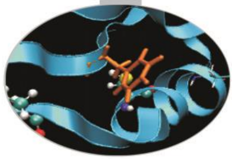
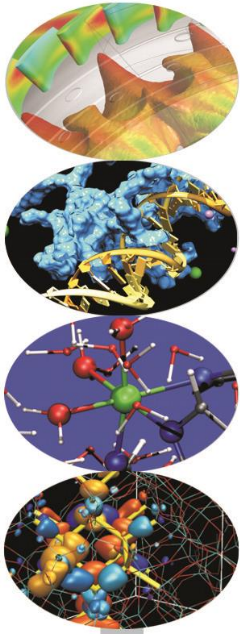
CINECA SCAI
SuperComputing Applications and Innovation

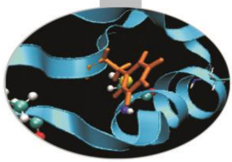


Overview of applications performance on Marconi

Piero Lanucara
p.lanucara@cineca.it
SCAI User Support team



CINECA

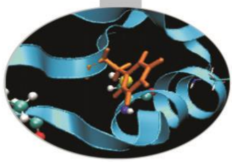


We would like to:

- Try to summarize the technological trend via benchmarks…
- …and use them to ***understand application performance issues, limitations and best practices on actual (Broadwell) and future architectures (KNL)***

CAVEAT

- ✓ All measurements was taken using HW at CINECA
- ✓ Sometimes there is an "unfair" comparison e.g.:
  - Sandy Bridge HW used was very "powerful", HPC oriented
  - Ivy Bridge HW used was devoted to "data crunching", not HPC oriented

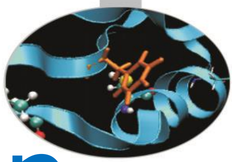


Tick/Tock

Intel CPU roadmap: two step evolution

- Tock phase:
  - ✓ New architecture
  - ✓ New instructions (ISA)

- Tick phase:
  - ✓ Keep previous architecture
  - ✓ New technological step (e.g. Broadwell → 14nm)
  - ✓ Core "optimization"
  - ✓ Usually increasing core number, keeping Thermal Dissipation (TDP) constant

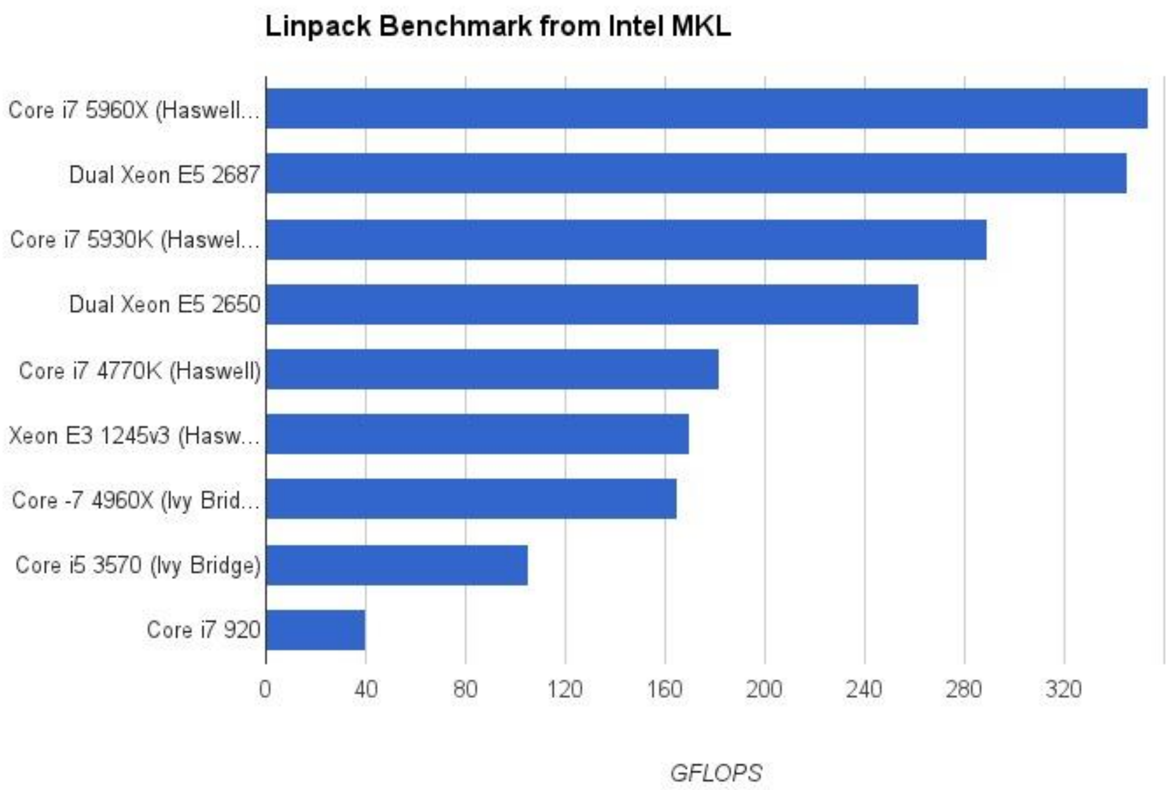


The Roadmap

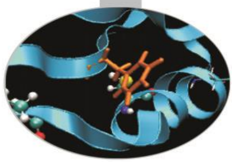
- **Westmere (tick, a.k.a. plx.cineca.it)**
  - Intel(R) Xeon(R) CPU E5645 @2.40GHz, 6 Core per CPU
  - Only serial performance figure
- **Sandy Bridge (tock, a.k.a. eurora.cineca.it)**
  - Intel(R) Xeon(R) CPU E5-2687W 0 @3.10GHz, 8 core per CPU
  - Serial/Node performance figure
- **Ivy Bridge (tick, a.k.a pico.cineca.it)**
  - Intel(R) Xeon(R) CPU E5-2670 v2 @2.50GHz, 10 core per CPU
  - Serial/Node/Cluster performance
  - Infiniband FDR
- **Hashwell (tock, a.k.a. galileo.cineca.it)**
  - Intel(R) Xeon(R) CPU E5-2630 v3 @2.40GHz, 8 core per CPU
  - Serial/Node/Cluster performance
  - Infiniband QDR
- **Broadwell (tick)**
  - Intel(R) Xeon(R) CPU E5-2699 v4 @ 2.20GHz, 22 core per CPU
  - Serial/Node performance figure

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

Benchmarks



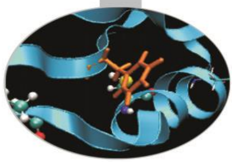
Linpack Benchmark from Intel MKL
Core i7 5960X (Haswell...
Dual Xeon E5 2687
Core i7 5930K (Haswel...
Dual Xeon E5 2650
Core i7 4770K (Haswell)
Xeon E3 1245v3 (Hasw...
Core -7 4960X (Ivy Brid...
Core i5 3570 (Ivy Bridge)
Core i7 920
0
40
80
120
160
200
240
280
320
GFLOPS



Performances

- Empirically tested on different HW at CINECA
  - **LINPACK**
    - Intel optimized benchmark, rel. 11.3
    - Stress Floating point performance, no Bandwidth limitation
  - **STREAM**
    - Rel. 3.6, OMP version
    - Bandwidth, no Floating point limitation
  - **HPCG**
    - Intel optimized benchmark, rel. 11.3
    - CFD oriented benchmark with Bandwidth Limitation

SCAI
SuperComputing Applications and Innovation

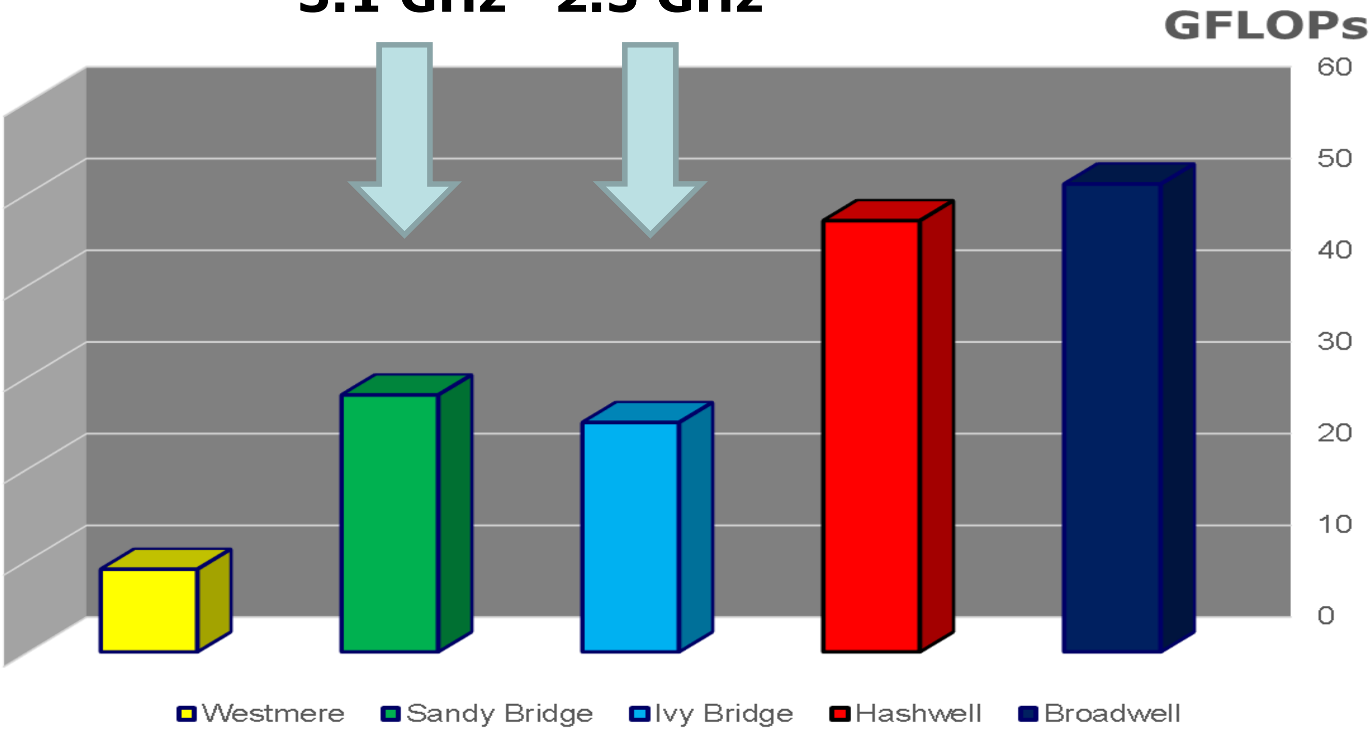


LINPACK

- Best result obtained, single core

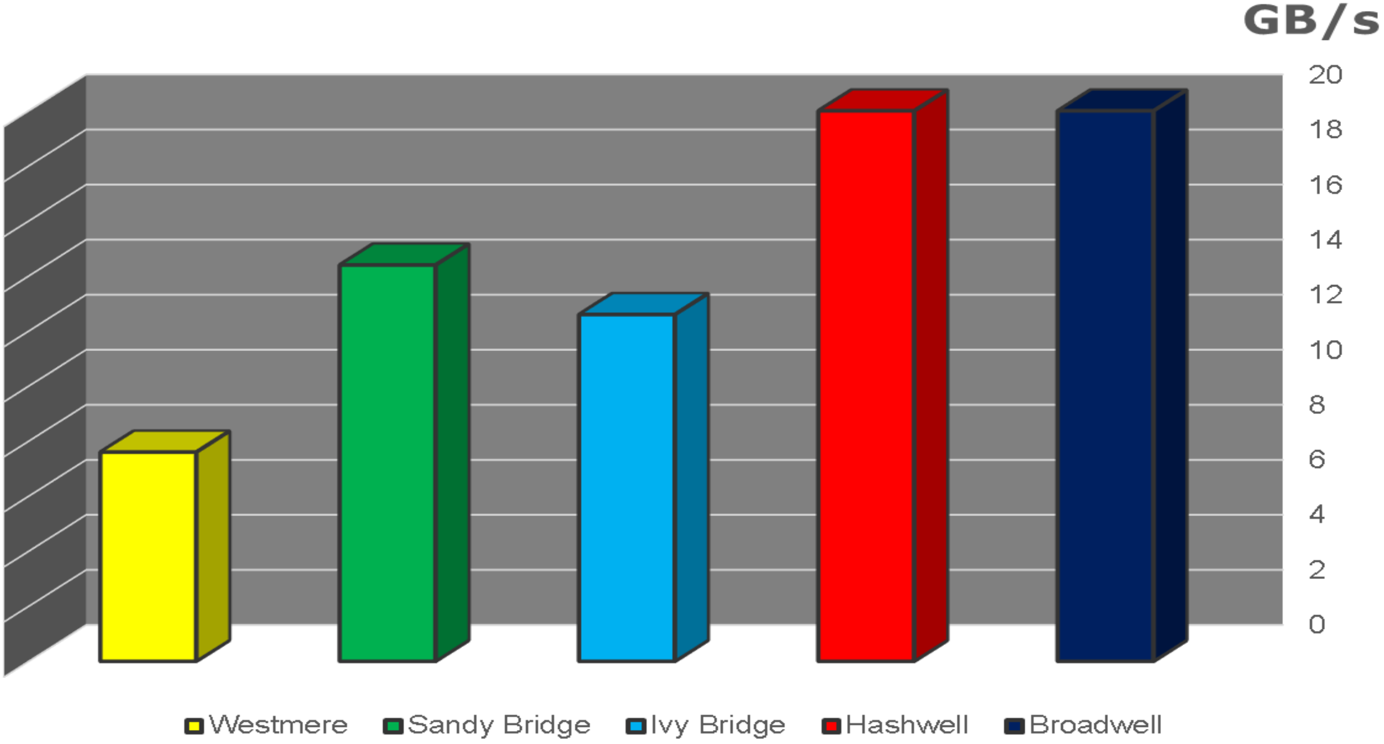
✓ 5.6x increase in 6 years (Q1-2010, Q1-2016)

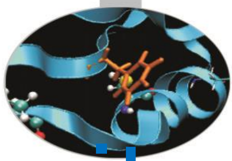
3.1 GHz 2.5 GHz



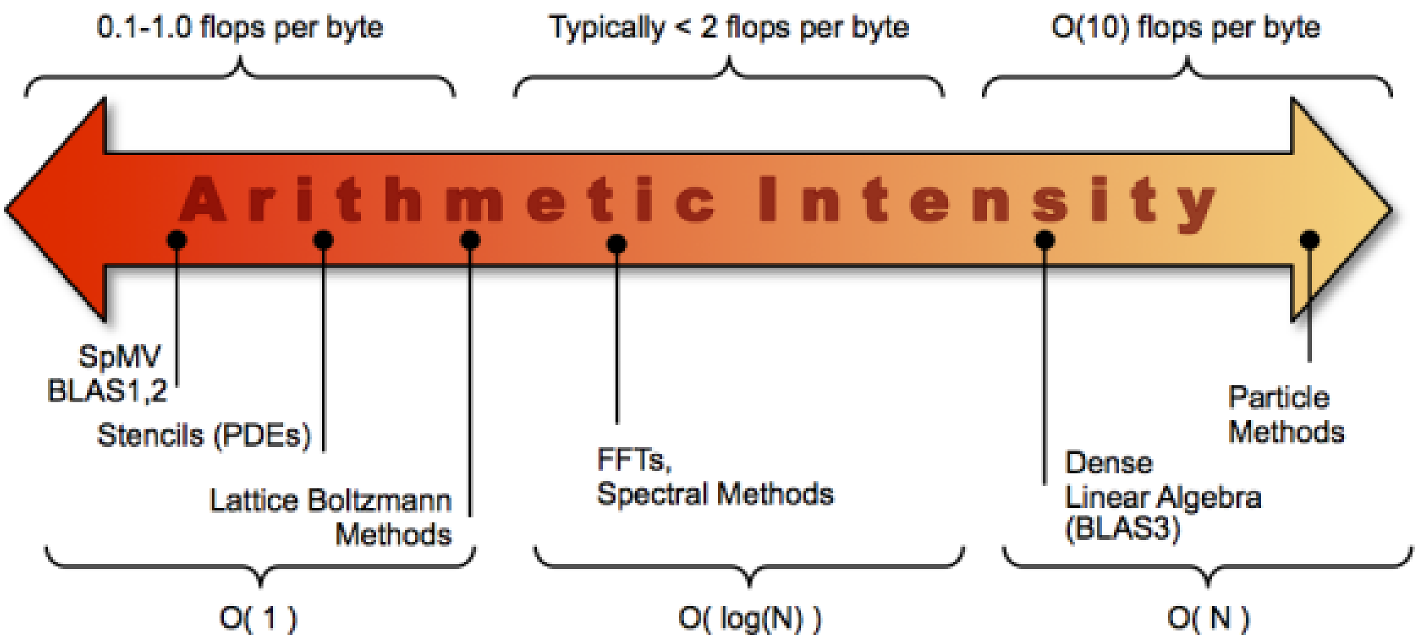
STREAM

- Best result obtained (using intel/gnu), single core
- 2.6x speed-up in 6 years ……☹





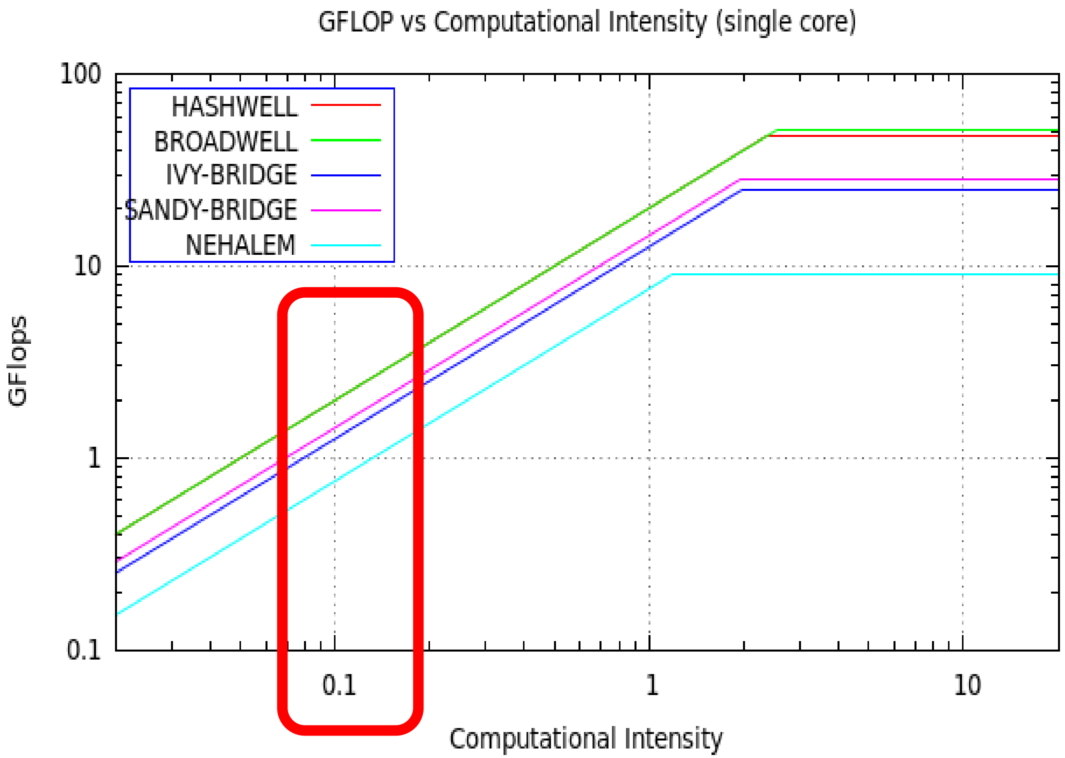
Roofline Model: Arithmetic Intensity



- Which is the typical application arithmetic intensity?
- About 0.1, may be less…. ☹
- It depends on application domain, solver, method,…

Roofline Mode: serial figure

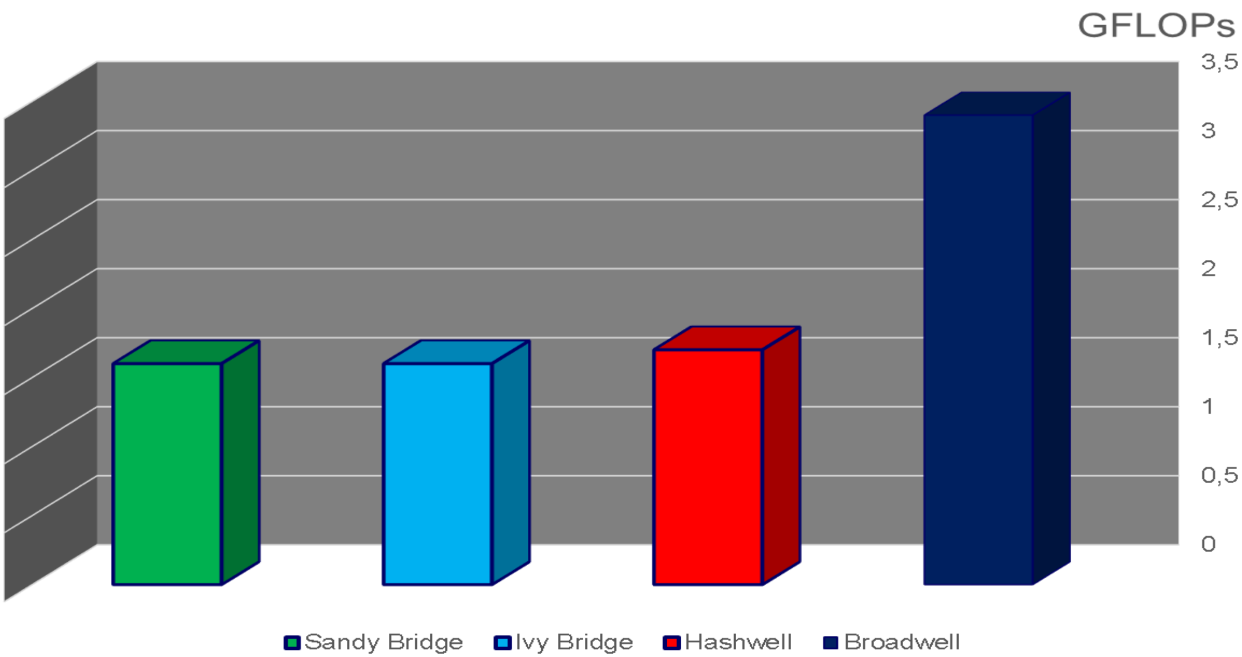
- Using the figures obtained on different HW (LINPACK, STREAM)



SCAI
SuperComputing Applications and Innovation

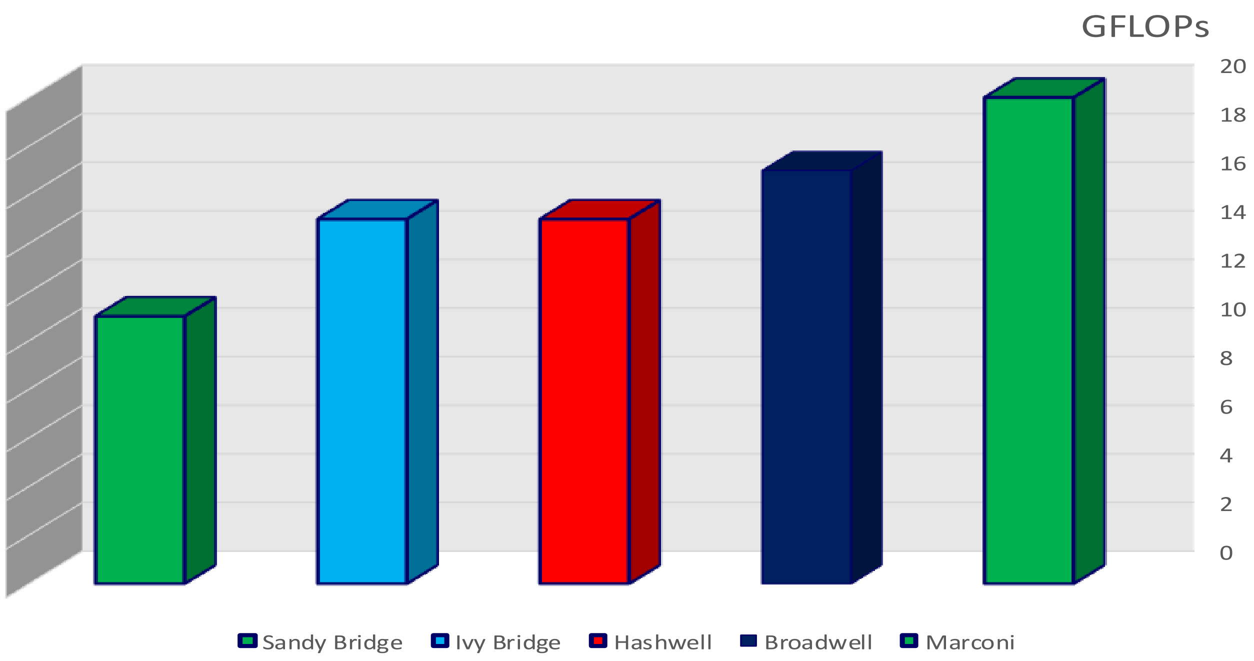
HPCG

- Conjugate Gradient Benchmark (http://hpcg-benchmark.org/)
- Intel benchmark: Westmere not supported
- 2x speed-up only for Broadwell



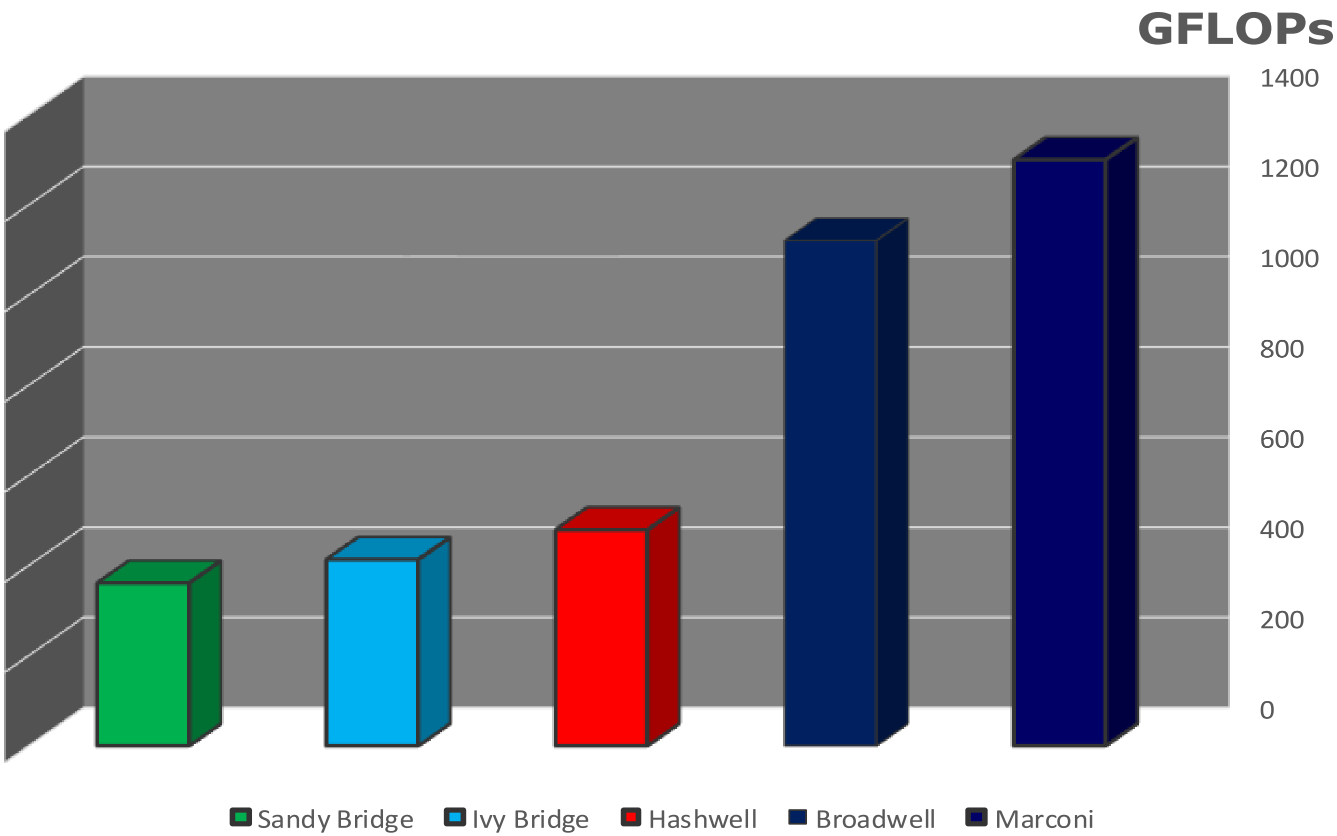
HPCG parallel figure

- Best performance with #tasks and #threads



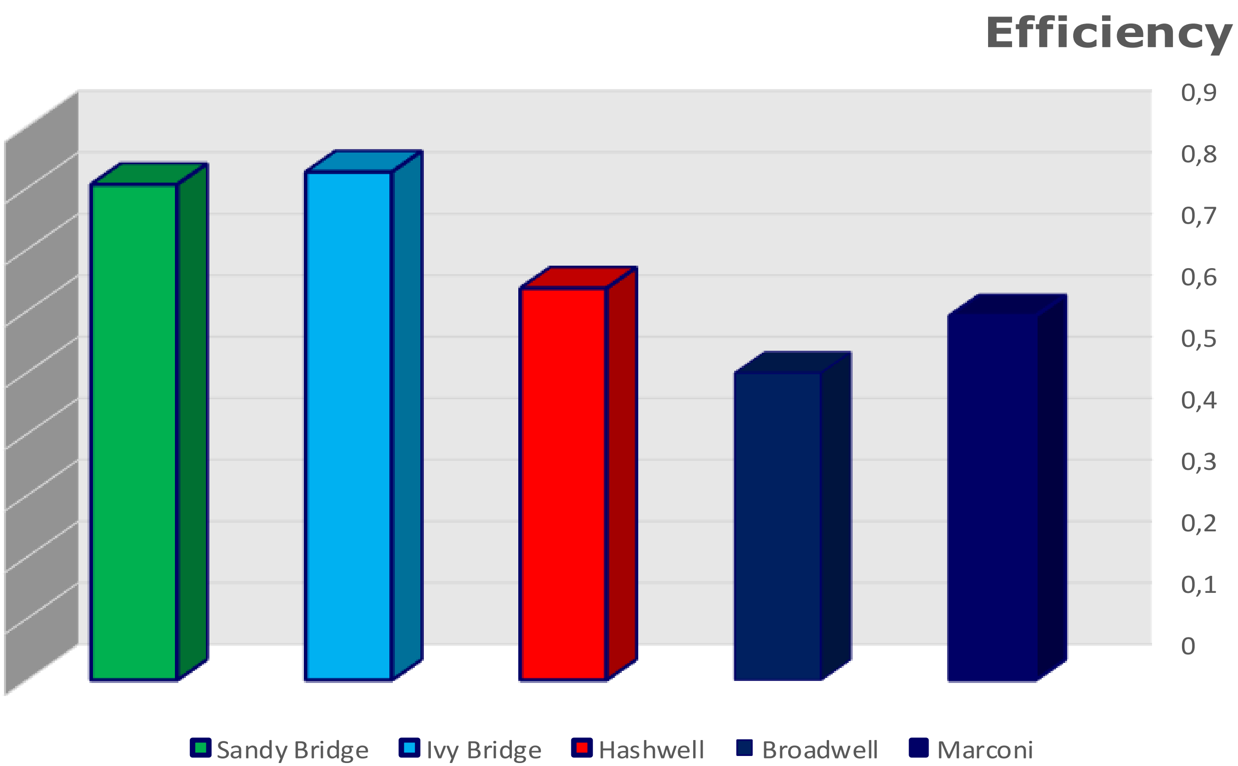
LINPACK parallel figure

- Best result obtained: Marconi (1 MPI, 36 threads)



LINPACK parallel figure/2

- Best result obtained
- Efficiency = Parallel_Flops/(#core*Serial_Flops)
  - 1 → Linear speed-up



SCAI SuperComputing Applications and Innovation

Marconi – A1 HPL

Full system Linpack:

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

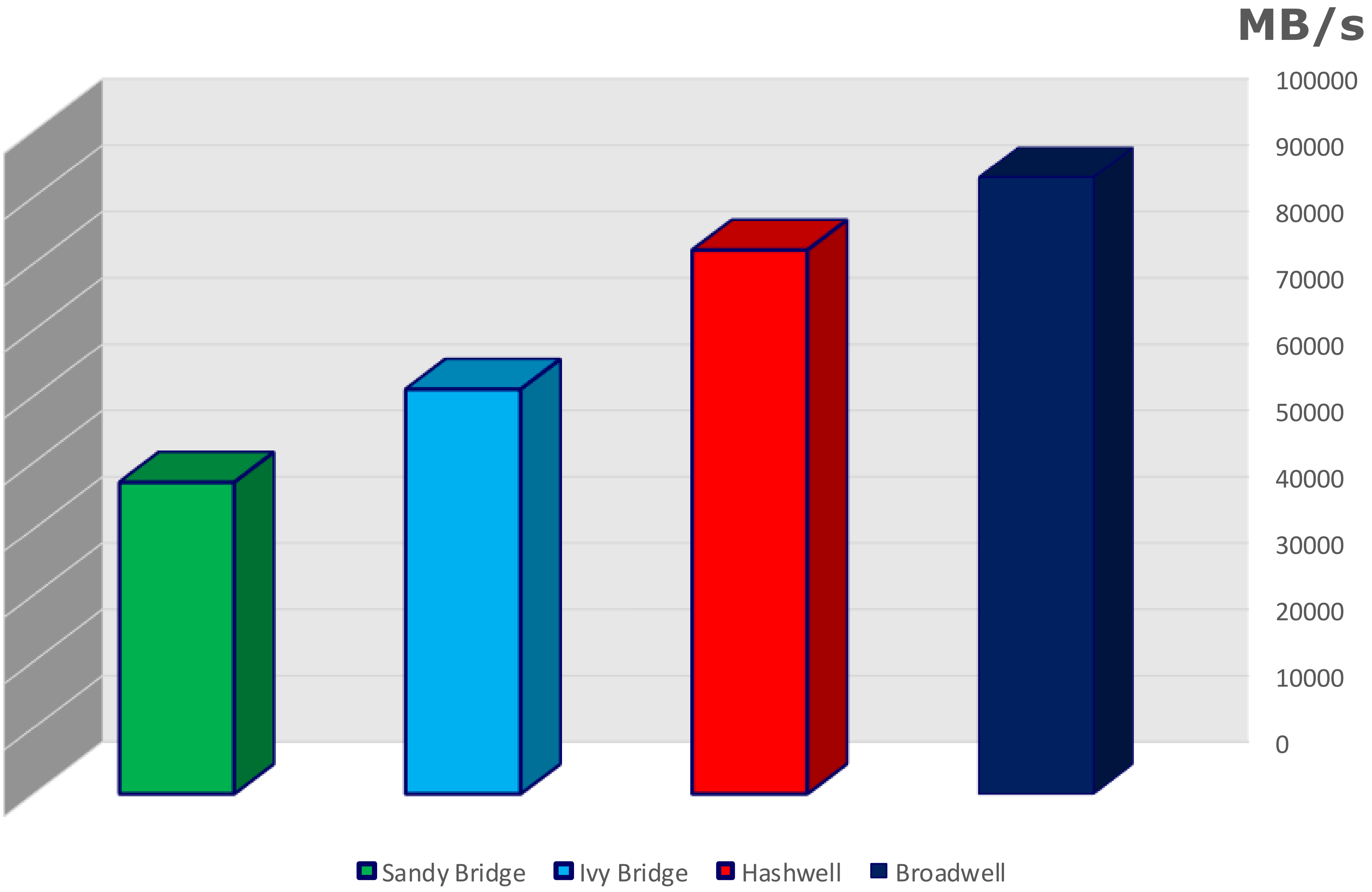
TOP500 The List.

June 2016:Number 46

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

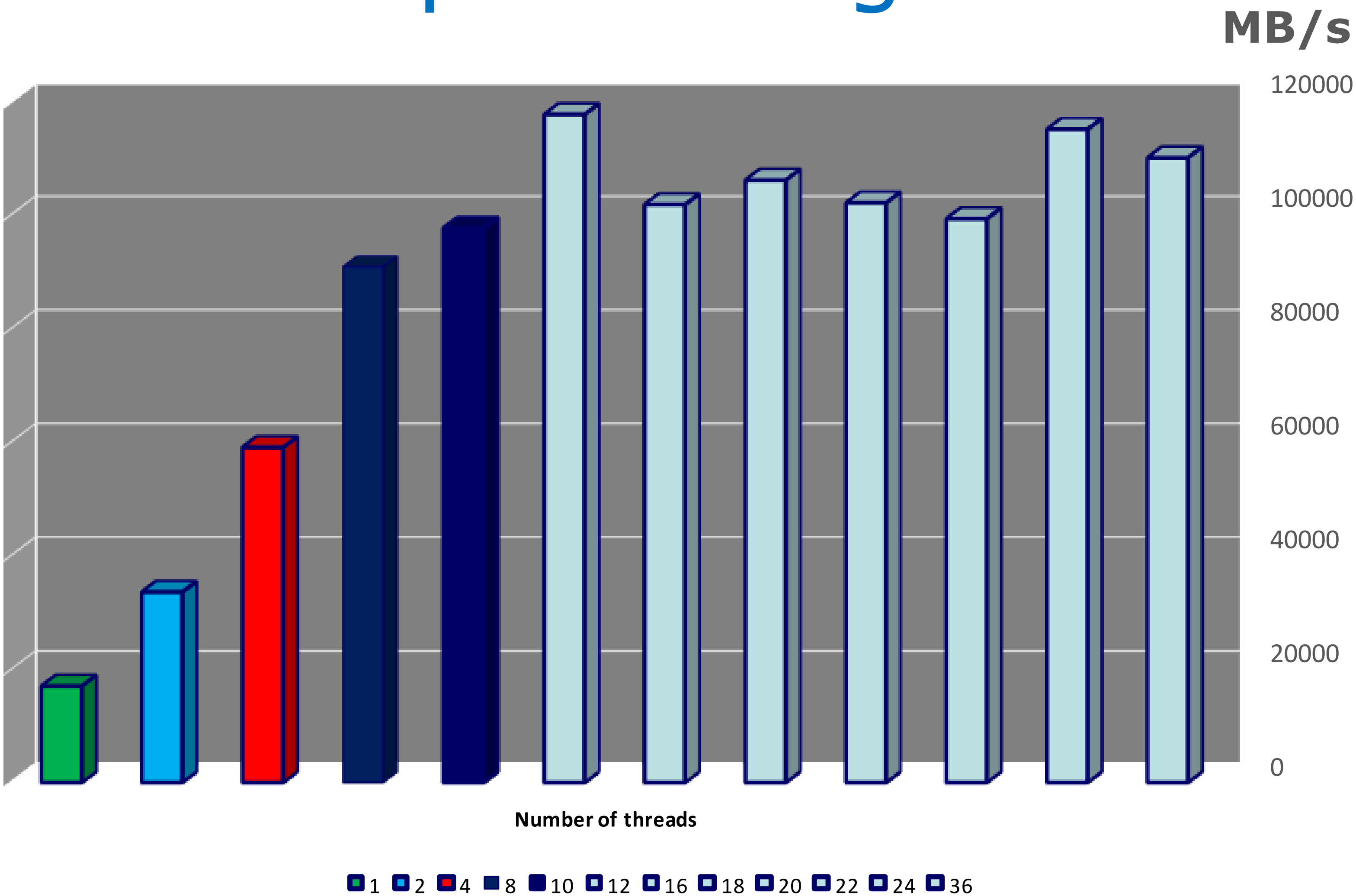
SuperComputing Applications and Innovation

STREAM parallel figure



CINECA SCAI
SuperComputing Applications and Innovation

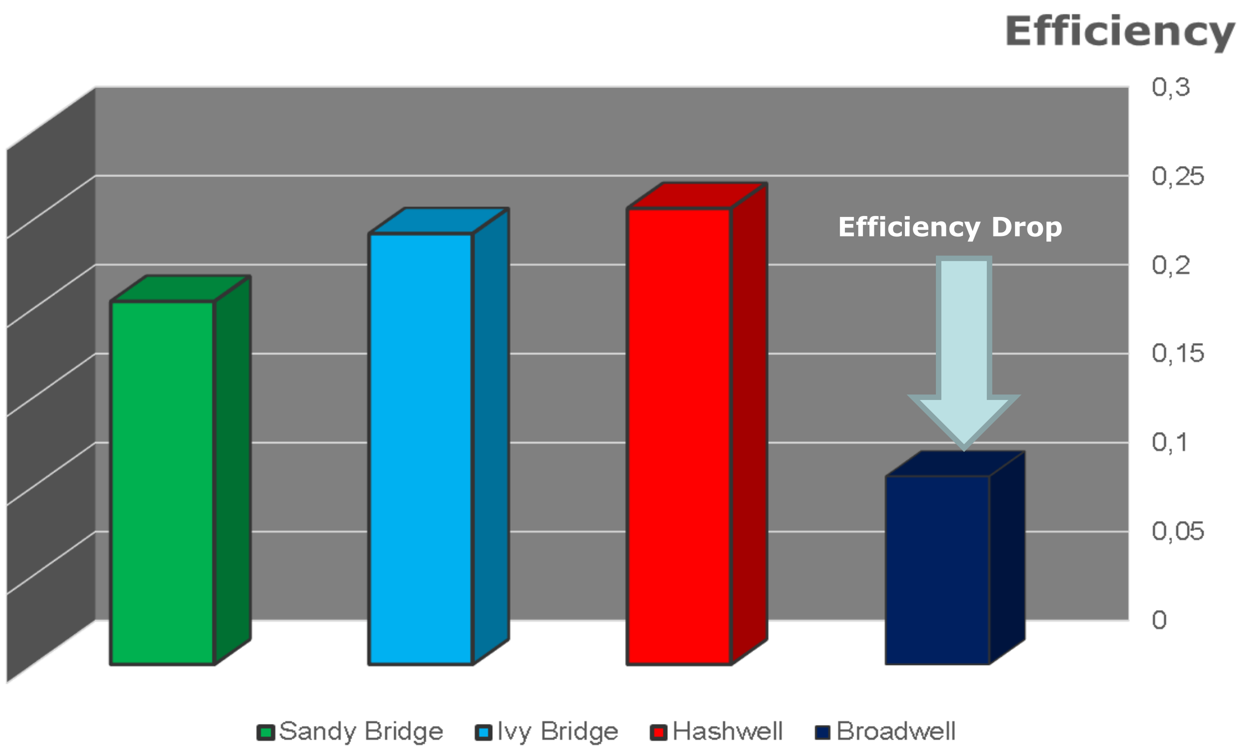
STREAM parallel figure: Marconi



CINECA

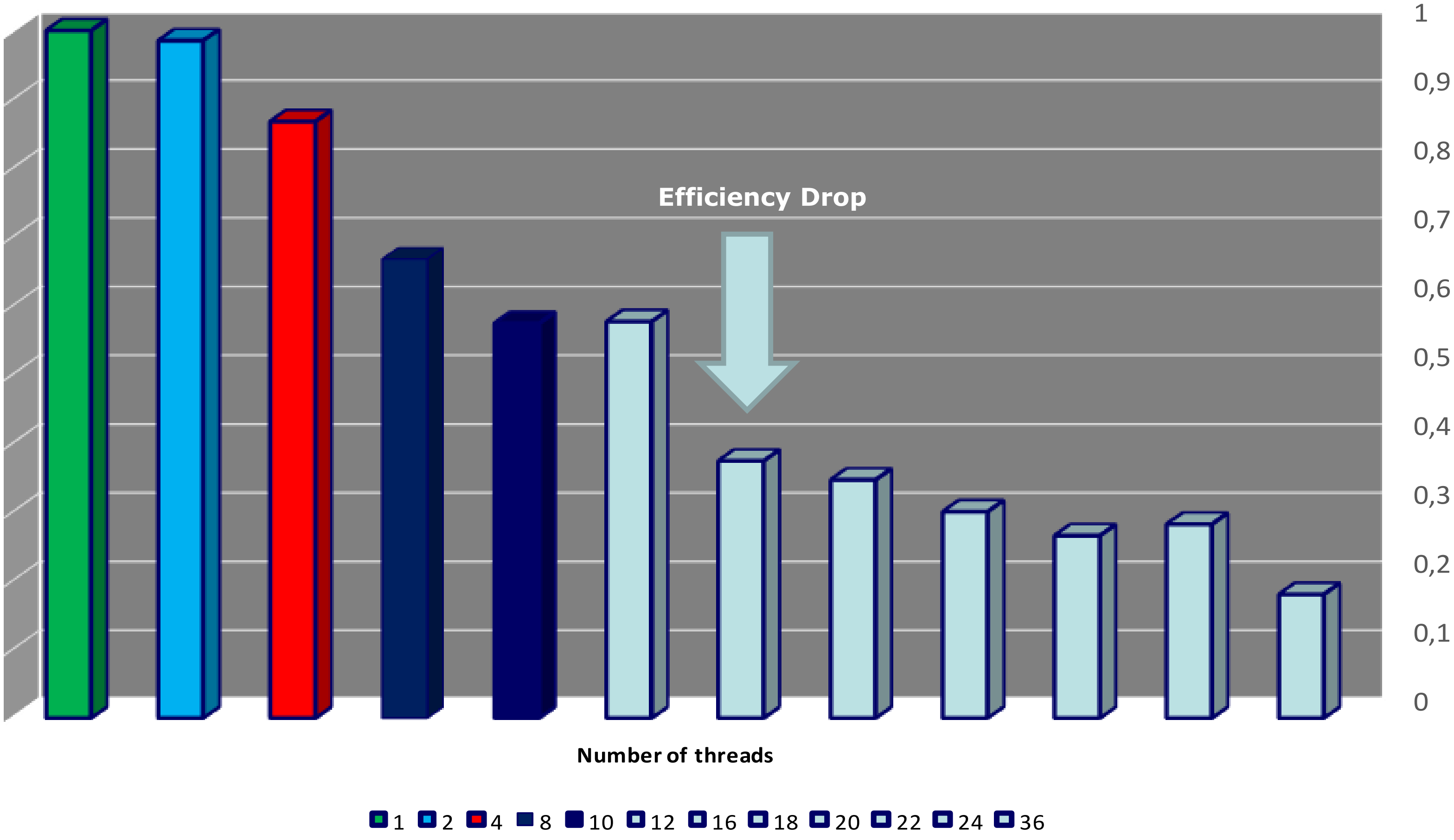
STREAM parallel figure/2

- Best result obtained (intel/gnu compiler)
- Efficiency = Parallel_BW/(#core*Serial_BW)
  - 1 → Linear Speed-up



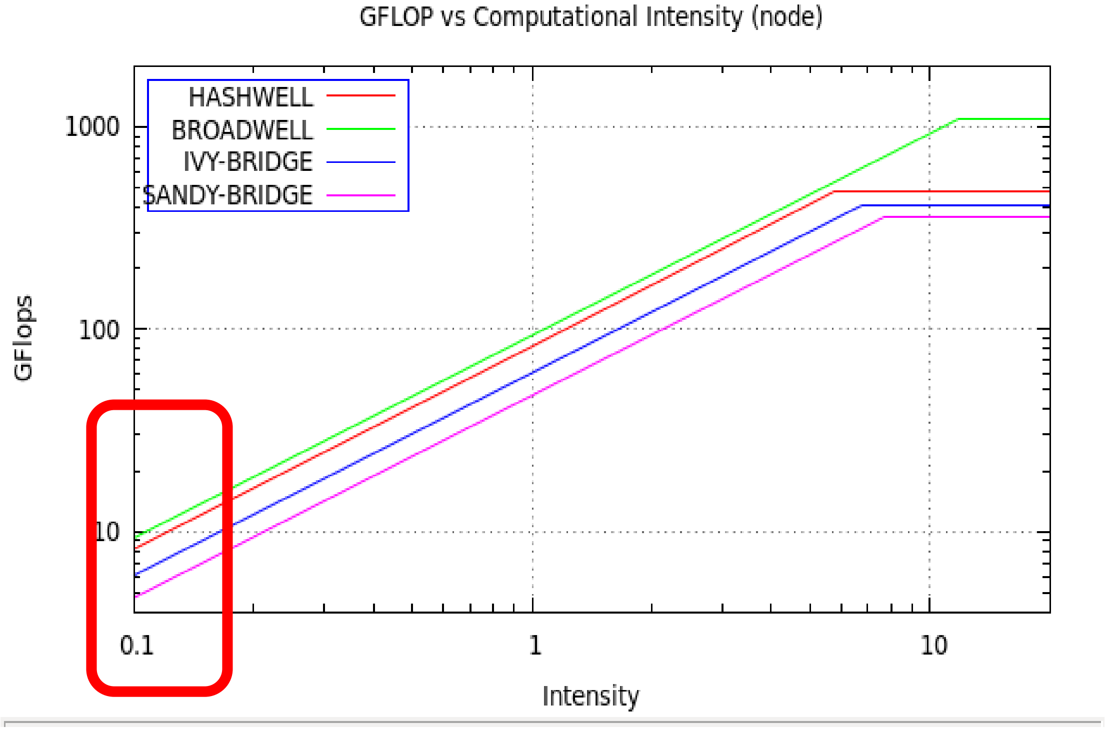
SCAISTREAM parallel figure/2: Marconi

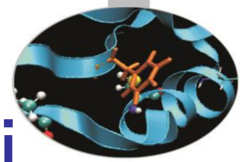
- Best result obtained (intel/gnu compiler)
- Efficiency = Parallel_BW/(#core*Serial_BW)
  - 1 → Linear Speed-up



Roofline: parallel graph

- Using the figures obtained on different HW (LINPACK, STREAM)



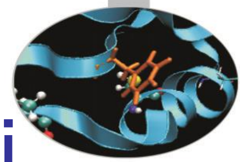


Intel Matrix Benchmarks@Marconi

- Preliminary investigation: try to check network performances (OPA)
- Different Benchmarks (PingPong, send-recv, collectives…) and message sizes

PingPong	MB/s (maximum size)
Same node	11305
Close node	10904
Far node	11246

- **1 or 2 nodes**
- **Same node: processes on the same node**
- **Close node: processes on different nodes but onto the same edge switch**
- **Far node: processes on different nodes and different edge switches (must use the Director OPA switch)**



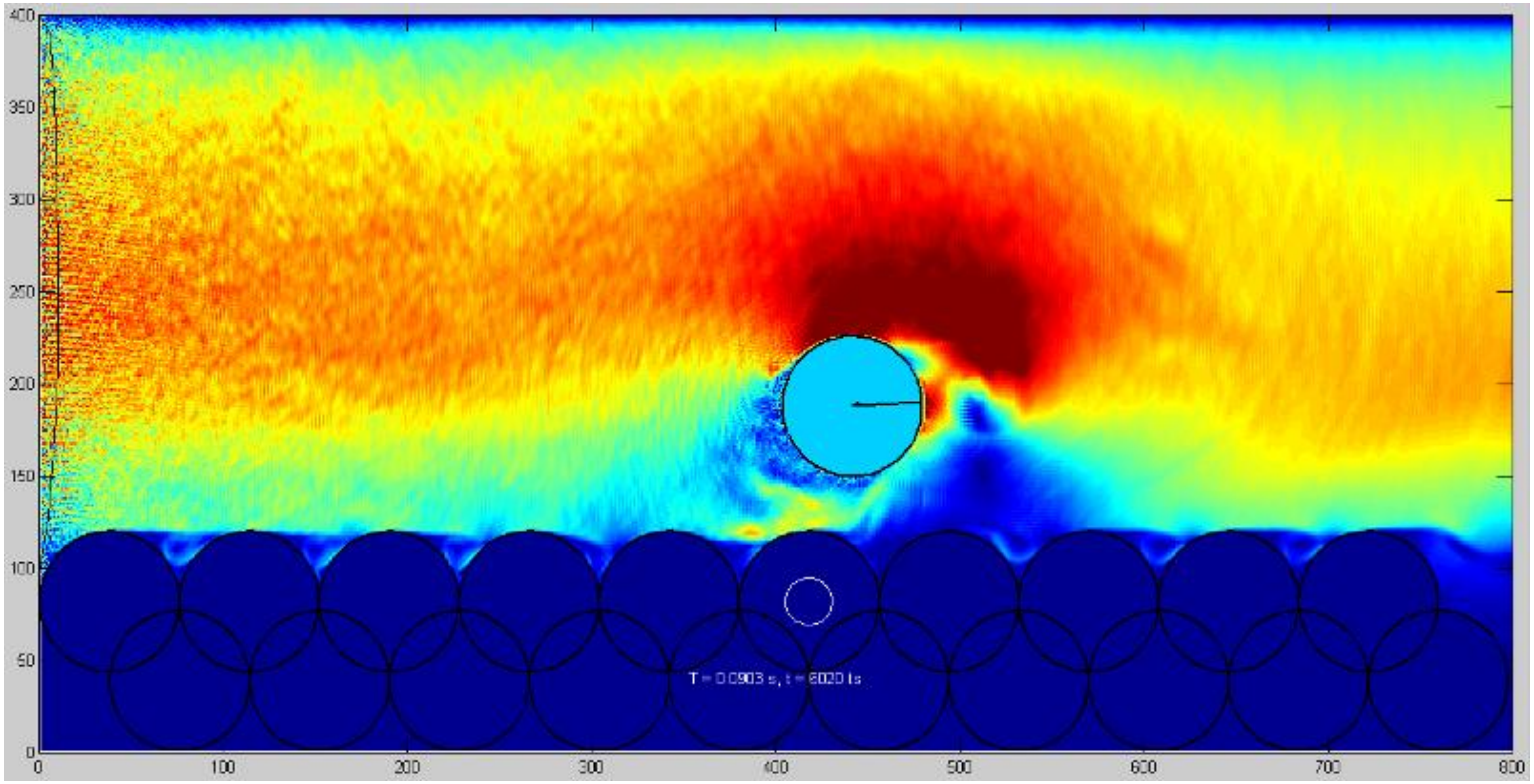
Intel Matrix Benchmarks@Marconi

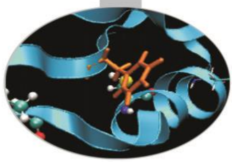
- Preliminary investigation: try to check network performances (OPA)
- Different Benchmarks (PingPong, send-recv, collectives…) and message sizes

AlltoAll	T_average (maximum size, microsec.)
Same node	962
Close node	803
Far node	804

- **1 or 2 nodes**
- **Same node: processes on the same node**
- **Close node: processes on different nodes but onto the same edge switch**
- **Far node: processes on different nodes and different edge switches (must use the Director OPA switch)**

Computational Fluid Dynamics





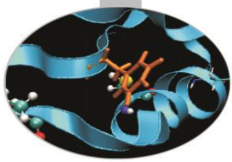
Roofline Mode: LBM

- LBM: hand-made code (3D Multiblock-MPI/OpenMP version)
- Three step serial optimization (an example)

1. Move+Streaming: Computational intensity → 0.36
   - Playing with compilers flag (-O1,-O2,-O3,-fast)
2. Fused: Computational intensity → 0.7
   - Playing with compilers flag (-O1,-O2,-O3,-fast)
3. Fused+single precision: Computational intensity → 1.4
   - Playing with compilers flag (-O1,-O2,-O3,-fast)

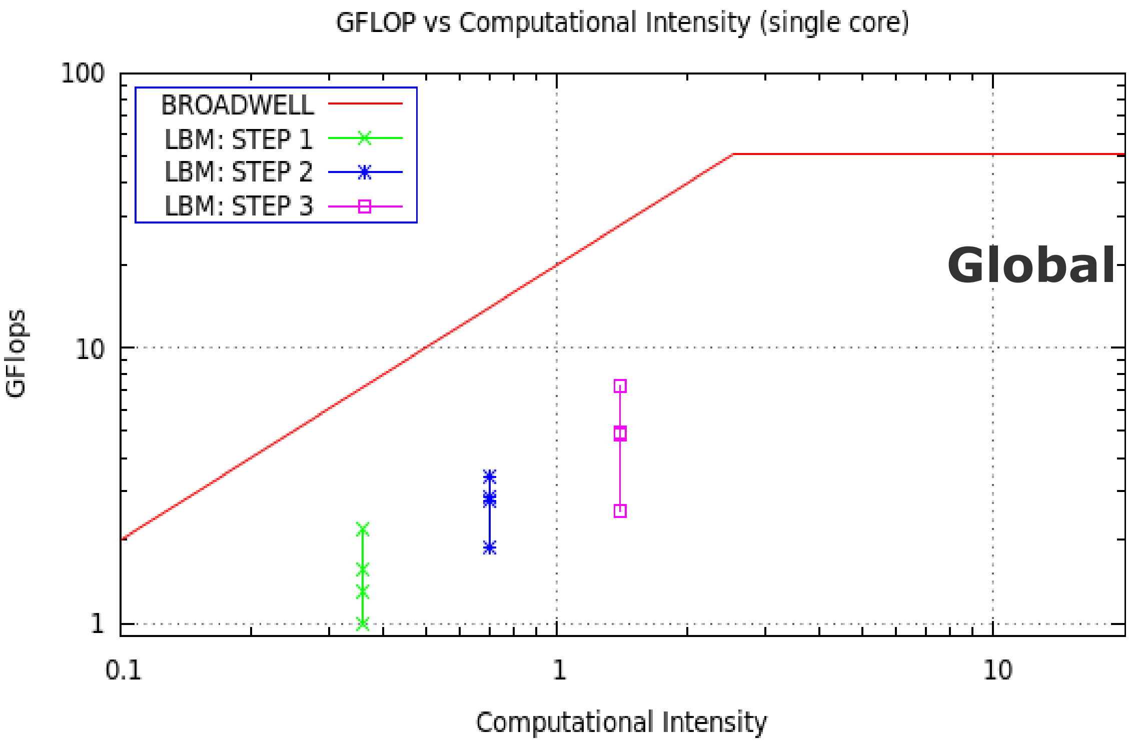
- Test case:
  - 3D driven cavity
  - 128^3

SCAI
SuperComputing Applications and Innovation



Roofline Mode: LBM/2

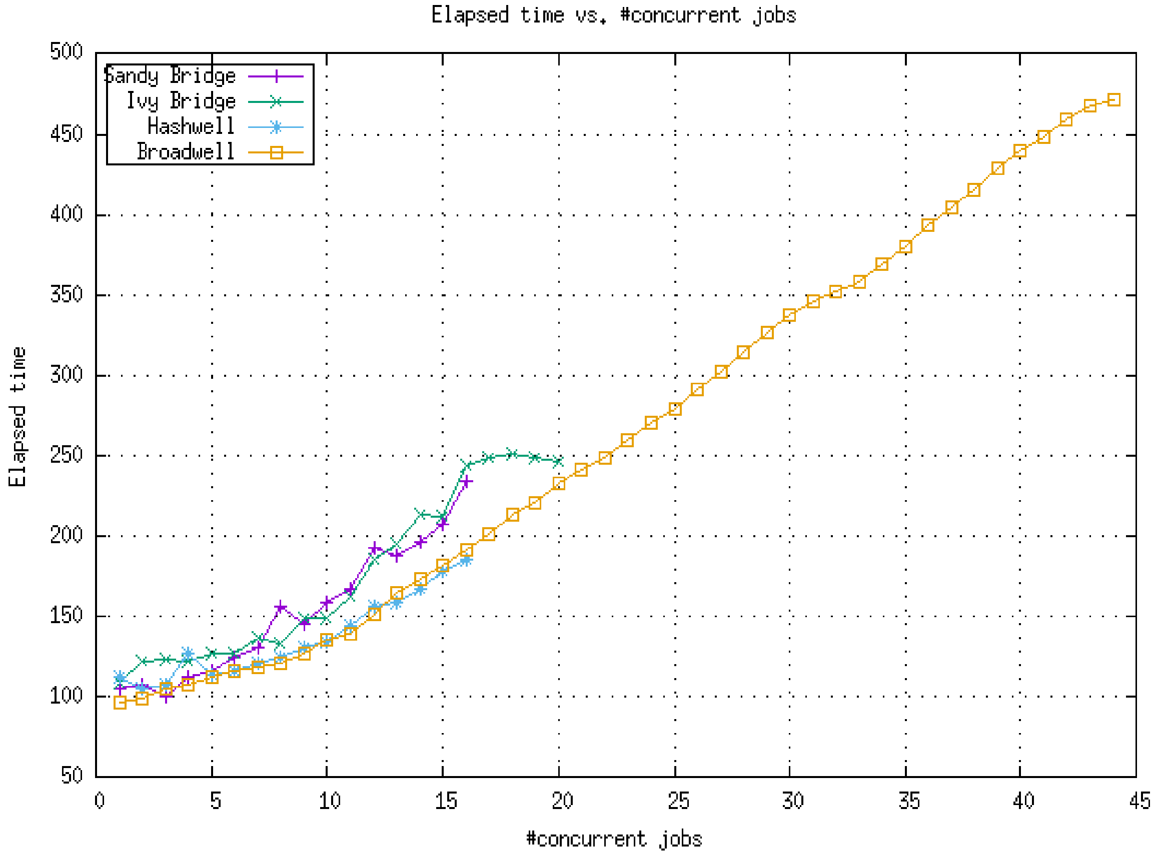
1. Move+Streaming: Computational intensity → 0.36 (2.2x)
2. Fused: Computational intensity → 0.7 (1.8x)
3. Fused+single precision: Computational intensity → 1.4 (2.8x)



Global improvement → 7.3x

Cuncurrent jobs

- LBM code, 3D Driven cavity, Mean value
- From 1 to n equivalent concurrent jobs

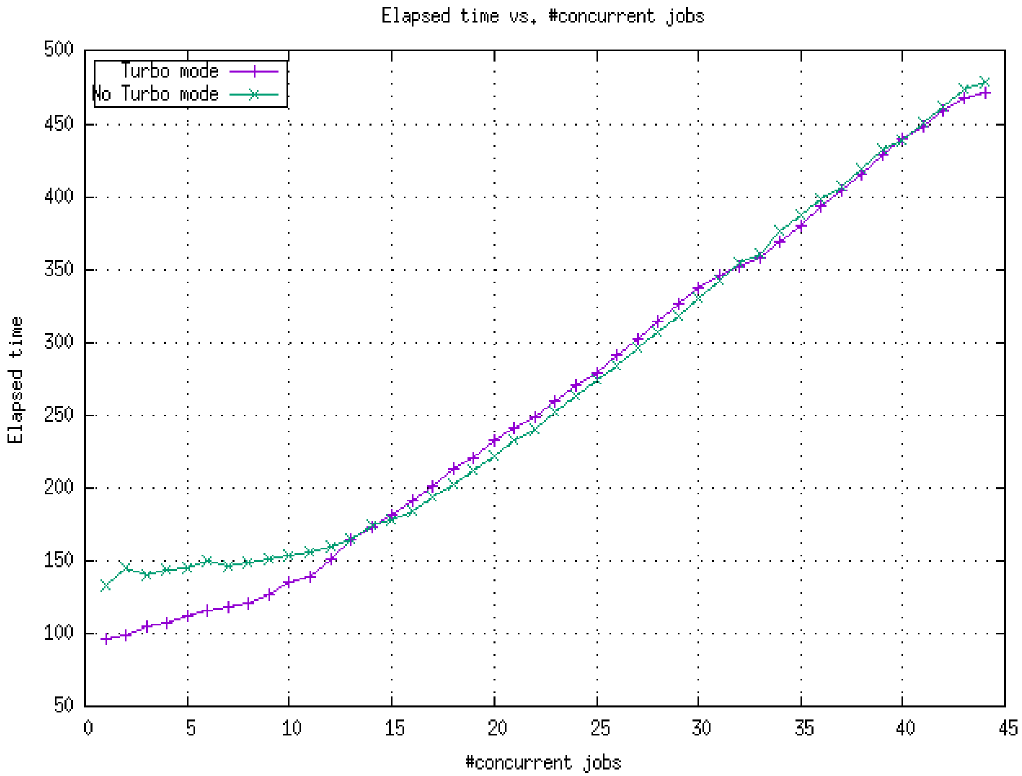


Intel Turbo mode

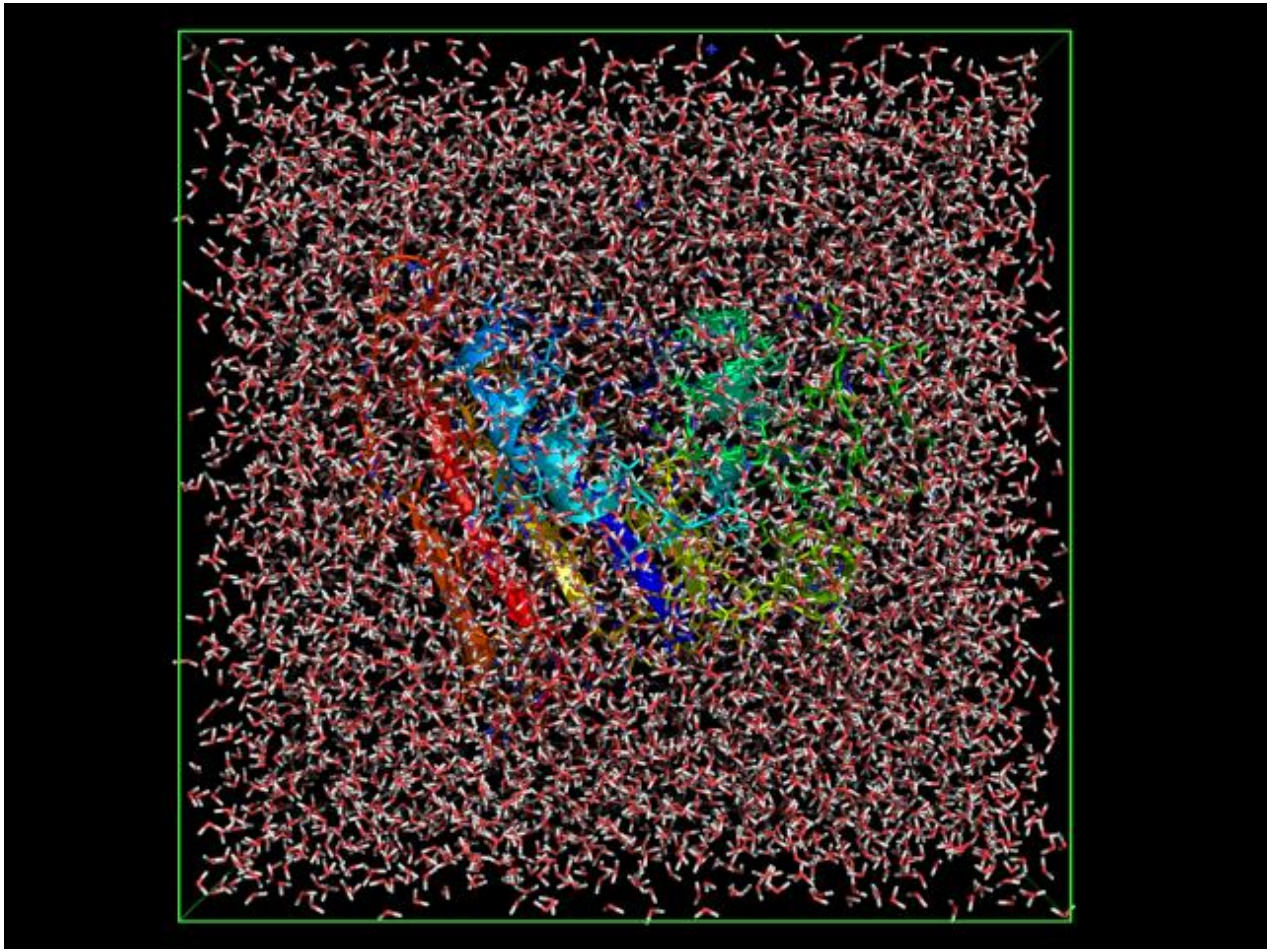
- i.e. Clock increase
- Starting from Hashwell the increase depends from the number of the core involved
- For CINECA Hashwell:
  - ✓ Core 1,2: 3.2 GHz
  - ✓ Core 3: 3.0 GHz
  - ✓ Core 4: 2.9 GHz
  - ✓ Core 5: 2.8 GHz
  - ✓ Core 6: 2.7 GHz
  - ✓ Core 7: 2.6 GHz
  - ✓ Core 8: 2.6 GHz
- Now It's hard to make a "honest" speedup!!!!!

Turbo mode & Concurrent jobs

- LBM code, 3D Driven cavity. Mean value, Broadwell



Molecular Dynamics

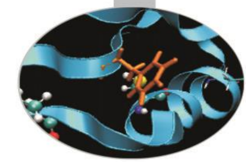


Using MD on Marconi – Phase I

- **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

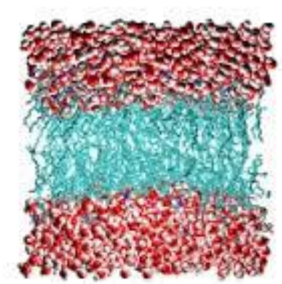
MD Broadwell benchmarks



Gromacs DPPC (1 core)

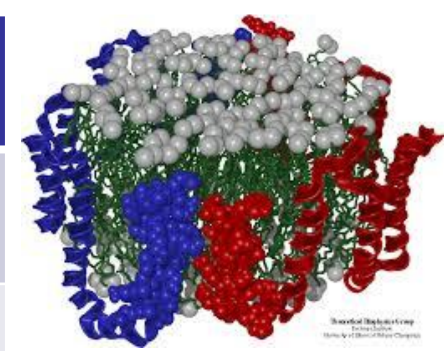
Computer system	ns/day	Speedup wrt Fermi
Haswell (5.0.4, Galileo)	1.364	13.64
Fermi (5.0.4)	0.100	1.00
Broadwell (5.1.2)	1.977	19.77

Based on a 1-node Broadwell partition (40 cores, hyperthreading on).

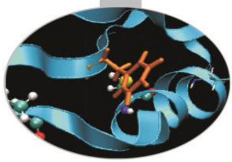


NAMD APOA1 (16 tasks)

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



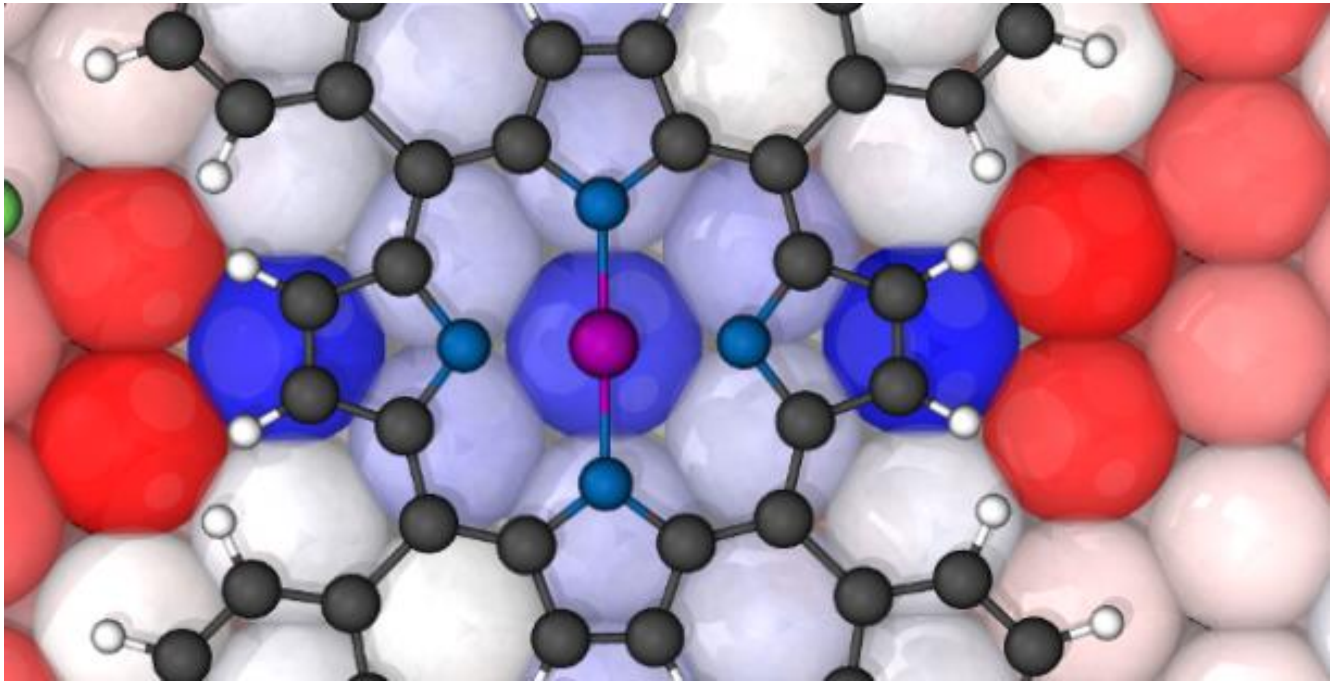
Using MD on Marconi-Phase II



- **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. Some example:
  - 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.

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

Material Science



Preliminary QE benchmarks

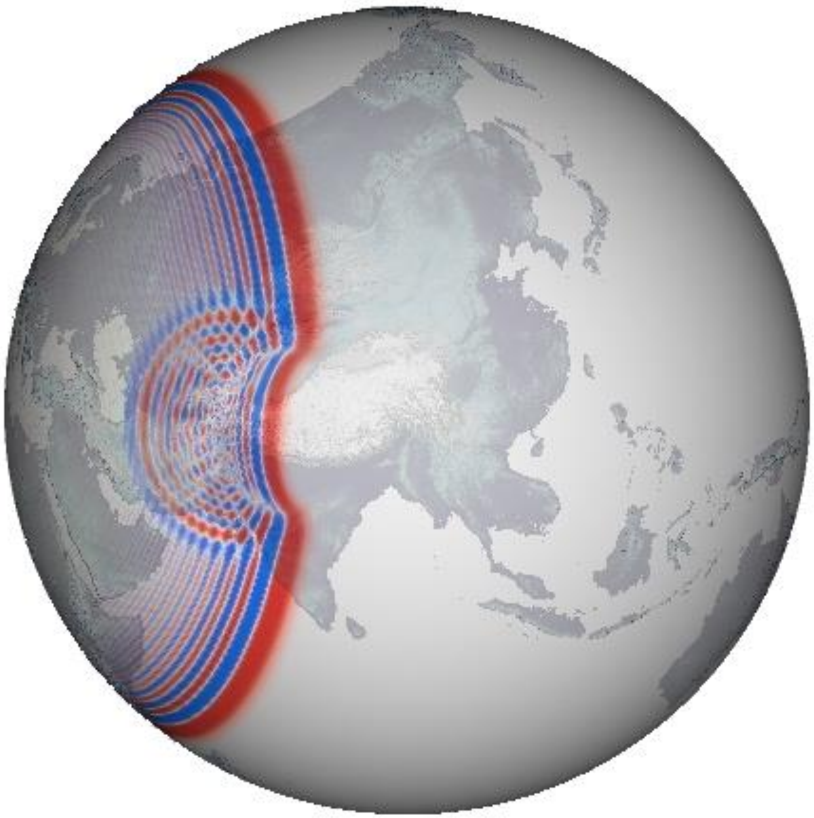
QE benchmark	Galileo	Marconi
W64@64pe	13.50s WALL	10.76s WALL
W256@1024	37.38s WALL	38.83s WALL*
W256@1024	37.38s WALL	28.23s WALL**
W256@1024	37.38s WALL	30.81s WALL
W256@2048	---	22.79s WALL***
W256@512	---	45.05s WALL
W256@256	1m 7.78s WALL	1m11.62s WALL

* Without tuning parallelization parameters

** 32 proc per node

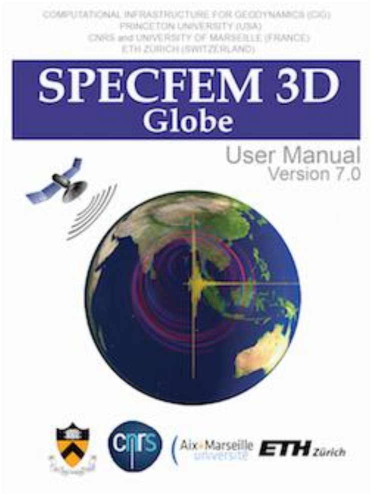
*** 1024-MPI x 2-OpenMP

Global Seismology



Global seismology activity on Marconi-Phase II

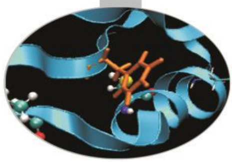
- Global seismology developers must utilise vectorisation (SIMD) and OpenMP threads, and possibly the fast memory of KNL.
- For the user, global seismology experience will depend on how software developers are able to exploit the KNL architecture:
  - SPECFEM3D_GLOBE. **Already reasonable results with KNC ("native" and "offload" version** in the framework of the IPCC@CINECA activity). Good amount of vectorisation (FORCE_VECTORIZATION preprocessing enabling) and SIMD optimization **suitable for KNC and future KNL**. **High number of OpenMP threads scaling (up to more than 60 on**



Worth noting that up to now KNC MICs haven't been widely supported by Global seismology software developers and users. A remarkable exception is SPECFEM3D_GLOBE software CIG repo where the "native" version is maintained and tested. Again, this should be fine for KNL startup.

SuperComputing Applications and Innovation

Global seismology benchmarks



SPECFEM3D_GLOBE Regional_MiddleEast test case: forward simulation

Computer system	e.t. (sec.)	Speedup wrt Haswell
Haswell (Galileo)	570.20	1.00
KNC (Galileo)	430.35	1.32

Based on a 4-node Galileo partition (16 MPI processes, 4 and 60 OpenMP threads on Haswell and KNC respectively).

*SPECFEM3D_GLOBE Regional_MiddleEast test case: **no vectorisation***

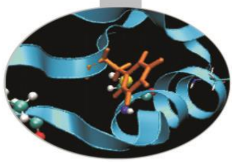
Computer System	e.t. (sec.)	Slowdown factor wrt vectorised
Haswell (Galileo)	687.14	1.20
KNC (Galileo)	848.12	1.97 **<- 2x Slowdown factor**

The impact of vectorisation: on Haswell and KNC respectively).

Conclusions

- Marconi A1 Single core: moderate improvements over the years…. **but a big improvements compared to Fermi.**
- Target is always LINPACK performances.
- Bandwidth grows more slowly than expected.

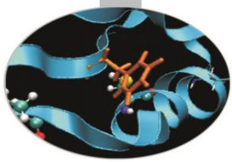
- **High expectations of Marconi A2 KNL performances**.
- KNC paves the way for increasing performances…
- ….try to manage domain parallelism, increase threading, exploit data parallelism (vectorisation) and improve data locality (new chance: use on package memory)



Credits

- Giorgio Amati, Ivan Spisso (Benchmarks, CFD)
- Carlo Cavazzoni (Benchmarks, Material Science)
- Andrew Emerson (Molecular Dynamics)
- Vittorio Ruggiero (Global Seismology)

Some Links



- TICK-TOCK: http://www.intel.com/content/www/us/en/silicon-innovations/intel-tick-tock-model-general.html
- WESTMERE: http://ark.intel.com/it/products/family/28144/Intel-Xeon-Processor-5000-Sequence#@Server
- SANDY BRIDGE: http://ark.intel.com/it/products/family/59138/Intel-Xeon-Processor-E5-Family#@Server
- IVY BRIDGE: http://ark.intel.com/it/products/family/78582/Intel-Xeon-Processor-E5-v2-Family#@Server
- HASHWELL: http://ark.intel.com/it/products/family/78583/Intel-Xeon-Processor-E5-v3-Family#@Server
- BROADWELL: http://ark.intel.com/it/products/family/91287/Intel-Xeon-Processor-E5-v4-Family#@Server
- LINPACK: https://en.wikipedia.org/wiki/LINPACK
- STREAM: https://www.cs.virginia.edu/stream/ref.html
- HPCG: http://hpcg-benchmark.org/
- ROOFLINE: http://crd.lbl.gov/departments/computer-science/PAR/research/roofline/
- TURBO MODE: http://cdn.wccftech.com/wp-content/uploads/2016/03/Intel-Broadwell-EP-Xeon-E5-2600-V4_Non_AVX.png