Marconi- Molecular Dynamics and New technologies

Preview of the new Tier-0 resource at Cineca and what it means for researchers in MD

BG/Q (Fermi) as a Tier0 Resource

- Many advantages as a supercomputing resource:
  - Low energy consumption.
  - Limited floor space requirements
  - Fast internal network
  - Homogeneous architecture → simple usage model.
- But
  - Low, single core performance + I/O structure meant very high parallelism necessary (at least 1024 cores).
  - For some applications (e.g QM) low memory/core (1Gb) and I/O performance also a problem. Also limited capabilities of O.S. on compute cores (e.g. no interactive access)
  - Cross compilation, because login nodes different to compute nodes, can complicate some build procedures.

FERMI scheduled to be decommissioned mid-end 2016

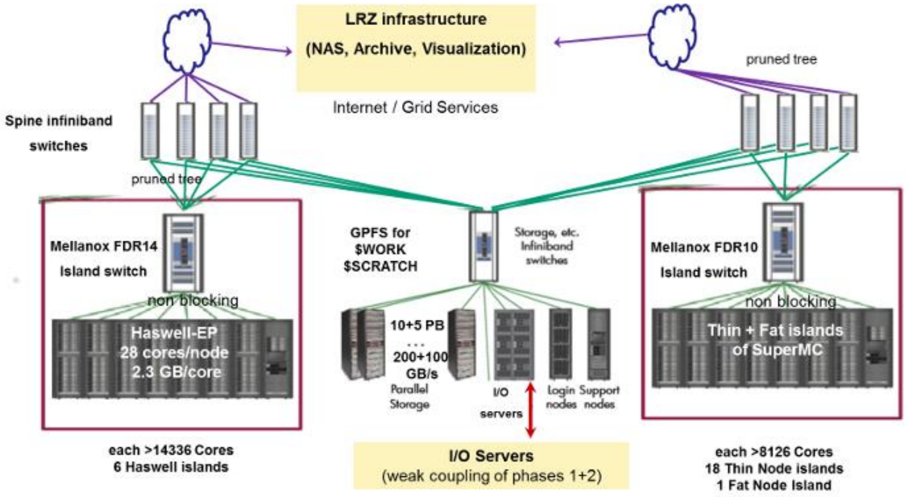
Replacing Fermi at Cineca - considerations

- A new procurement is a complicated process and considers many factors but must include (together with the price):
  - Minimum peak compute power
  - Power consumption
  - Floor space required
  - Availability
  - Disk space, internal network, etc.
- IBM no longer offers the BlueGene range for supercomputers so cannot be a solution.
- Many computer centres are adopting instead a *heterogenous* model for computer clusters

Heterogenous clusters

SuperMUC, LRZ (Germany)

System type	Bullx system built by Bull
Full system	40,960 cores + 132 GPUs: 1.559 Pflop/s (peak performance)
Thin nodes (Haswell)	25,920 cores: 1.078 Pflop/s
Thin nodes (Ivy Bridge)	12,960 cores: 249 Tflop/s
GPU nodes (K40m)	1,056 cores + 132 GPUs: 210 Tflop/s
Fat nodes (Sandy Bridge)	1,024 cores: 22 Tflop/s
Memory	117 TB memory (CPU + GPGPU)
Disk space	180 TB home file systems, 7.7 PB scratch and project



1. Fat node islands
2. Thin node islands
3. Haswell node islands

Cartesius, SurfSara (the Netherlands)

Different hardware resources according to application needs

Replacing Fermi – the Marconi solution

- The Fermi replacement, called Marconi, will be built by Lenovo using Intel CPUs.
- Some details still being decided but Marconi will consist of different types of processors arriving in phases:
  1. ***Phase 1*** (by Summer 2016*): 1512 Broadwell nodes, 2Pflops.
  2. ***Phase 2*** (end 2016*): 3600 Intel Knights Landing (KNL), 11 Pflops.
- The whole system will also have a total storage of ~10Pb of disk. All nodes will be connected via Intel Omni-Path network.
- Possible upgrade in 2017 with Intel Skylake processors.

* **All dates are, of course, approximate.**

Replacing Fermi with Marconi

But what will this mean for Molecular Dynamics@Cineca ?

*not Marconi obviously

Replacing Fermi with Marconi

- Without Marconi physically here we can only make predictions.
- General observations:
  - – Single processor cores more powerful than Fermi, so very high parallelism no longer essential (but may still be required by some calls).
  - – Different types of processors and nodes means must choose where to run simulations.
  - – Possibility of interactive access should help testing.
  - – Intel hardware more likely to be supported by application developers.

Using MD 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 MD programmers and users.

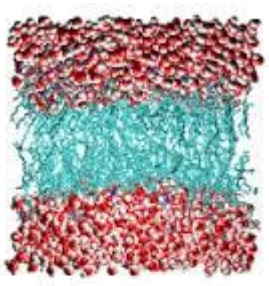
cores/node	36
Memory/node	128 GB

Broadwell benchmarks

Gromacs DPPC (1 core)

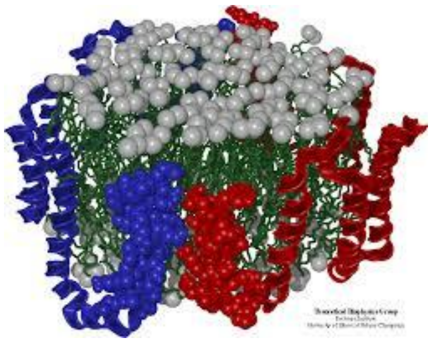
Computer system	ns/day	Speedup wrt Fermi
Haswell (5.0.4, Galileo)	1.364	13.64
Fermi (5.0.4)	0.100	1.00
Broadwell (5.1.2)	1.977	19.77

Based on a 2-node Broadwell partition (80 cores/node).



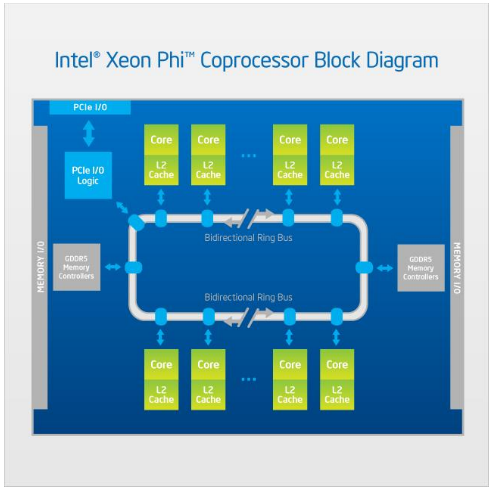
NAMD APOA1 (16 tasks)

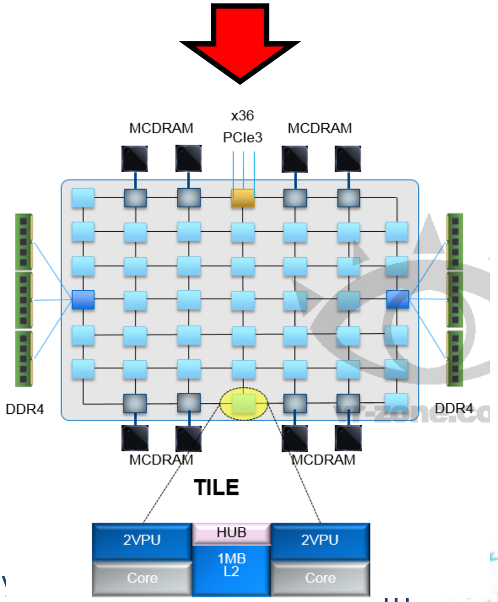
Computer System	ns/day	Speedup wrt Fermi
Haswell (2.10, Galileo)	1.425	7.27
Fermi (2.10)	0.196	1.00
Broadwell (2.11)	1.516	7.73



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 respect to KNC.
  - The main differences are:
    - KNL will be a standalone processor not an accelerator (unlike KNC)
    - KNL has more powerful cores and faster internal network.
    - On package high performance, memory (16Gb, MCDRAM).





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.

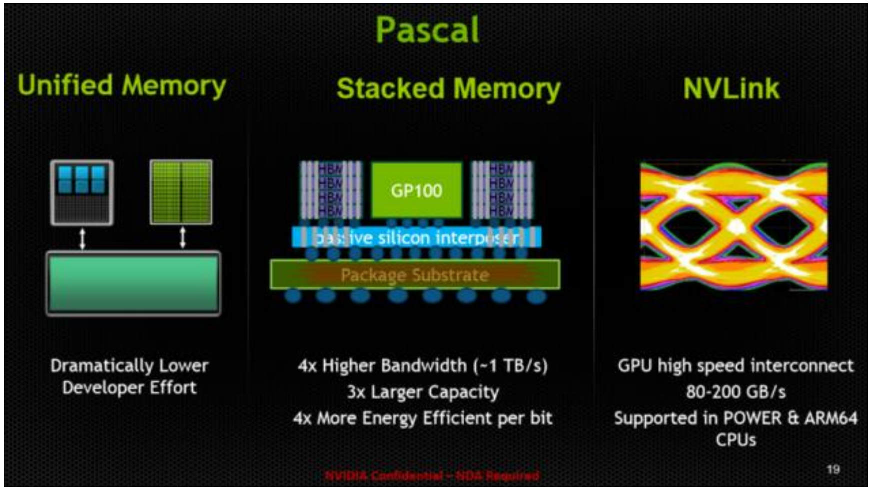
Using MD on Marconi-Phase II

- Programmers must utilise vectorisation (SIMD) and OpenMP threads, and possibly the fast memory of KNL.
- For the user, MD experience will depend on how software developers are able to exploit the KNL architecture:
  - NAMD. Already reasonable results with KNC. According to NAMD mailing list much effort being devoted to KNL version.
  - GROMACS. Developers didn't really bother with KNC Xeon Phi's (no offload version and poor symmetric mode). But since KNL is standalone and Gromacs can use OpenMP threads (which are advisable on KNL) should run well on KNL. Also GROMACS has good SIMD optimisation.
  - Amber. Already support for KNC and with OpenMP probably should be ok for KNL.
  - LAMMPS. Current support for KNC via Kokkos package. Plans for KNL unknown.
  - DL_POLY. Plans for KNL unknown.
  - Desmond: Also ?

Worth noting that up to now KNC MICs haven't been widely supported by software developers. But this should change for KNL.

And GPUs?

- In HPC, an alternative to Intel is focussed on the **OpenPower** initiative which promotes IBM PowerPCs and accelerators such as GPUs.
- Particularly important PowerPC+Nvidia GPU (Pascal) with NVLink which will be used in two US supercomputers.
- NAMD is one of the benchmark codes for these systems.
- Cineca likely to have a small prototype system to monitor the technology.



Pascal

Unified Memory

Dramatically Lower Developer Effort

Stacked Memory

4x Higher Bandwidth (~1 TB/s)
3x Larger Capacity
4x More Energy Efficient per bit

NVLink

GPU high speed interconnect
80-200 GB/s
Supported in POWER & ARM64 CPUs

NVIDIA Confidential – NDA Required

19

Summary

- For classical MD the new Marconi platform should be much more productive due to more powerful cores and less need for very high parallel scalability. Most users will find Marconi easier to use than Fermi.
- Compared to Galileo, the Broadwell partition is only likely to show a small performance increase.
- The performances of MD codes on the Xeon Phi (KNL) partition are unknown but should be good for the most popular MD programs (NAMD, GROMACS and AMBER).
- Future upgrades could also include very fast disk space (Non-Volatile Memory).

Questions: What improved hardware features (e.g disks, memory, accelerators, etc) would you like for your simulations?