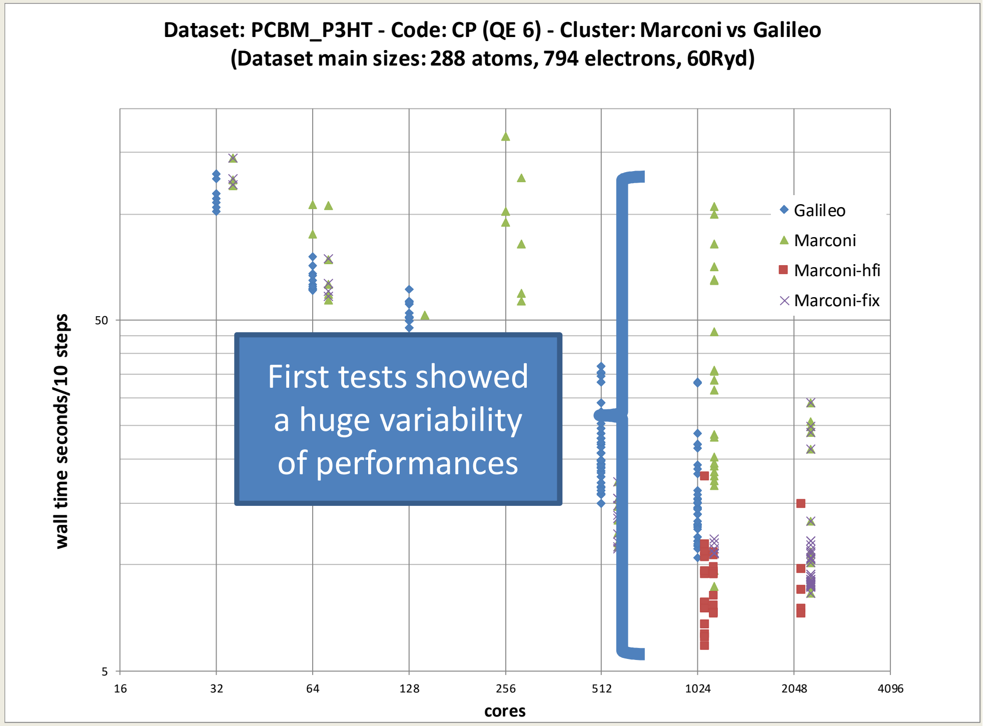
XEON®

Westmere	Sandy Bridge	Ivy Bridge	Haswell	Broadwell	Skylake	• • •
32nm SSE4.2 DDR3 PCIe2	32nm AVX DDR3 PCIe3	22nm	22nm AVX2 DDR4 PCIe3	14nm	14nm AVX3.2 DDR4 PCIe4	

But life ain’t easy

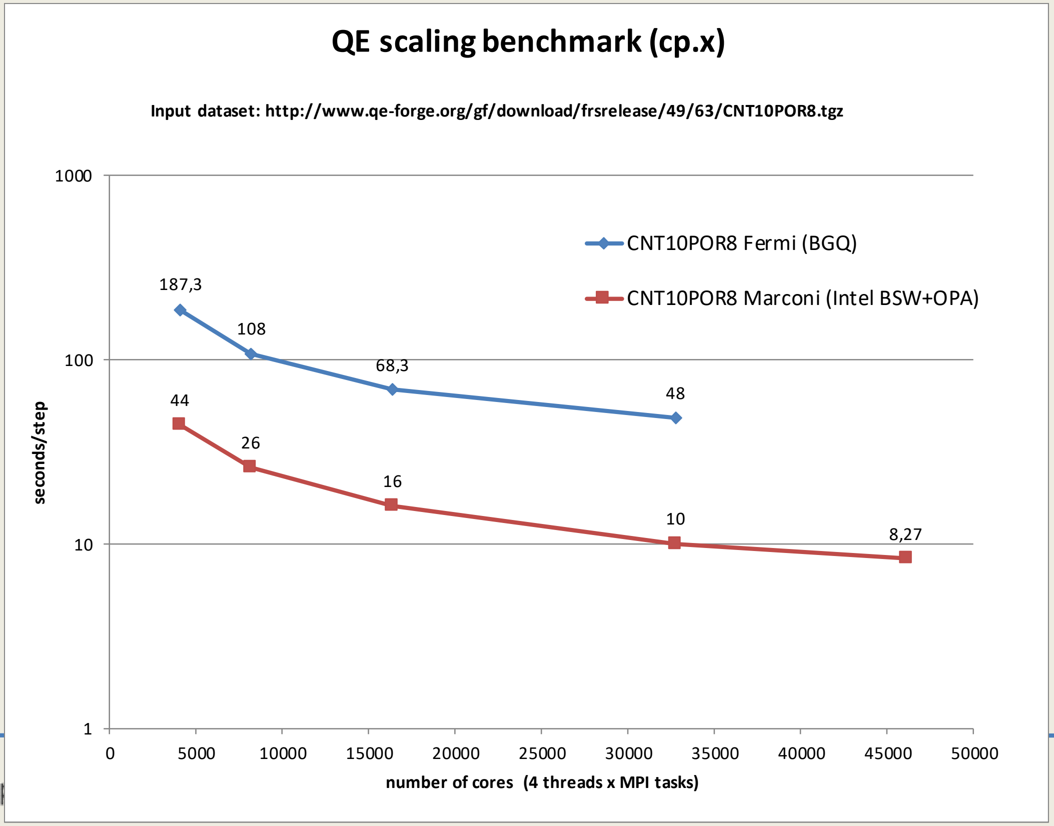


But life ain’t easy



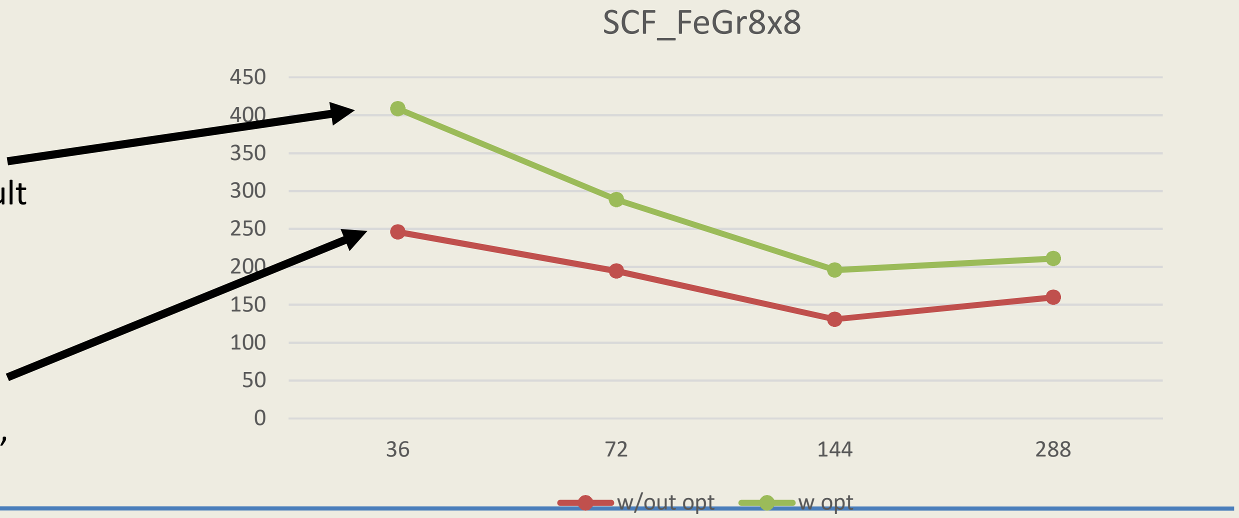
First tests showed a huge variability of performances

With a little patience…



Exploit parallelism levels

In order to run efficiently on MARCONI doesn't require any porting (i.e. coding, etc.), but you should be able to wisely exploit the existing parallelism



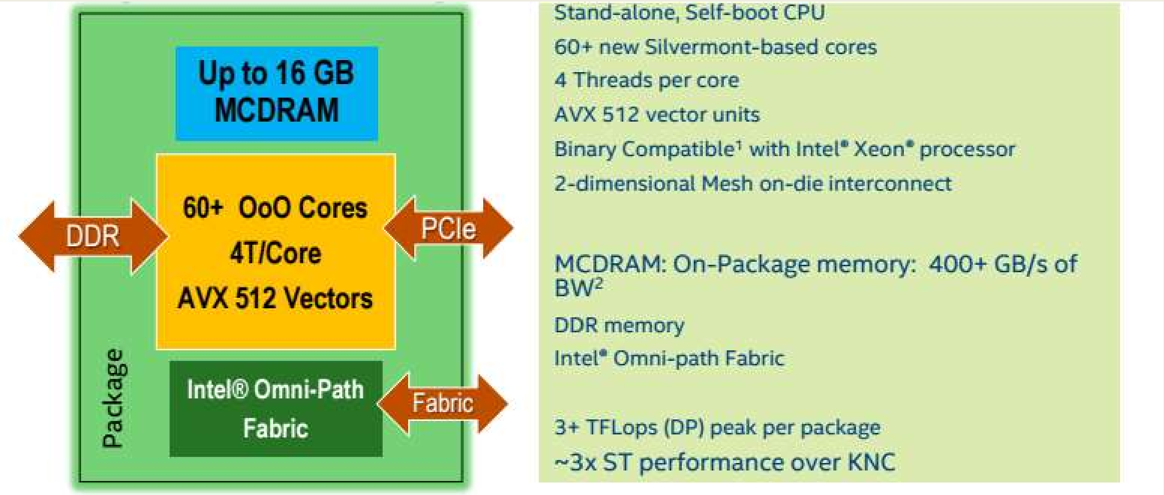
Run with default parameters

Run with the optimal choice of bgrp, npools, ntask and ndiag

MARCONI A2: Introducing KNL

Exploiting the parallelism is way more important with the KNL platform: 68 cores!

Differently from the Intel Xeon Phi KNC, this is not a co-processor.



Stand-alone, Self-boot CPU

60+ new Silvermont-based cores

4 Threads per core

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

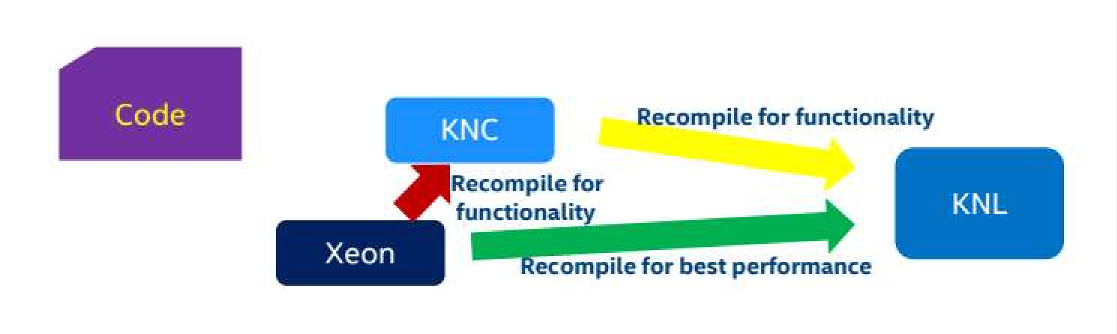
~3x ST performance over KNC

«living together in harmony» with the KNL

Key features:	How to survive them:
68 cores	When running exploit parallelism: Use all the hierarchical parallelism inside QE (i.e. pools, bands, taskgroups) using both MPI and OpenMP.
MCDRAM	When coding improve data locality: reuse data structures as much as possible. If MCDRAM is configured in cache-mode, this can improve the performances
AV512	Write loops such that they can be easily vectorized by the compiler. Use clean code techniques and check the vectorization report of the Intel compiler to get help

A smooth transition?

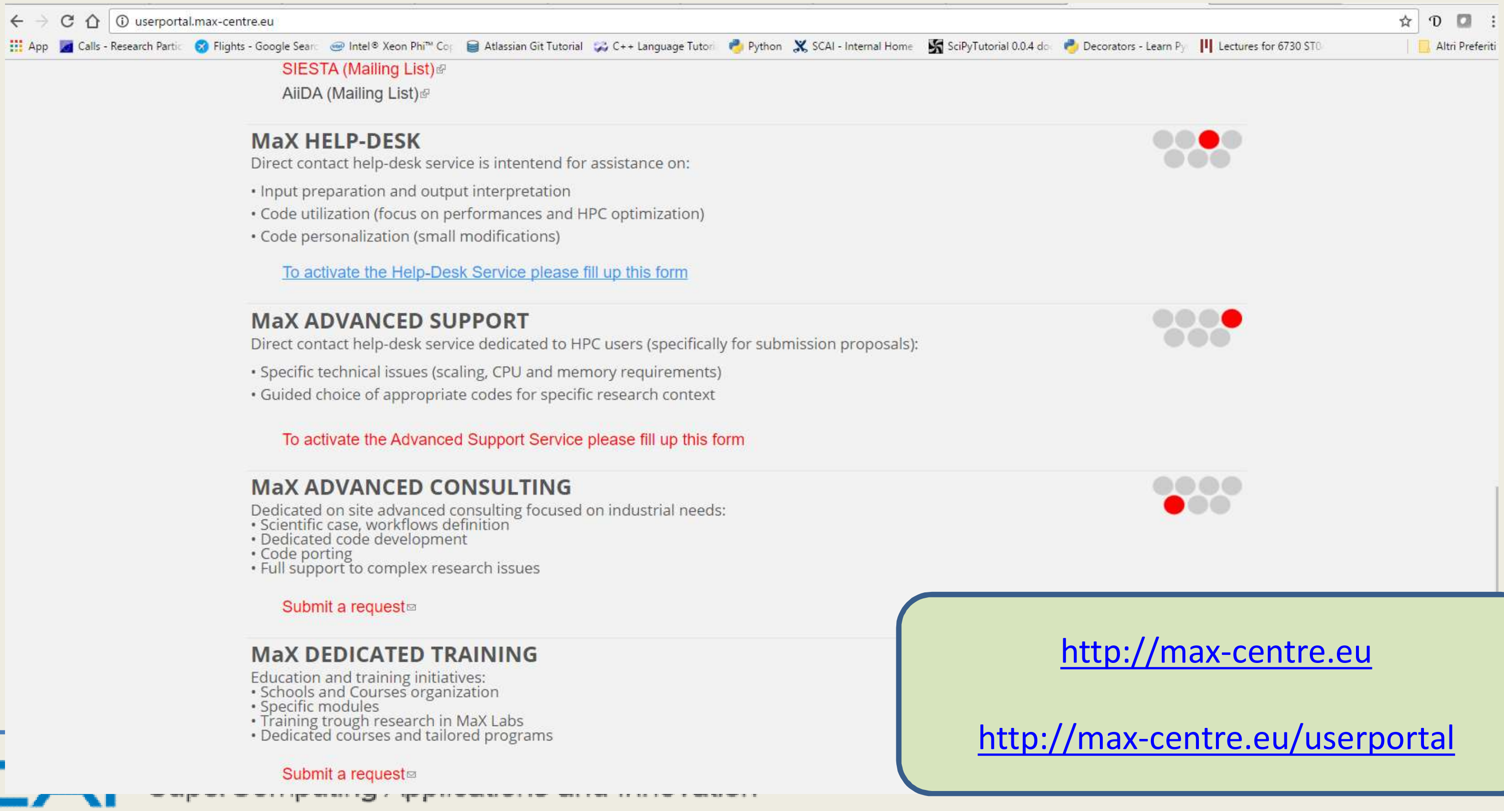
Yes, in principle


Code
KNC
Recompile for functionality
Recompile for functionality
Xeon
KNL
Recompile for best performance

But in order to properly use KNL:

- Use both MPI and OpenMP
- Don’t forget to compile with –xMIC-AVX512 to switch on AVX512

When in trouble, ask MaX’s support!



SIESTA (Mailing List)

AiiDA (Mailing List)

MaX HELP-DESK

Direct contact help-desk service is intentend for assistance on:

- Input preparation and output interpretation
- Code utilization (focus on performances and HPC optimization)
- Code personalization (small modifications)

To activate the Help-Desk Service please fill up this form

MaX ADVANCED SUPPORT

Direct contact help-desk service dedicated to HPC users (specifically for submission proposals):

- Specific technical issues (scaling, CPU and memory requirements)
- Guided choice of appropriate codes for specific research context

To activate the Advanced Support Service please fill up this form

MaX ADVANCED CONSULTING

Dedicated on site advanced consulting focused on industrial needs:

- Scientific case, workflows definition
- Dedicated code development
- Code porting
- Full support to complex research issues

Submit a request

MaX DEDICATED TRAINING

Education and training initiatives:

- Schools and Courses organization
- Specific modules
- Training trough research in MaX Labs
- Dedicated courses and tailored programs

Submit a request

http://max-centre.eu

http://max-centre.eu/userportal