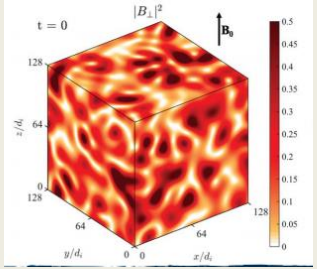
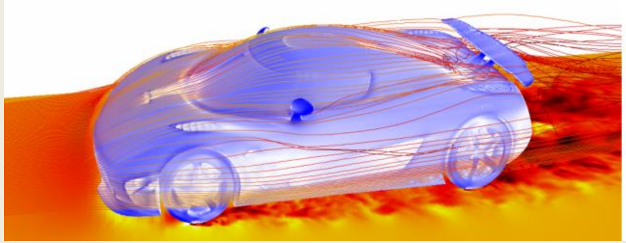
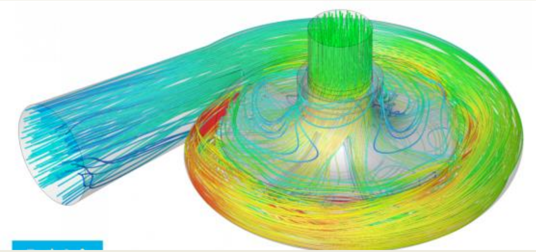
HPC Cineca Infrastructure: State of the art and towards the exascale

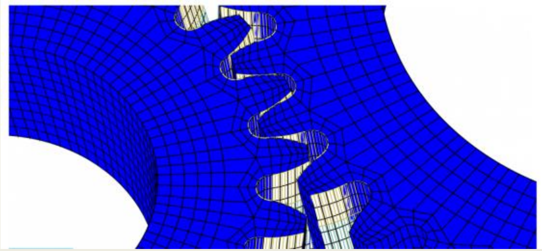


HPC Methods for CFD and Astrophysics
13 Nov. 2017, Casalecchio di Reno, Bologna

Ivan Spisso, i.spisso@cineca.it







Contents

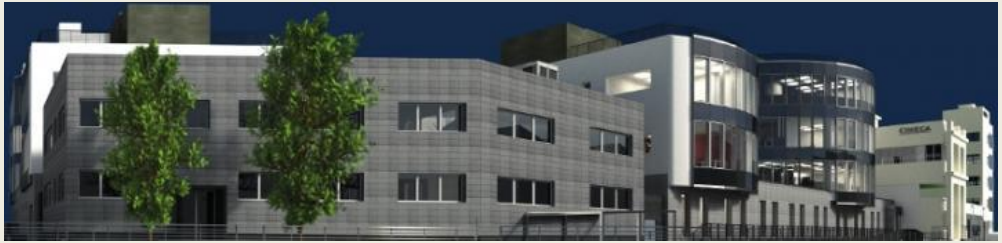
- CINECA in a nutshell and SCAI mission
- HPC ecosystem (up-to-date)
    - Galileo
    - Pico
    - Marconi
    - D.A.V.I.D.E.
- HPC future trends: towards the exascale

Cineca in a nutshell

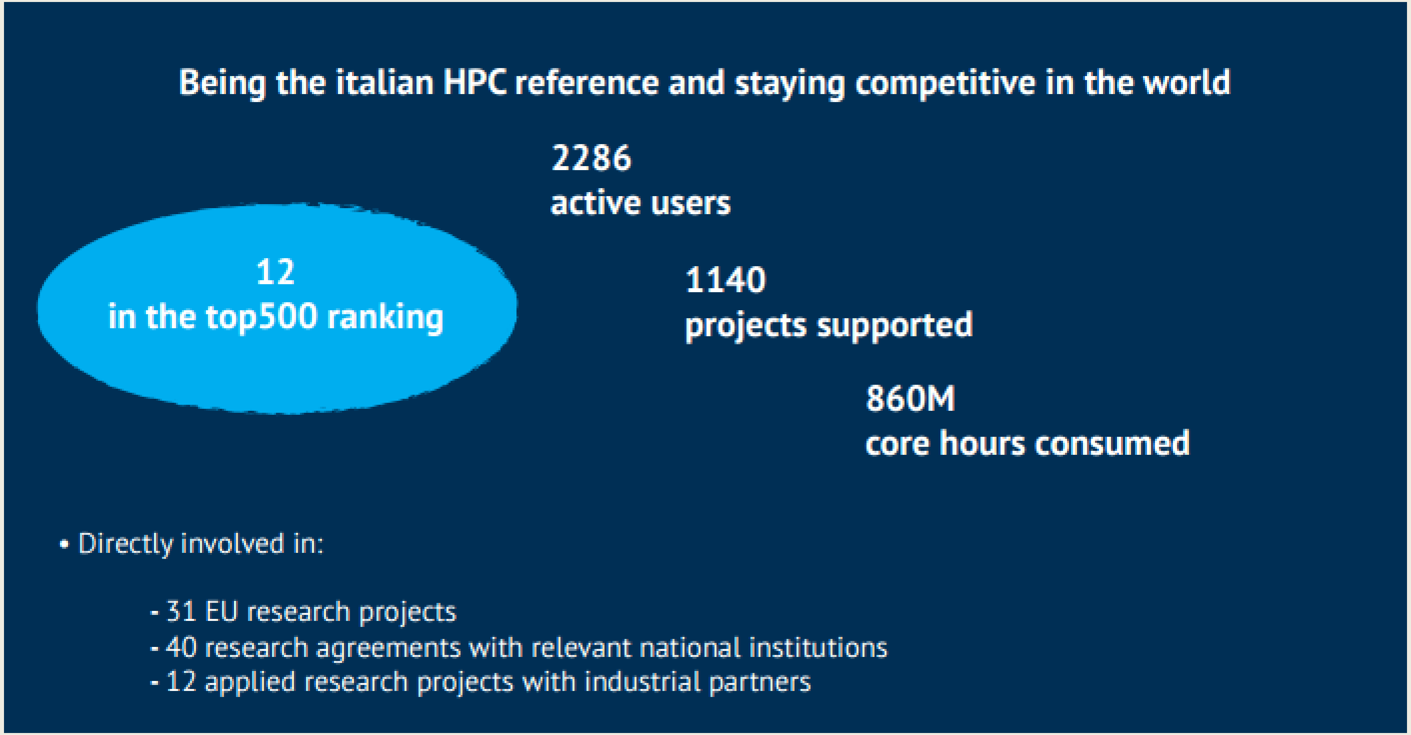
Cineca is a no-profit consortium composed by 70 italian universities, research institutions and the ministry of research.

Cineca provides IT services and it is the largest italian supercomputing facility

Cineca headquarters are in Bologna (selected for the new ECMWF datacenter) and it has offices in Rome and Milan.



SCAI department at Cineca



Being the italian HPC reference and staying competitive in the world

12
in the top500 ranking

2286
active users

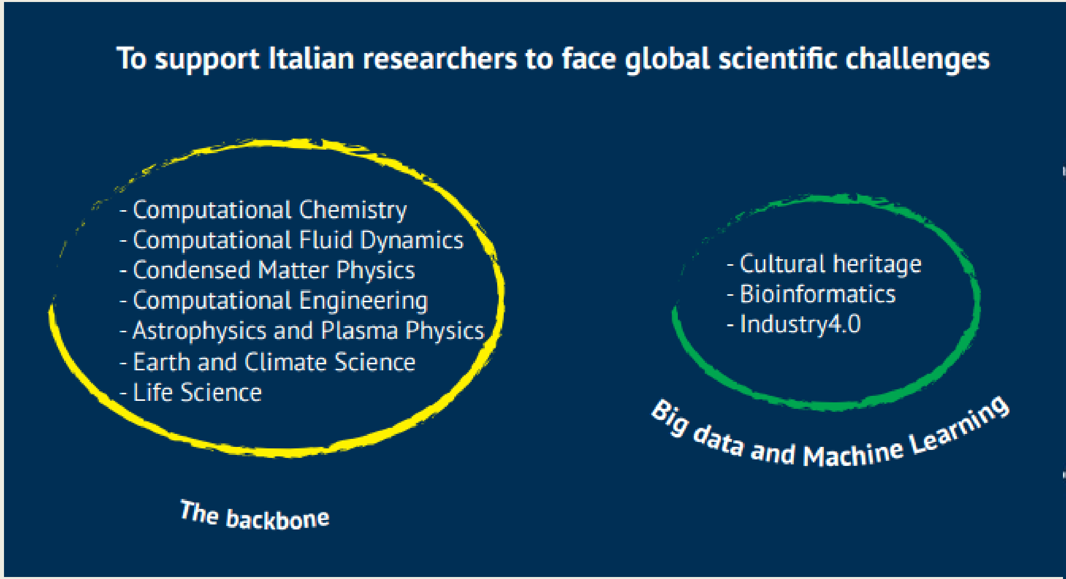
1140
projects supported

860M
core hours consumed

- Directly involved in:

  - 31 EU research projects
  - 40 research agreements with relevant national institutions
  - 12 applied research projects with industrial partners

SCAI mission



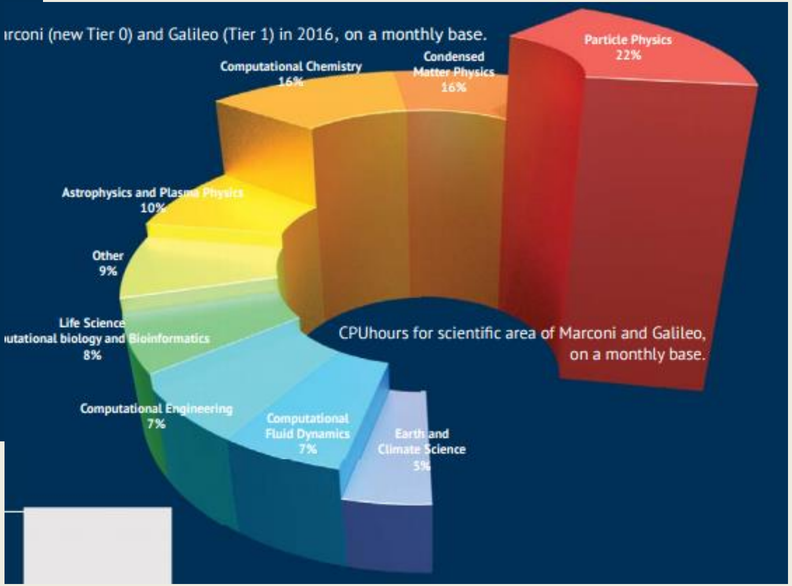
To support Italian researchers to face global scientific challenges

- Computational Chemistry
- Computational Fluid Dynamics
- Condensed Matter Physics
- Computational Engineering
- Astrophysics and Plasma Physics
- Earth and Climate Science
- Life Science

The backbone

- Cultural heritage
- Bioinformatics
- Industry4.0

Big data and Machine Learning



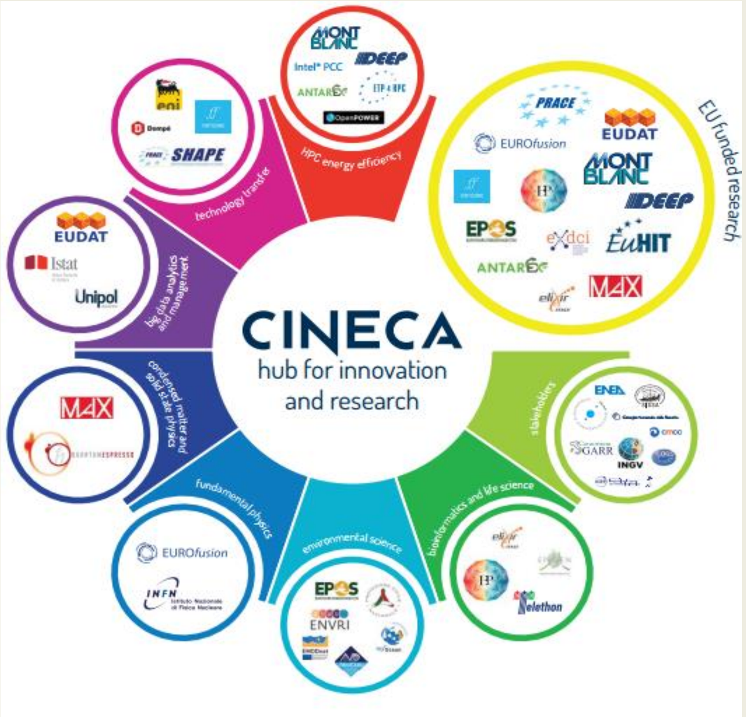
ırconi (new Tier 0) and Galileo (Tier 1) in 2016, on a monthly base.

CPUhours for scientific area of Marconi and Galileo, on a monthly base.

The Cineca ecosystem

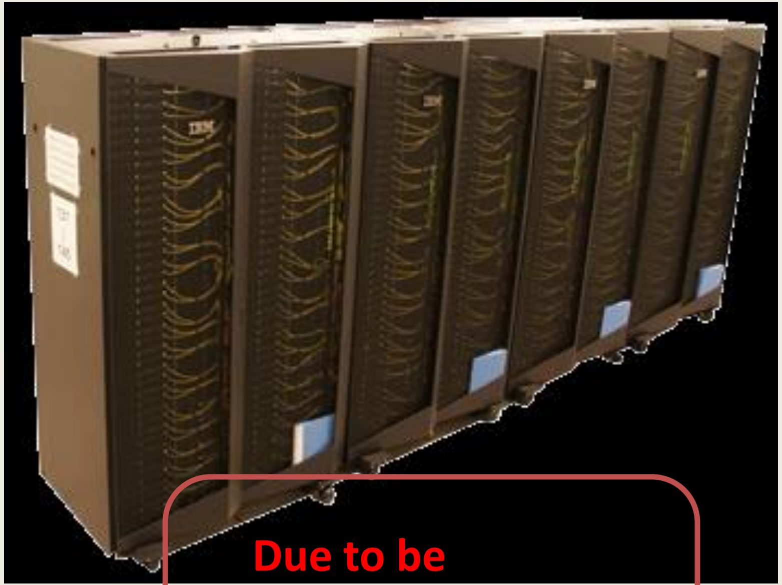
Cineca acts as a hub for innovation and research contributing to many scientifical and R&D projects on italian and european basis.

In particular, Cineca is a PRACE hosting member and a member of EUDAT.



HPC INFRASTRUCTURE: GALILEO

- IBM Cluster Linux
- 516 compute nodes
- 2 eight-core Intel Xeon 2630 (16 cores) @2.40 GHz a.k.a. Haswell
- 128GB RAM per node
- Infiniband with 4x QDR switch (40 Gb/s)
- TPP: 1 PFlop/s
- National and PRACE Tier-1 calls, FORTISSIMO, industrial customers

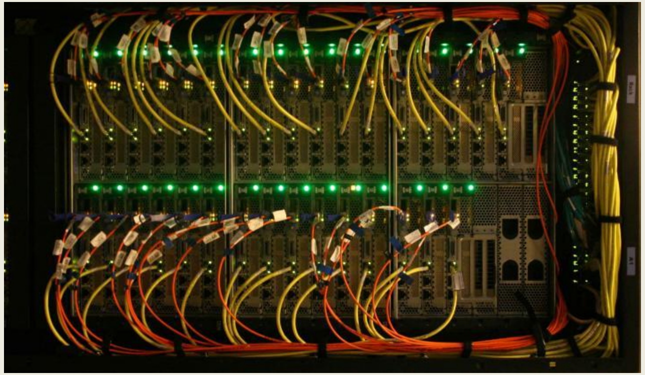


Due to be decommissioned End of November 2017

HPC INFRASTRUCTURE: PICO

- IBM Cluster Linux
- 74 nodes of different types
  - Compute nodes: 51 x (2 x Intel Xeon 10 Core E5-2670v2 2.50 GHz, 128 GB mem)
  - Visualization nodes: 2 x (20 core, 128 GB mem, 2 GPU Nvidia K40)
  - BigInsights nodes: 4 x (16 core, 64 GB mem, 32TB local disks)
  - BigMemory nodes:
    - 1 x (32 core, 520 GB mem)
    - 2 x (20 core, 510 GB mem, 1 GPU Nvidia K6000s)

- Infiniband high-performance network
- devoted to data analytics and large data visualization



HPC INFRASTRUCTURE: MARCONI

- Marconi is the new Tier-0 LENOVO system that replaced the FERMI BG/Q.
- Marconi is planned in two technological stages in a 5 years programme with the objective to reach a 50 Pflop/s system by the year 2019-2020.
- Marconi is a Lenovo NextScale system equipped with Intel Xeon, Intel Xeon Phi processors and Intel SkyLake with an Intel OmniPath network.
- The first stage of MARCONI is made of 3 different partitions (A1, A2 and A3) whose installation started in 2016.
- Marconi is part of the infrastructure provided by Cineca to the EUROFUSION project
- [UserGuide]()

MARCONI A1 : Intel Broadwell

- Started in april 2016 and opened to the production in july 2016
- 1512 compute nodes
- 2 sockets Intel(R) Xeon(R) CPU E5-2697 v4 @2.30 GHz, 18 cores
- 128GB RAM per node
- S.O. Linux Centos 7.2
- PBSpro 13 batch scheduler
- TPP: 2 PFlop/s

MARCONI A2: Intel KNL

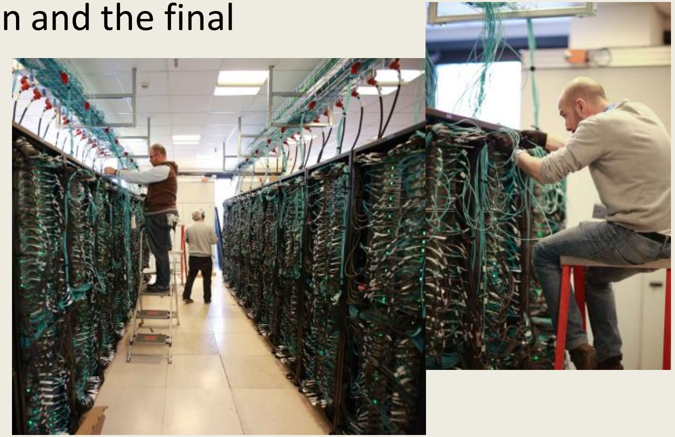
- Opened to production at the end of 2016
- 3600 Knights Landing compute nodes
- Intel Xeon Phi 7250 (68 cores) @1.40 GHz a.k.a. KNL
- 120GB RAM per node
- Default configuration: Cache/Quadrant
- TPP: 11 PFlop/s

MARCONI A3: Intel Skylake

- Full installation end of November
- Racks: 21
- Nodes: 1512 + 792
- Processors: 2 x 24-cores Intel Xeon 8160 CPU (Skylake) at 2.10 GHz
- Cores: 48 cores/node
- 72.576 + 38.016 cores in total
- RAM: 192 GB/node of DDR4
- TPP: 11 PFlop/s

MARCONI’s outlook

- In 2017 MARCONI will evolve with the installation of the A3 partition and the final configuration will have:



- 3024 Intel Skylake nodes (approx. 120960 cores)
- 3600 Intel Knights Landing (approx. 244800 cores)
- Peak performance: about 20 PFlop/s
- Internal network: Intel OPA

- In 2019 we expect the convergence of the HPDA infrastructure and the HPC infrastructure towards the target of 50 PFlop/s

HPC INFRASTRUCTURE: D.A.V.I.D.E.

- Development of an Added-Value Infrastucture Designed in Europe
- PCP (Pre-Commercial Procurement) by PRACE
- OpenPOWER-based HPC cluster
- Power8 processors with NVLink bus + Nvidia Tesla P100 SXM2
- Designed, integrated and tested by E4. Installation in CINECA’s data center
- Available for research projects starting from Septmber

PRACE

E4 COMPUTER ENGINEERING

OpenPOWER™

HPC future trends: towards the exascale

HPC & CPU
Intel evolution: 2010-2016

Westmere (a.k.a. plx.cineca.it)
- Intel(R) Xeon(R) CPU E5645 @2.40GHz, 6 Core per CPU

Sandy Bridge (a.k.a. eurora.cineca.it)
- Intel(R) Xeon(R) CPU E5-2687W 0 @3.10GHz, 8 core per CPU

Ivy Bridge (a.k.a pico.cineca.it)
- Intel(R) Xeon(R) CPU E5-2670 v2 @2.50GHz, 10 core per CPU
- Infiniband FDR

Hashwell (a.k.a. galileo.cineca.it)
- Intel(R) Xeon(R) CPU E5-2630 v3 @2.40GHz, 8 core per CPU
- Infiniband QDR/True Scale (x 2)

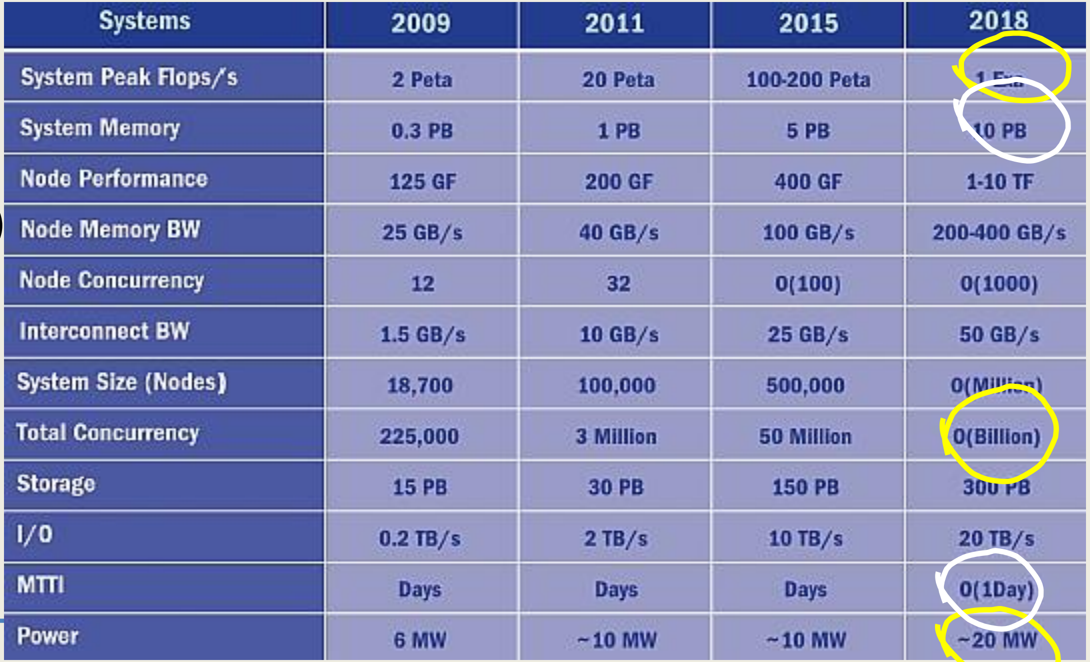
Broadwell (a.k.a marconi.cineca.it)
- Intel(R) Xeon(R) CPU E5-2697 v4 @ 2.30GHz, 18 core per CPU (x2)
- OmniPath

Increasing # of cores,
Same clock

Roadmap to Exascale

(architectural trends)

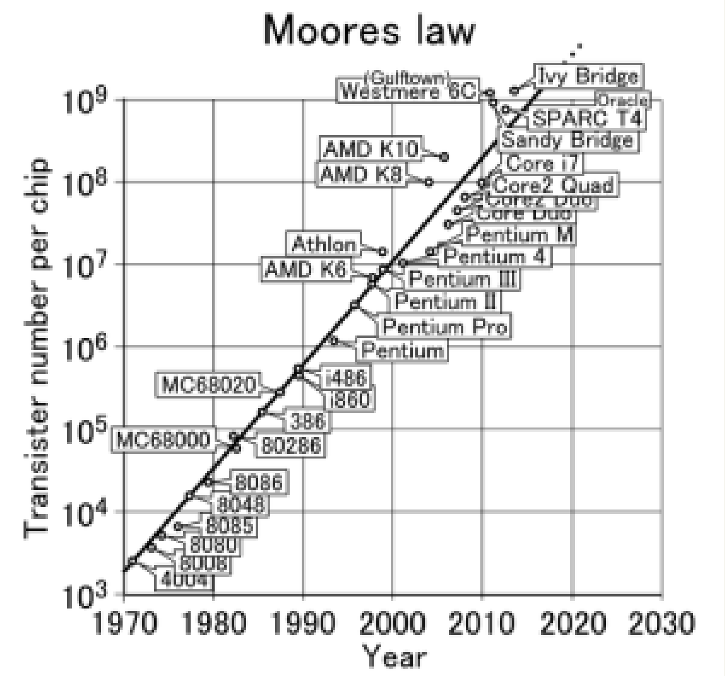
- exascale: computing system capable of al least one exaFLOPs calculation per second.
- exaFLOPs = 10^18 FLOPS or a billion of billion calculations per seconds
- As clock speeds may for reasons of power efficiency be as low as 1 Ghz
- to Performe 1 Eflop/s peak performance
- Need to execute 1 billion floating-point operations concurrently (Total Concurrency)
- MTTI = Mean Time to interrupt, order of day(s)

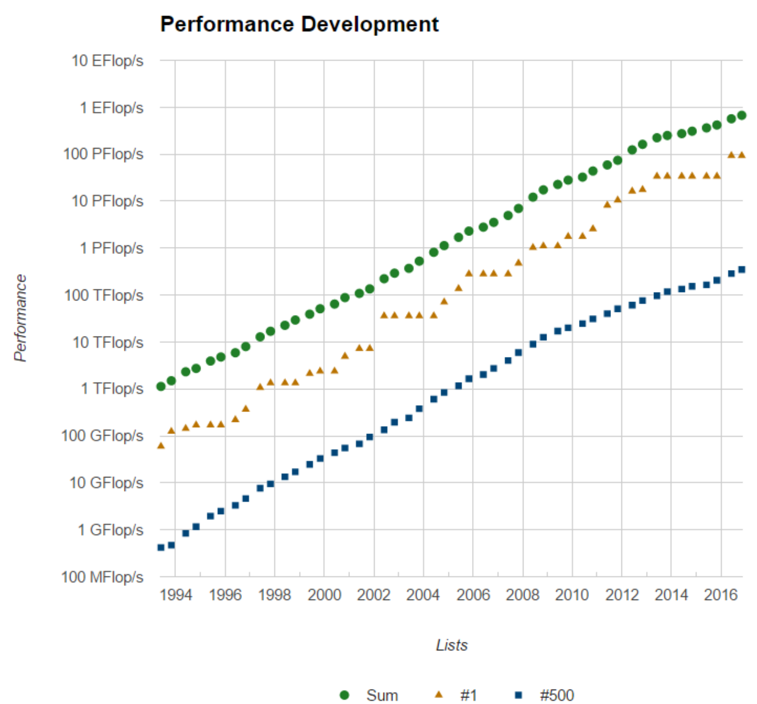


Systems	2009	2011	2015	2018
System Peak Flops/s	2 Peta	20 Peta	100-200 Peta	1 Exa
System Memory	0.3 PB	1 PB	5 PB	10 PB
Node Performance	125 GF	200 GF	400 GF	1-10 TF
Node Memory BW	25 GB/s	40 GB/s	100 GB/s	200-400 GB/s
Node Concurrency	12	32	O(100)	O(1000)
Interconnect BW	1.5 GB/s	10 GB/s	25 GB/s	50 GB/s
System Size (Nodes)	18,700	100,000	500,000	O(Million)
Total Concurrency	225,000	3 Million	50 Million	O(Billion)
Storage	15 PB	30 PB	150 PB	300 PB
I/O	0.2 TB/s	2 TB/s	10 TB/s	20 TB/s
MTTI	Days	Days	Days	O(1Day)
Power	6 MW	~10 MW	~10 MW	~20 MW

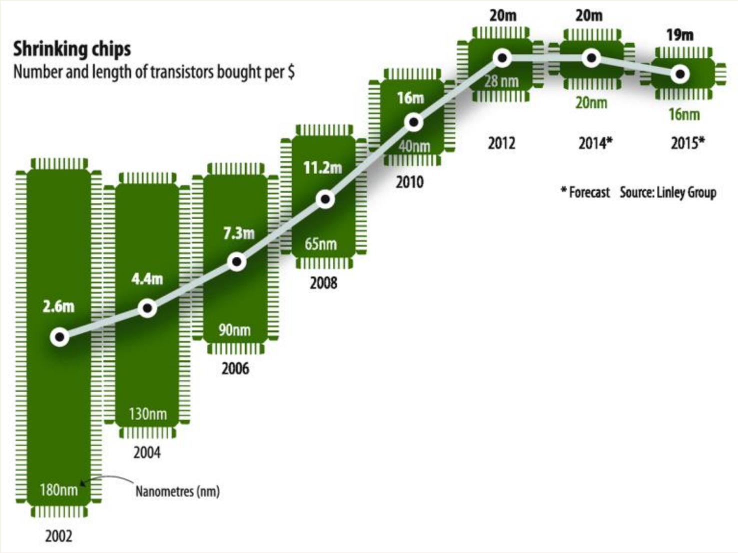
Moore’s Law - Chips

Moore's law is the observation that the number of transistors in a dense integrated circuit doubles approximately every two years (18 months, Intel executive David House)





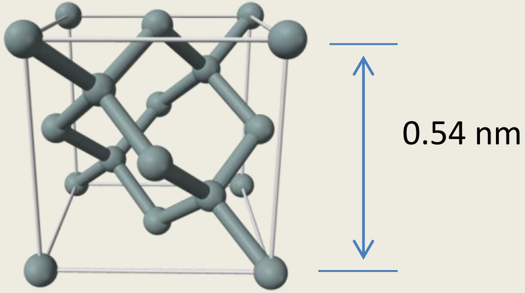
Moore’s Law - Dollars



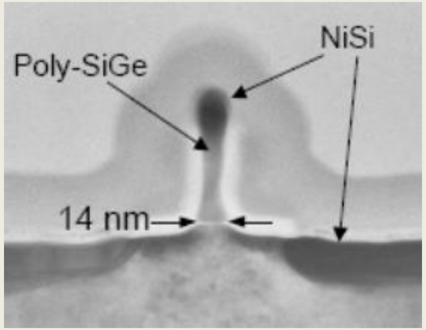
* Forecast Source: Linley Group

Oh-oh! Houston! we have a problem….

The silicon lattice



Si lattice



50 atoms!

There will be still 4~6 cycles (or technology generations) left until
we reach 11 ~ 5.5 nm technologies, at which we will reach downscaling limit, in some year
between 2020-30 (H. Iwai, IWJT2008).

Dennard scaling law (downscaling)

also known as **MOSFET scaling** states that as transistors get smaller their power density (P) stays constant, so that the power (D) use stays in proportion with area: both voltage (V) and current scale (downward) with length.

old VLSI gen.

The key effect of Dennard scaling was that as transistors got smaller the power density was constant – so if there was a reduction in a transistor's linear size by 2, the power it used fell by 4 (with voltage and current both halving. "the Golden Age of transistor"

$L' = L / 2$

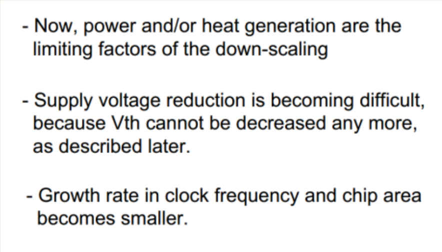
$V' = V / 2$

$F' = F * 2$

$D' = 1 / L^2 = 4D$

$P' = P$

do not hold anymore!



- Now, power and/or heat generation are the limiting factors of the down-scaling
- Supply voltage reduction is becoming difficult, because Vth cannot be decreased any more, as described later.
- Growth rate in clock frequency and chip area becomes smaller.

new VLSI gen.

$L' = L / 2$

$V' = \sim V$

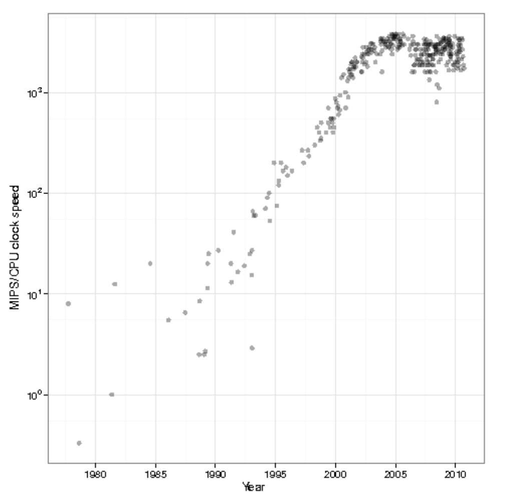
$F' = \sim F * 2$

$D' = 1 / L^2 = 4 * D$

$P' = 4 * P$

The power crisis!

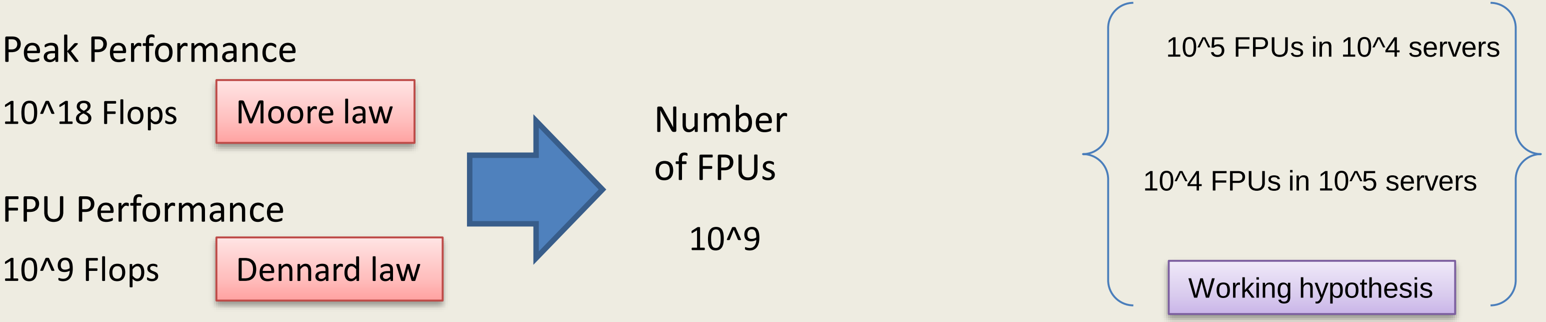
The core frequency and performance do not grow following the Moore's law any longer



Increase the number of cores to maintain the architectures evolution on the Moore's law

Programming crisis!

Exascale How serious the situation is?



Peak Performance

10^18 Flops — Moore law

FPU Performance

10^9 Flops — Dennard law

Number of FPUs

10^9

10^5 FPUs in 10^4 servers

10^4 FPUs in 10^5 servers

Working hypothesis

- Exascale is not (only) about scalability and Flops performance!
- In an exascale machine there will be 10^9 FPUs, bring data in and out will be the main challenge.
- 10^4 nodes, but 10^5 FPUs inside the node!
- heterogeneity is here to stay
- deeper memory hierarchies

POWER is the limit!

At 7nm Power will be the main limit for chip designers, not number of transistors

-> I cannot power all transistors all together -> dark silicon, how to use it?
-> Memory? I/O interface? Different cores? Core & GPU?

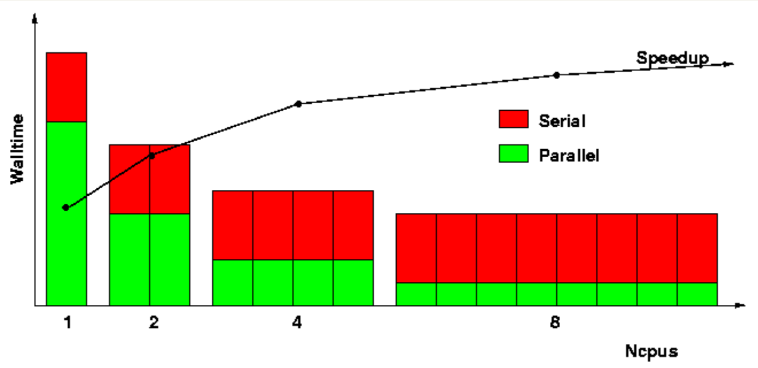
Very Big co-design Problem!

Amdahl's law

Amdahl's law is a formula which gives the theoretical speedup in latency of the execution of a task at fixed workload that can be expected of a system whose resources are improved

In a massively parallel context, an upper limit for the scalability of parallel applications is determined by the fraction of the overall execution time spent in non-scalable operations (Amdahl's law).

For example, if a program needs 20 hours using a single processor core, and a particular part of the program which takes one hour to execute cannot be parallelized, while the remaining 19 hours (p = 0.95) of execution time can be parallelized, then regardless of how many processors are devoted to a parallelized execution of this program, the minimum execution time cannot be less than that critical one hour. Hence, the theoretical speedup is limited to at most 20 times (1/(1 − p) = 20). For this reason parallel computing with many processors is useful only for very parallelizable programs.



maximum speedup tends to
$1 / (1 - P)$
P= parallel fraction

1,000,000 core

$P = 0.999999$

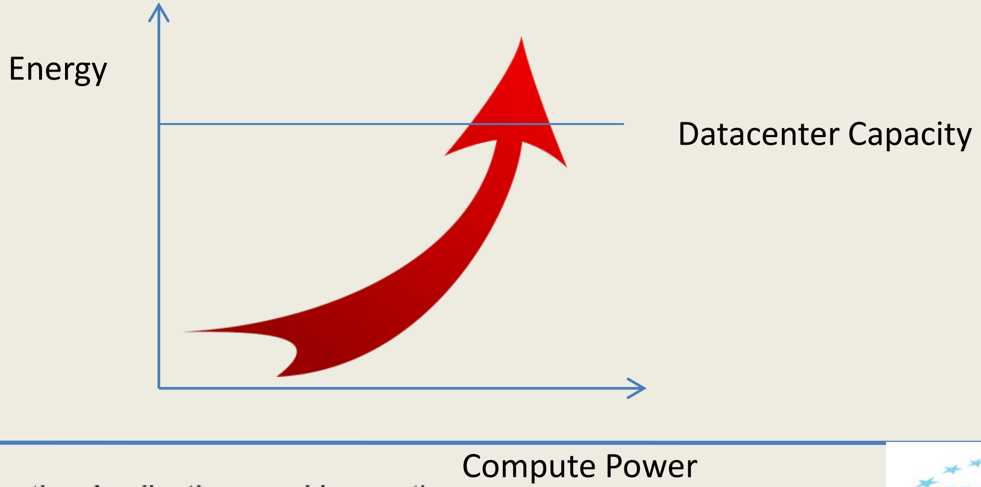
serial fraction= 0.000001

Oh-oh! Houston! we have an another problem….

Energy trends

- “traditional” RISC and CISC chips are designed for maximum performance for all possible workloads
- RISC = Reduced Instruction Set Computer
- CISC = Complex Instruction Set Computer

A lot of silicon to maximize single thread performace

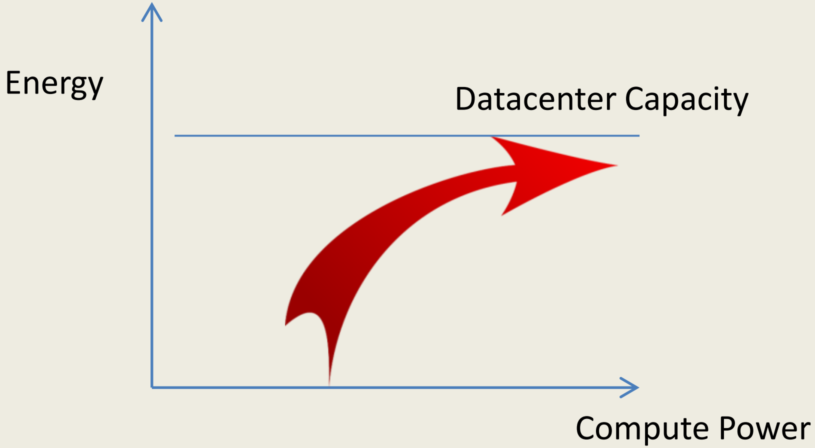


Change of Paradigm: Energy Efficiency

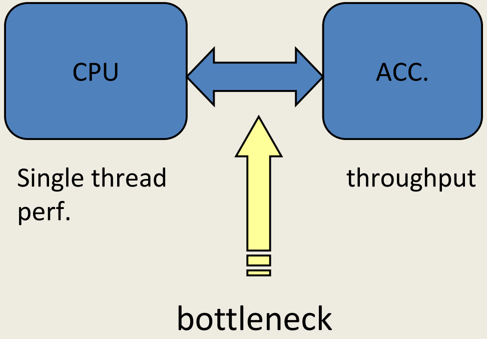
New chips designed for maximum performance in a small set of workloads

Simple functional units, poor single thread performance, but maximum throughput

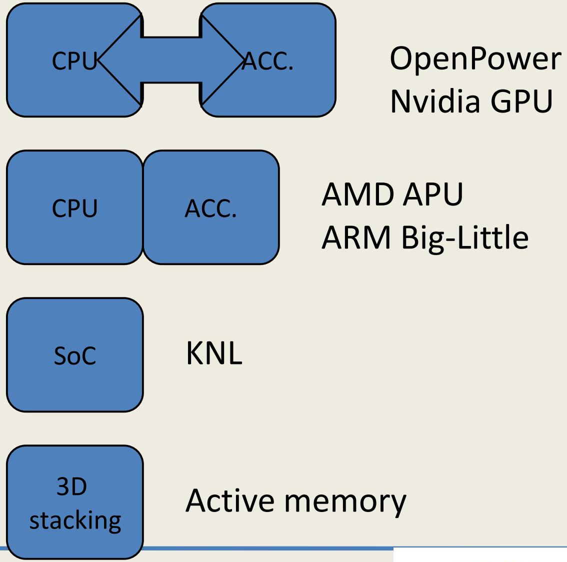
- HPC centres are vast and greedy consumers of electricity, requiring MW of energy (for example, Cineca is the largest consumer of power in the Emilia-Romagna region)
- Energy efficiency is clearly an important topic and there is much interest in renewable energy sources, re-using waste heat for builing, use of hot water cooling (see old [Eurora cluster](), top rank in the Green500 in June 2013)
- Many EU projects, in the quest for Exascale performances, are studying strategies for reducing energy



Architecture toward exascale



GPU/MIC/FPGA

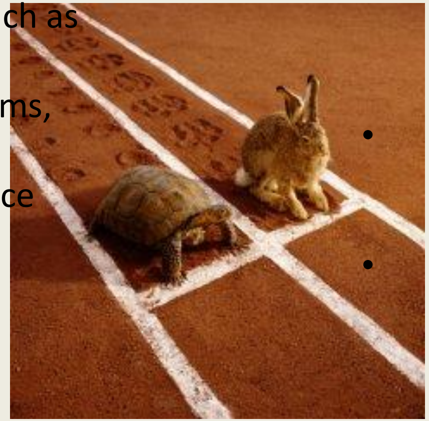


Photonic -> platform flexibility
TSV -> stacking

Towards the exascale: Summary and trends

Software (turtle)

- As usual software lags behind hardware but must learn to exploit accelerators and other innovative technologies such as FGPAs, PGAS
- Reluctance by some software devs to learn new languages such as CUDA, OpenCL is driving interest in compiler-directive languages such as OpenAcc and OpenMP (4.x)
- Continued investment in efficient filesystems, checkpointing, resilience, parallel I/O
- **co-design** is the way the reduce the distance between hardware and software for HPC

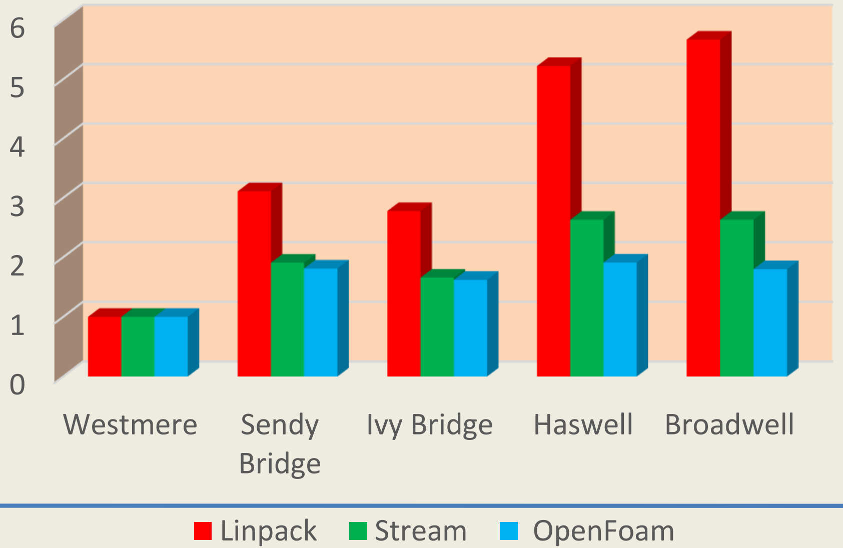


Hardware (hare)

- Reaching physical limits of transistor densities and increasing clock frequencies further is too expensive and difficult (energy consumption, heat dissipation)
- Parallelism only solution in HPC but the Blue Gene road is no longer being persued. Hybrid with accelerators such as GPUs or Xeon Phi become the norm
- Accelerator technologies advancing to remove limits associated with, (Intel KNL or Nvidia NVLINK)
- A range of novel architectures being explored (e.g. Mont Blanc, DEEP) and technologies in many areas

HPC status and future trends. Which impact for OpenFoam?

- ✓ About 6 year CPU evolution
  - ✓ Linpack (Floating point Benchmark)
  - ✓ Stream (Memory BW benchmark)
  - ✓ OpenFoam (3D lid driven cavity, 80^3)



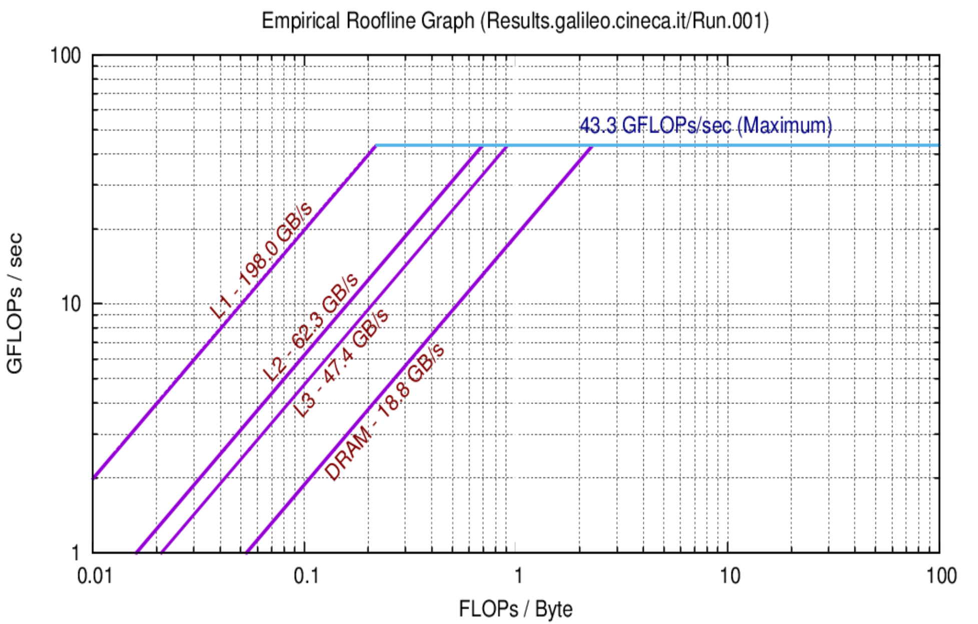
← 5.5x Linpack

← 2.5x Stream

← 1.8x OpenFoam

HPC status and future trends: roofline model

- The roofline model
- Performance bound (y-axis) ordered according to arithmetic intensity (x-axis) (i.e. GFLOPs/Byte)



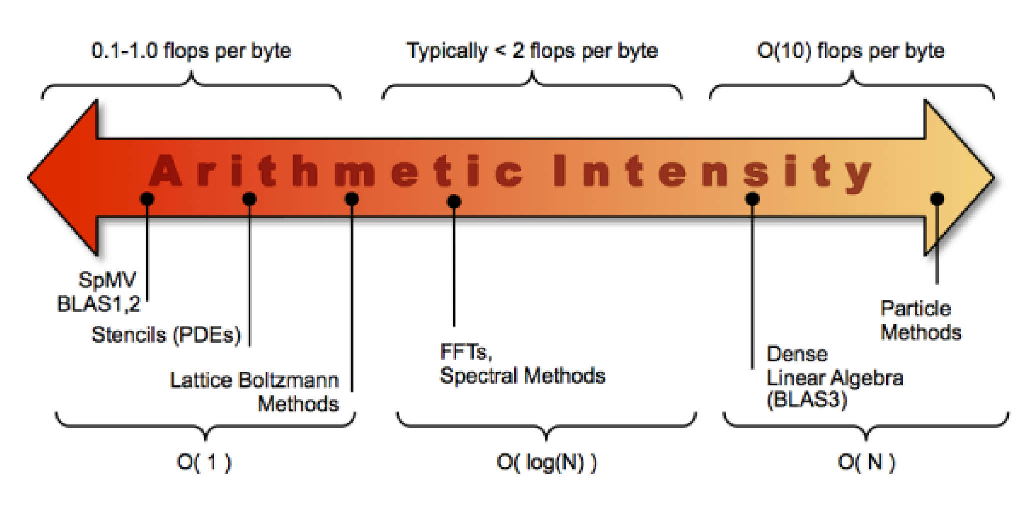
HPC status and future trends: Arithmetic intensity

Arithmetic Intensity: is the ratio of total floating-point operations to total data movement (bytes): i.e. flops per byte

Which is the OpenFoam arithmetic intensity?

- About 0.1, may be less…. ☹

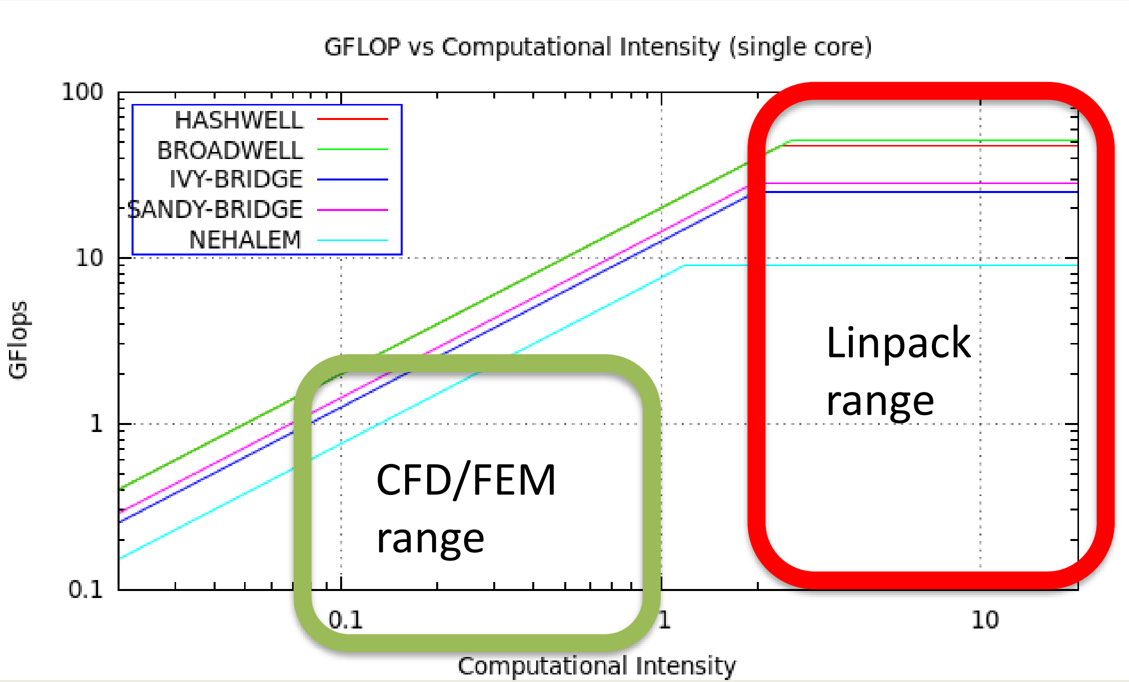
“Design and Optimization of OpenFOAM-based CFD Applications for Hybrid and Heterogeneous HPC Platforms”. Onazi et al, ParCFD14



HPC status and future trends. Which impact for OpenFoam?

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

"Theoretical FLOP/s isn't therefore a good indicator of how applications such as CFD ones (and many more) will perform"


GFLOP vs Computational Intensity (single core)
GFlops
Computational Intensity
HASHWELL
BROADWELL
IVY-BRIDGE
SANDY-BRIDGE
NEHALEM
Linpack range
CFD/FEM range