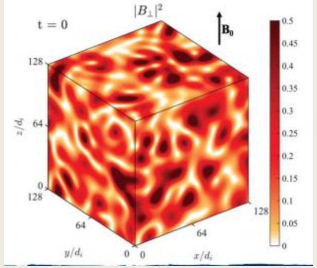
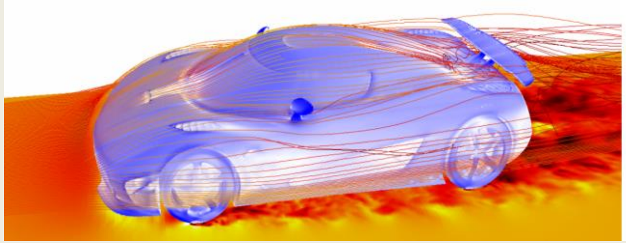
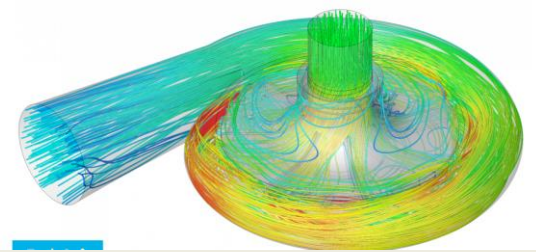
HPC Cineca Infrastructure:
State of the art, towards the exascale and OpenFOAM perspective

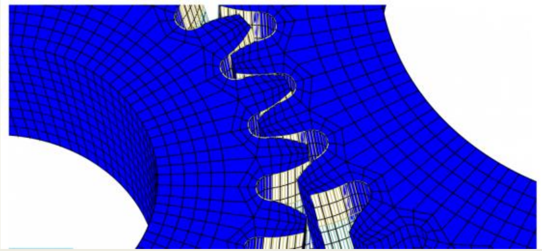


HPC Methods for Eng Applications
20 June 2017, Milan Italy

Ivan Spisso, Giorgio Amati
i.spisso@cineca.it, g.amati@cineca.it







CINECA SCAI SuperComputing Applications and Innovation

SuperComputing Applications and Innovation

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
- OpenFOAM perspective
  - Figure about performances on CINECA's HPC ecosystem
  - Parallel aspect and actual bottlenecks
  - Suggested future work: CFD4exascale

Who Am I?

Academic Achievements

- 01/12/2013 PhD in Computational Aeroacoustics, University of Leicester, UK.
- 26/09/2005 Master of second level (MSc) in `Satellites and Orbiting Platforms', Universita of Roma, La Sapienza
- 17/11/2003 Degree in Aeronautical Engineering (MEng) with full marks, University of Palermo, Italy.

Professional activities

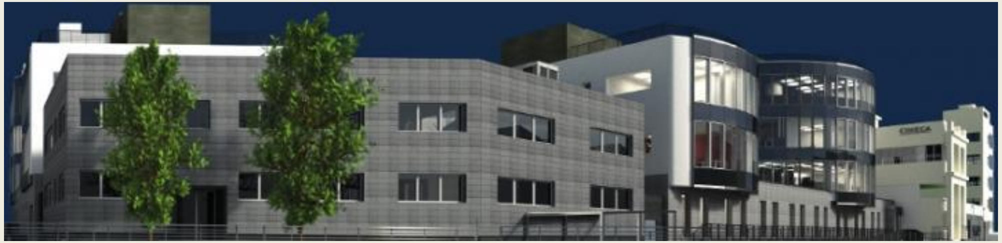
- From 07/2010 to date Staff member of SCAI (**SuperComputing Applications and Innovation**) HPC Department at CINECA, as consultant for academic and industrial CFD applications. CINECA, Casalecchio di Reno, Bologna, Italy.
- From 20/09/2009 to 20/12/2009 HPC-Europa2 Transnational Access fellowship at CINECA Casalecchio di Reno, Bologna, Italy.
- From 9/04/2009 to 02/05/2009 Visiting fellow at the IMFT (Institut de Mecanique de Fluides de Toulouse), Toulouse, France.
- From 01/2010 to 04/2010 Teaching assistant for the course of Fluid Dynamics, Introduction to Computing and Vector Calculus and Applications. University of Leicester.
- From 01/08/2006 to 31/07/2009 Marie Curie EST Fellow, Marie Curie multi-host EST network Aero-TraNet at the University of Leicester. Project title: Development of aprefactored high-order compact scheme for low-speed aeroacoustics.
- From 01/07/2005 to 31/07/2006 Qualied tutor for the CEPU centre of San Giovanni, Rome, for tuition
- in Engineering and Applied Sciences.
- From 11/04/2005 to 24/09/2005 Stage at the R&D Department of Aerosekur s.p.a., Latina, Italy. Computational Fluid Dynamics analysis of the SPEM re-entry system.

Cineca in a nutshell

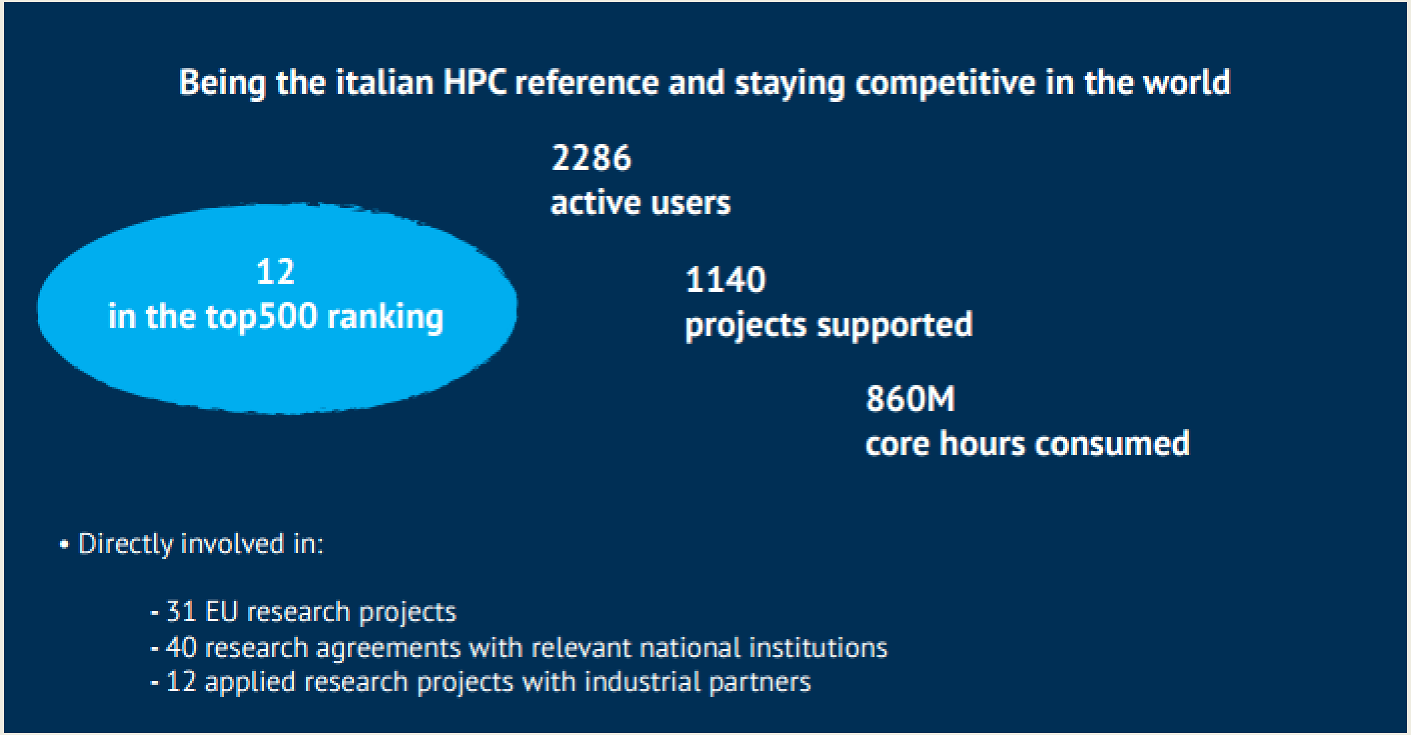
Cineca is a no-profit consortium composed by 70 italian universities, research institutions and the ministry of research.

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

2286
active users

12
in the top500 ranking

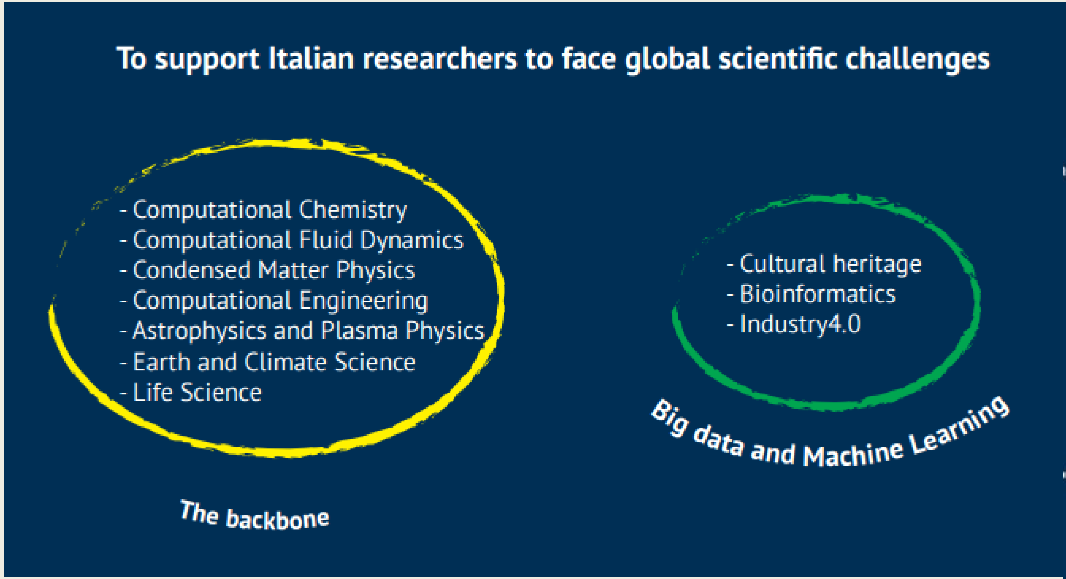
1140
projects supported

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



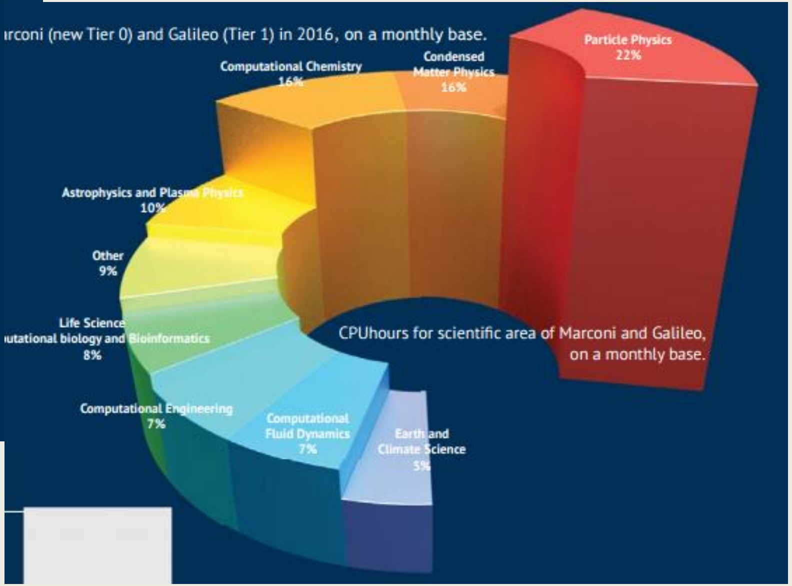
To support Italian researchers to face global scientific challenges

- 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



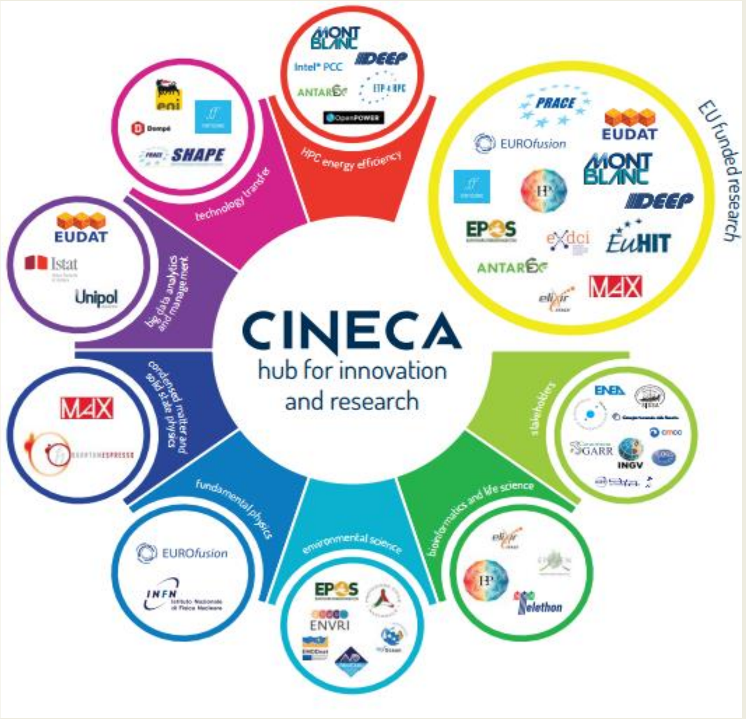
ırconi (new Tier 0) and Galileo (Tier 1) in 2016, on a monthly base.

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

The Cineca ecosystem

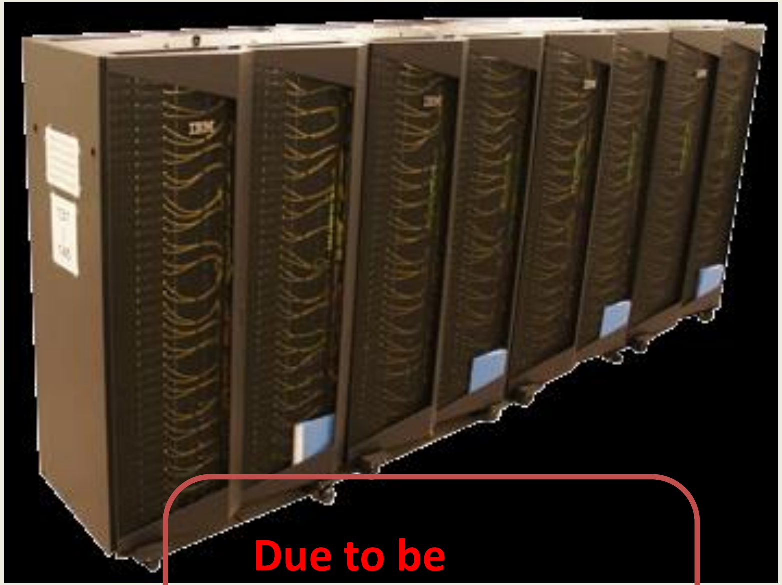
Cineca acts as a hub for innovation and research contributing to many scientifical and R&D projects on italian and european basis.

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


CINECA
hub for innovation and research
HPC energy efficiency
technology transfer
big data analytics and management
condensed matter and solid state physics
fundamental physics
environmental science
bioinformatics and life science
stakeholders
EU funded research

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 Summer 2017

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

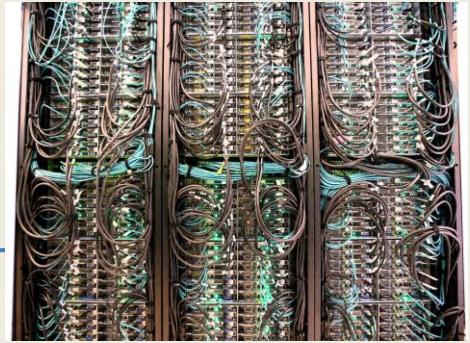
- Opened to production at the end of 2016
- 3600 Knights Landing compute nodes
- Intel Xeon Phi 7250 (68 cores) @1.40 GHz a.k.a. KNL
- 120GB RAM per node
- Default configuration: Cache/Quadrant
- TPP: 11 PFlop/s

MARCONI’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

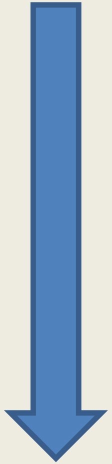
Westmere (a.k.a. plx.cineca.it)
  - Intel(R) Xeon(R) CPU E5645 @2.40GHz, 6 Core per CPU

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

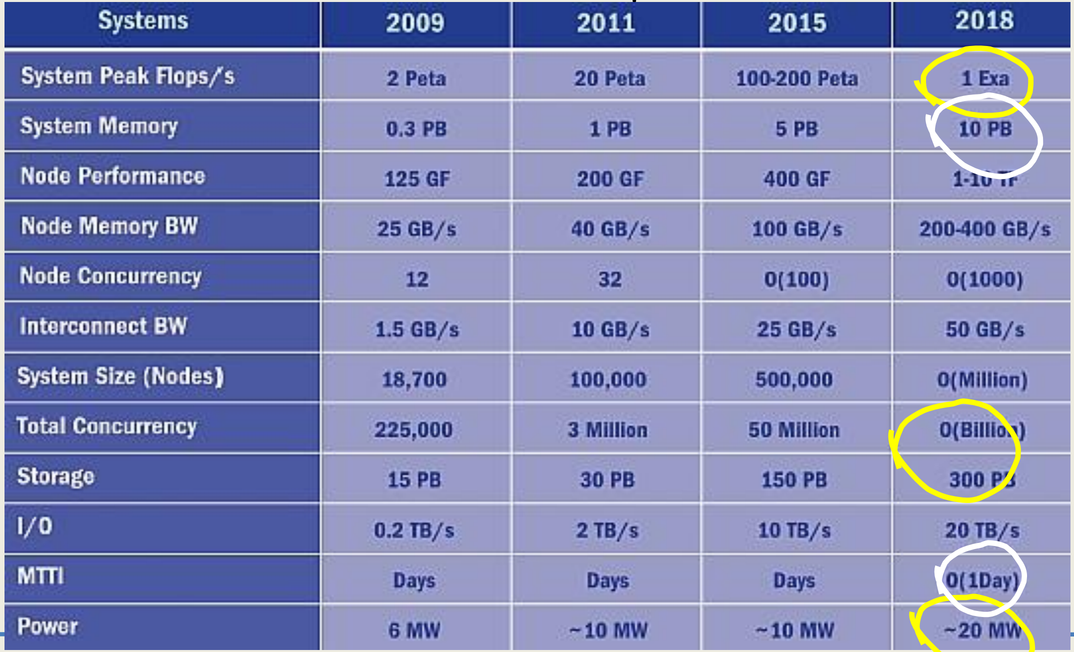


Increasing # of cores,
Same clock

Roadmap to Exascale

(architectural trends)

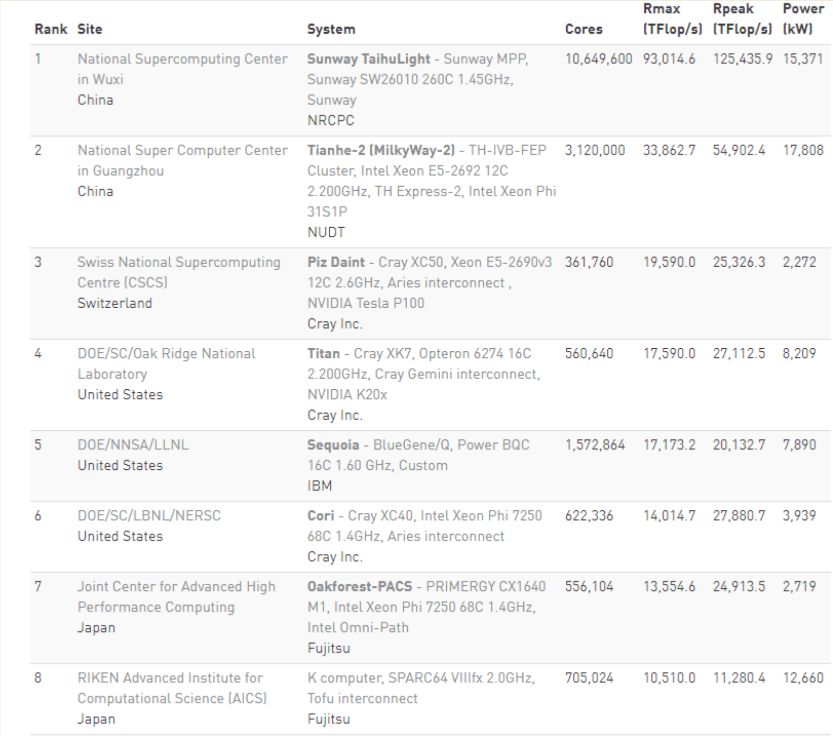
exascale: computing system capable of al least one exaFLOPs calculation per second.
exaFLOPs = 10^18 FLOPS or a billion of billion calculations per seconds

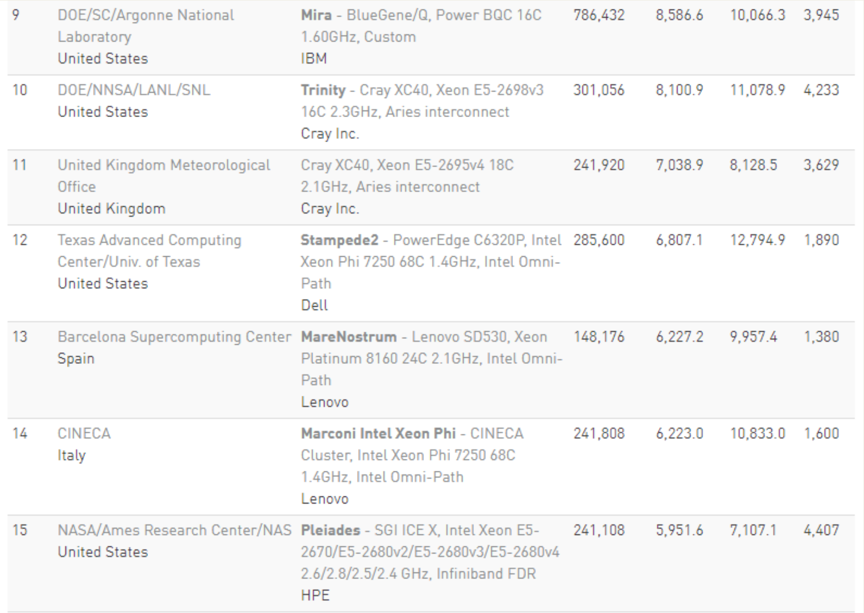


Systems	2009	2011	2015	2018
System Peak Flops/'s	2 Peta	20 Peta	100-200 Peta	1 Exa
System Memory	0.3 PB	1 PB	5 PB	10 PB
Node Performance	125 GF	200 GF	400 GF	1-10 TF
Node Memory BW	25 GB/s	40 GB/s	100 GB/s	200-400 GB/s
Node Concurrency	12	32	0(100)	0(1000)
Interconnect BW	1.5 GB/s	10 GB/s	25 GB/s	50 GB/s
System Size (Nodes)	18,700	100,000	500,000	0(Million)
Total Concurrency	225,000	3 Million	50 Million	0(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	0(1Day)
Power	6 MW	~10 MW	~10 MW	~20 MW

Top 500

(June 2017)

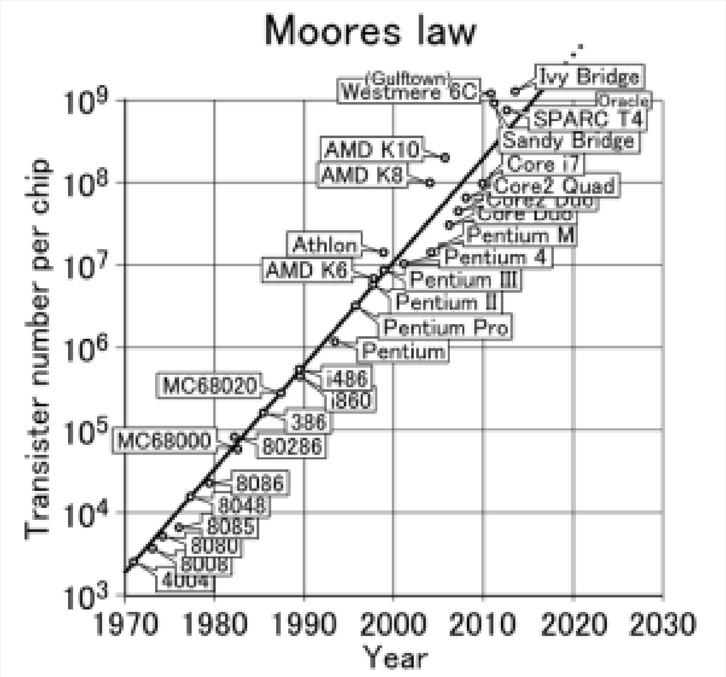


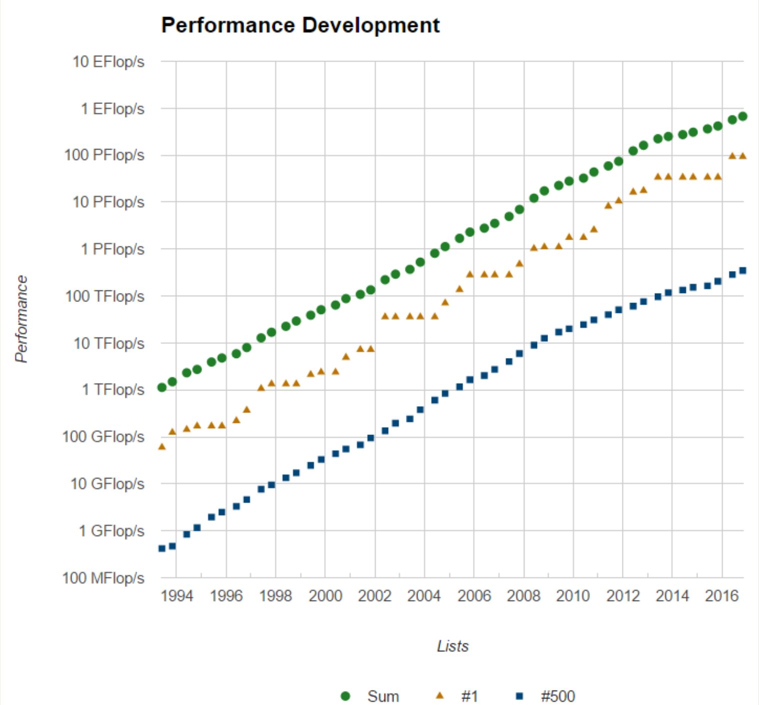


Rank	Site	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	National Supercomputing Center in Wuxi China	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway NRCPC	10,649,600	93,014.6	125,435.9	15,371
2	National Super Computer Center in Guangzhou China	**Tianhe-2 (MilkyWay-2)** - 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	Swiss National Supercomputing Centre (CSCS) Switzerland	**Piz Daint** - Cray XC50, Xeon E5-2690v3 12C 2.6GHz, Aries interconnect , NVIDIA Tesla P100 Cray Inc.	361,760	19,590.0	25,326.3	2,272
4	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
5	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
6	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
7	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
8	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
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	7,038.9	8,128.5	3,629
12	Texas Advanced Computing Center/Univ. of Texas United States	**Stampede2** - PowerEdge C6320P, Intel Xeon Phi 7250 68C 1.4GHz, Intel Omni-Path Dell	285,600	6,807.1	12,794.9	1,890
13	Barcelona Supercomputing Center Spain	**MareNostrum** - Lenovo SD530, Xeon Platinum 8160 24C 2.1GHz, Intel Omni-Path Lenovo	148,176	6,227.2	9,957.4	1,380
14	CINECA Italy	**Marconi Intel Xeon Phi** - CINECA Cluster, Intel Xeon Phi 7250 68C 1.4GHz, Intel Omni-Path Lenovo	241,808	6,223.0	10,833.0	1,600
15	NASA/Ames Research Center/NAS United States	**Pleiades** - SGI ICE X, Intel Xeon E5-2670/E5-2680v2/E5-2680v3/E5-2680v4 2.6/2.8/2.5/2.4 GHz, Infiniband FDR HPE	241,108	5,951.6	7,107.1	4,407

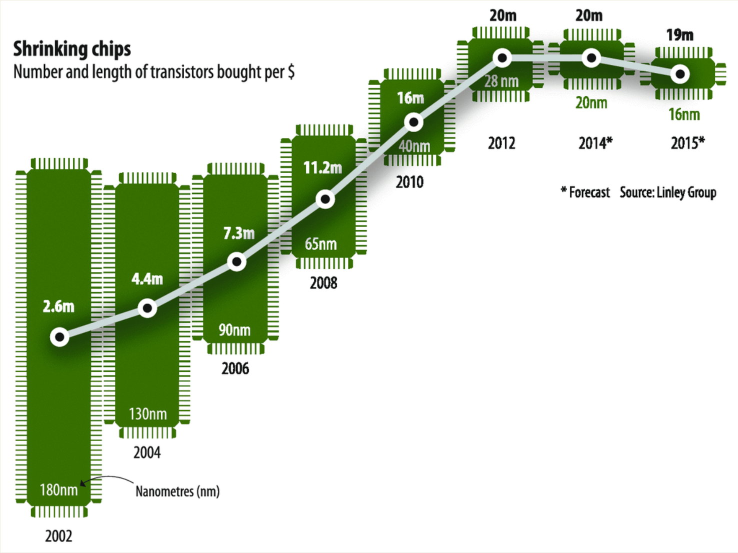
Moore’s Law - Chips

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)



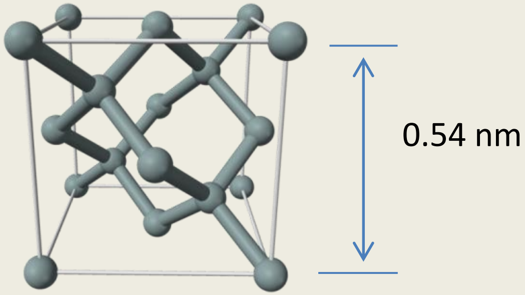


Moore’s Law - Dollars

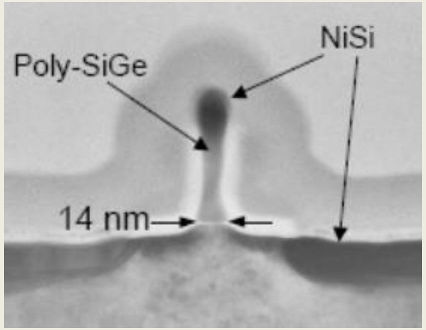


Oh-oh! Houston! we have a problem….

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.

$L' = L / 2$

$V' = V / 2$

$F' = F * 2$

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

$P' = P$

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.

$L' = L / 2$

$V' = \sim V$

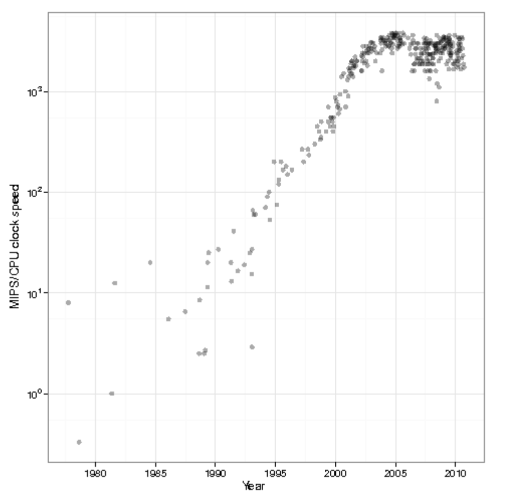
$F' = \sim F * 2$

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

$P' = 4 * P$

The power crisis!

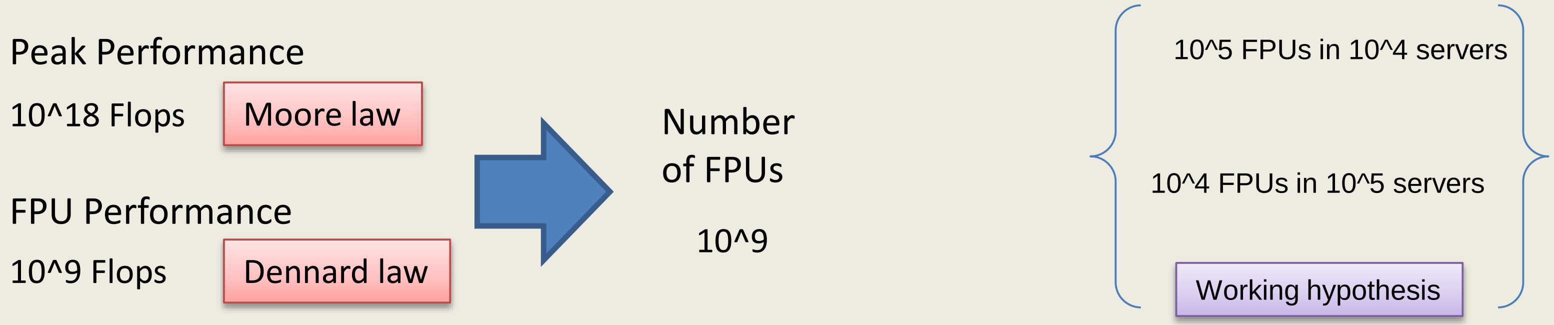
The core frequency and performance do not grow following the Moore's law any longer



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?

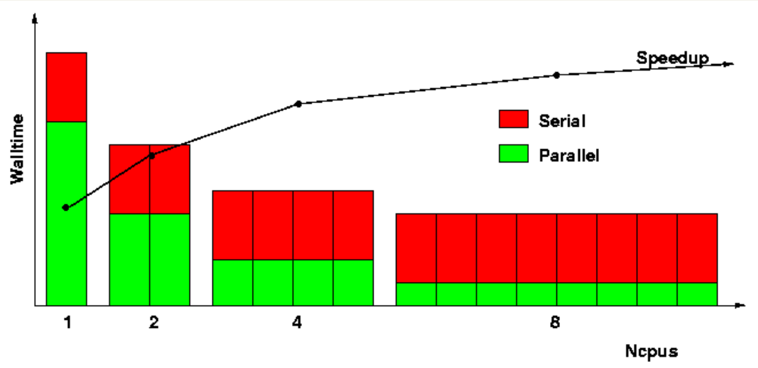
Very Big co-design Problem!

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

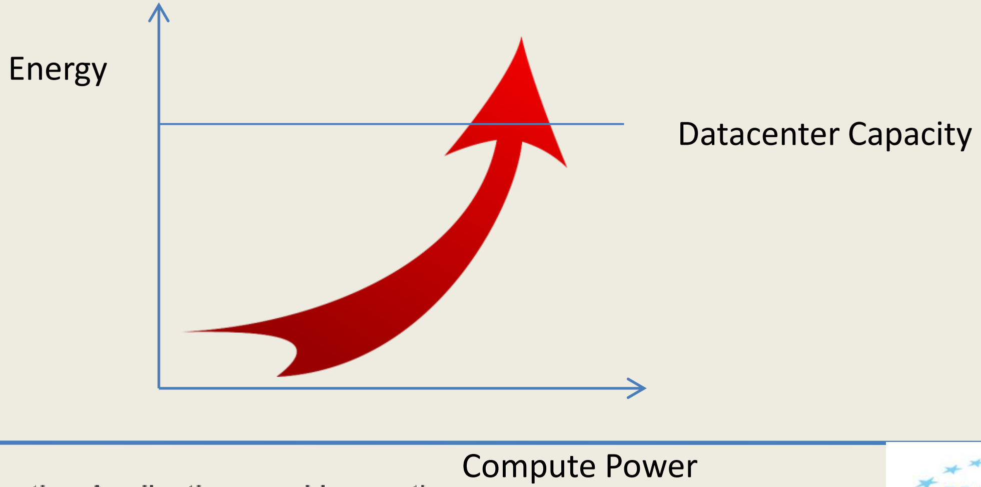
serial fraction= 0.000001

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



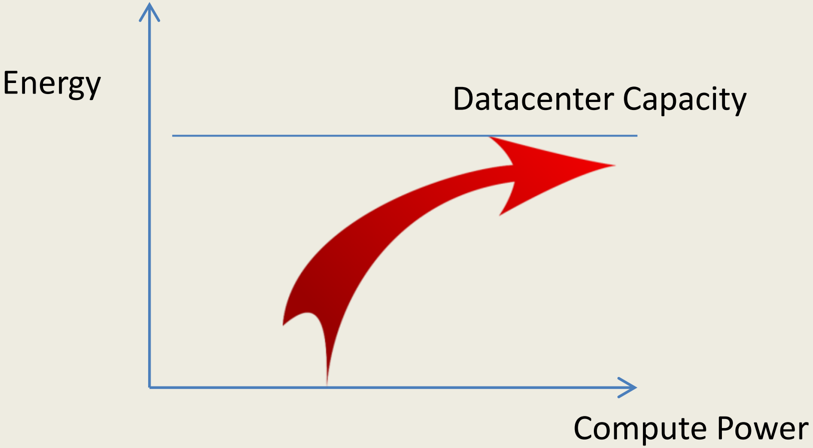
Change of Paradigm: Energy Efficiency

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

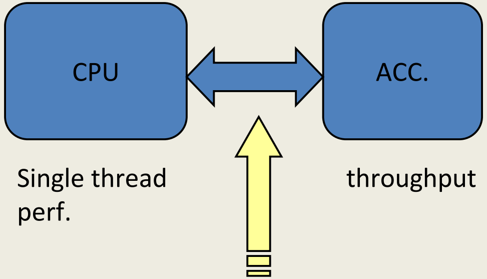
→

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

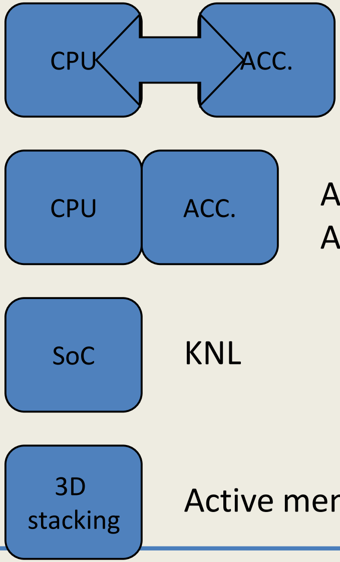


Architecture toward exascale



GPU/MIC/FPGA

bottleneck



OpenPower
Nvidia GPU

AMD APU
ARM Big-Little

KNL

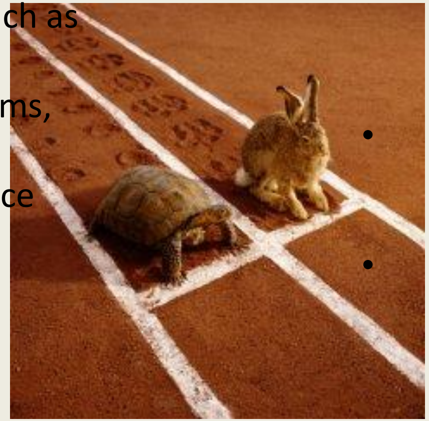
Active memory

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

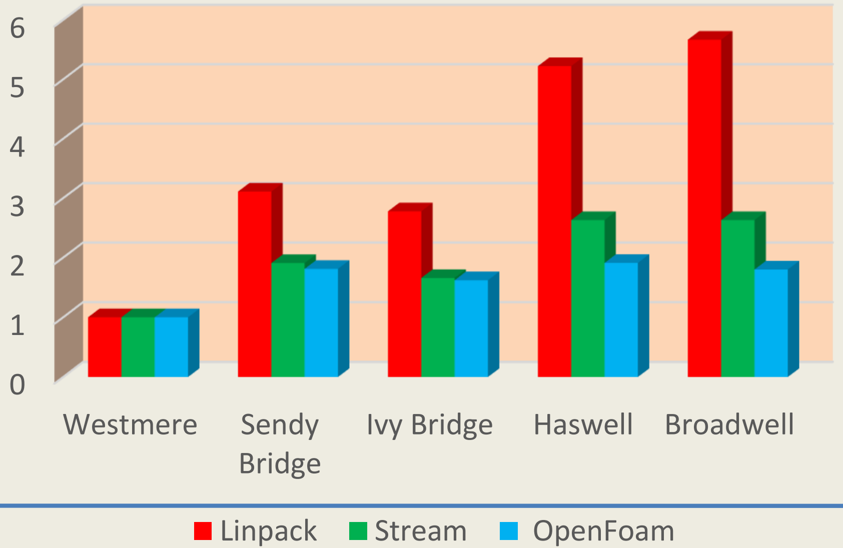


Hardware (hare)

- 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)



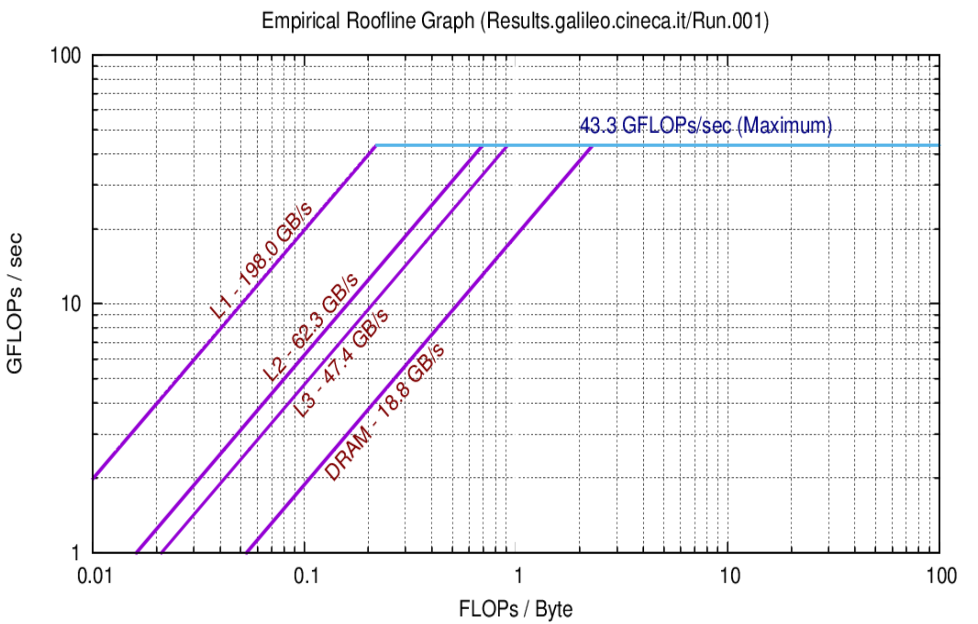
← 5.5x Linpack

← 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)



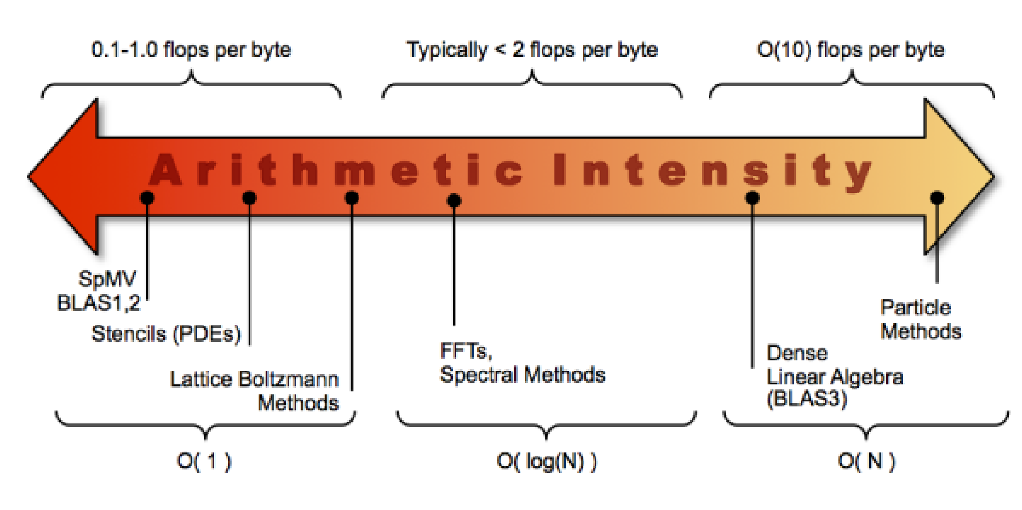
HPC status and future trends: Arithmetic intensity

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"

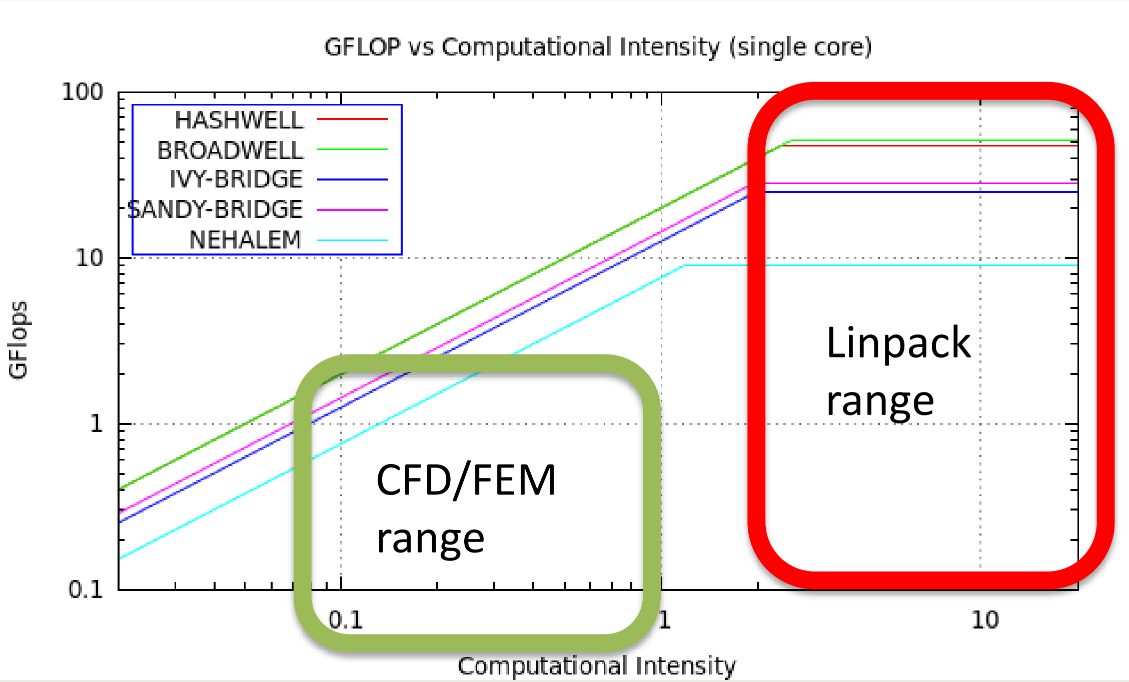


Figure about performances on CINECA's HPC ecosystem

- Our aim is to stress Marconi Machine (a Petascale KNL-based machine) in order to understand the bottleneck & theoretical limits for an efficient performance using future exa-scale machine

CAVEAT

- The test performed is a 3D lid-driven cavity, performance can be really different for different test-case
- Test-case info
  - KNL: OF rel. v1612+, compiled with intel, flag= -xMIC-avx512
  - BDW (reference): : OF rel. v1612+, compiled with intel,
  - 3D lid driven cavity

  - Size=300^3 (27M point)
  - T=0.20, dt=0.005, no output, viscosity=0.01
  - Size=400^3 (64M point)
  - T=0.10, dt=0.0025, no output, viscosity=0.01
  - Size=500^3 (125 MPoint)
  - T=0.10, dt=0.00125, no output, viscosity=0.01

1) Intranode performance

- Testing with 100^3 and 200^3 we found that
  - ✓ 64 task is the maximum intranode decomposition
- Some noisy measurements…

2) 300^3

- KNL Total time

Node	Task per node	Time	Speed-up
2	64	2240″	-
4	64	1159″	1.93
8	64	646″	3.47
16	64	404″	5.54
32	64	228″	9.82
64	64	167″	13.4
128	64	142″	15.8
256	64	109″	20.6
512	64	122″	18.4

- BDW Total time (KNL version = `-axMIC-AVX512`)

Node	Task per node	Time	Speed-up
8	32	733″ (735″)	-
16	32	302″ (302″)	-

3) 400^3

- KNL: Total time

Node	Task per node	Time	Speed-up
2	64	5130″	-
4	64	2605″	1.97
8	64	1427″	3.59
16	64	787″	6.52
32	64	447″	11.4
64	64	276″	18.6
128	64	192″	26.7
256	64	157″	32.3
512	64	141″	36.4

- BDW: total (KNL version = **`-axMIC-AVX512`**)

Node	Task per node	Time	Speed-up
16	32	971″ (967″)	
32	32	459″ (395″)	-
64	32	(159″)	
128	32	(90″)	
256	32	(95″)	

4) 500^3

- KNL: Total time (with faster all_reduce)

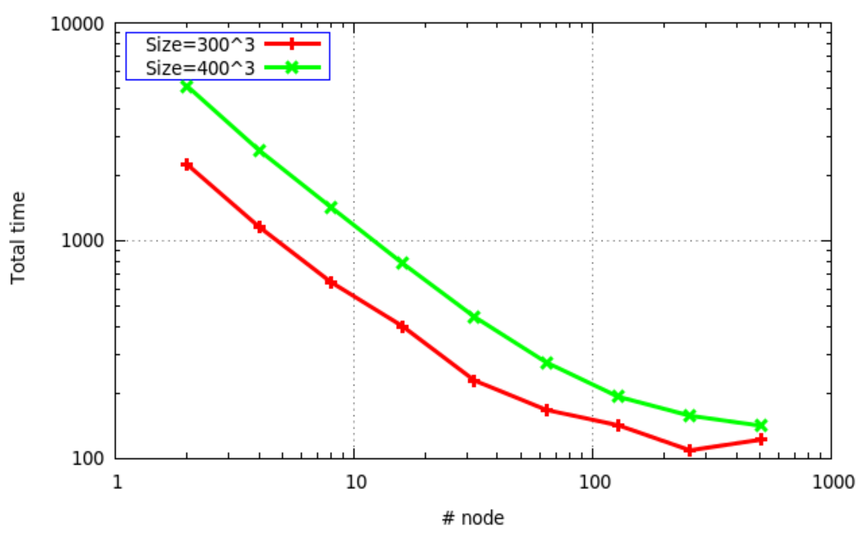
Node	Task per node	Time	Speed-up
2	64	-	
4	64	-	
8	64	2108”	
16	64	1084”	
32	64	751” (709”)	
64	64	536” (536”)	
128	64	332” (303”)	
256	64	247” (268”)	
512	64	-	

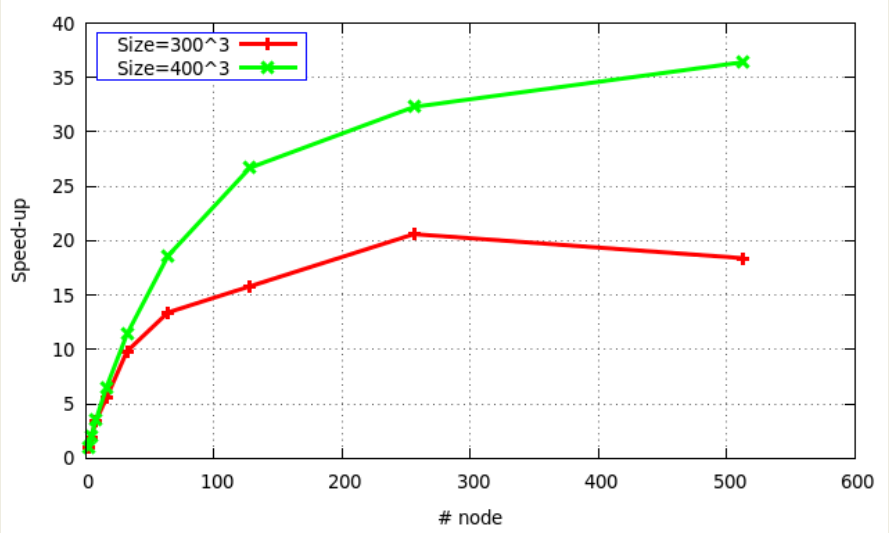
- BDW: total (KNL version = **–axMIC–AVX512**)

Node	Task per node	Time	Speed-up
16	32	1463” (1460”)	
32	32	561” (551”)	
64	32	- (235”)	

Figure about performances on CINECA’s HPC ecosystem

- Strong scaling (how the solution time varies with the number of processors for a fixed total problem size)
- Total time, 64 task per node





Fine Tuning

Total time in seconds

A. Original time

B. Different allreduce algorithm

C. Explicit taskset (bind to cpu)

D. Multilevel decomposition

- 300^3

Decom.	A	B	C	D	B+C	B+D	B+C+D
256x64	109”	101”	-	92”	89”	93”	86”

- 400^3

Decom.	A	B	C	D	B+C	B+D	B+C+D
256x64	157”	146”	147”	155”	143”	143”	140”
512x64	144”	133”	-	-	121”	-	-

Fine Tuning

Total time in seconds

A. Original time
B. Different allreduce algorithm
C. Explicit taskset (bind to cpu)
D. Multilevel decomposition

- 300^3
- 400^3

Parallel aspect

- OpenFOAM is first and foremost a C++ library used to solve in discretized form systems of Partial Differntial Equations (PDE).
- The “Engine" of OpenFOAM is the Numerical Method. To solve equations for a continuum, OpenFOAM uses a numerical approach with the following features: segregated, iterative solution, finite volume method, co-located variables, equation coupling.
- The method of parallel computing used by OpenFOAM is based on the standard Message Passing Interface (MPI) using the strategy of domain decomposition.

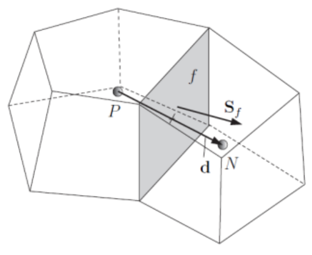


Figure: Finite Volume Discretization

Parallel aspect

- The geometry and the associated fields are broken into pieces and allocated to separate processors for solution.
- A convenient interface, Pstream, is used to plug any Message Passing Interface (MPI) library into OpenFOAM. It is a light wrapper around the selected MPI Interface

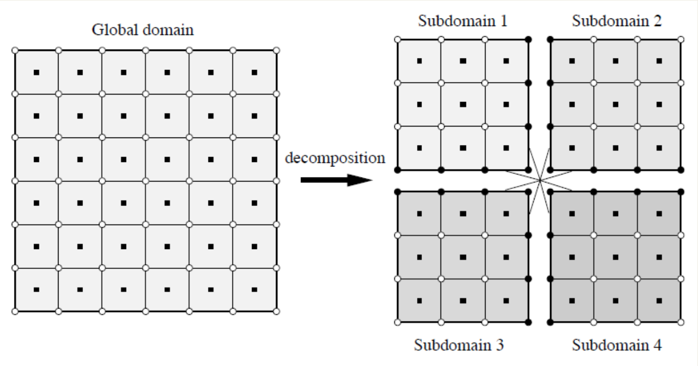


Figure: Zero Layer Domain Decomposition

Actual bottlenecks

An analysis has been done in the framework of PRACE 1IP to study the current bottlenecks in the scalability of OpenFOAM on Massively parallel clusters.

- Standard OpenFOAM scales reasonably well up to thousands of cores, upper limit order of 1,000 cores.
- An in-depth profiling identified the calls to the MPI AllReduce function in the linear algebra as core libraries as the main communication bottleneck
- A sub-optimal performance on-core is due the sparse matrices storage format that does not employ any cache blocking.

M. Culpo, Current Bottlenecks in the Scalability of OpenFOAM on Massively Parallel Clusters, PRACE White Paper, available on-line at [*www.prace-ri.eu*](http://www.prace-ri.eu)

Some references

http://www.prace-ri.eu/application-scalability/

P. Dagna, J.Hertzer: Evaluation of Multi-threaded OpenFOAM Hybridization for Massively Parallel Architectures,
PRACE White Paper, available on-line at http://www.prace-ri.eu/IMG/pdf/wp98.pdf

M. Moylesa, P. Nash, I. Girotto: Performance Analysis of Fluid-Structure Interactions using OpenFOAM
PRACE White Paper, available on-line at http://www.prace-ri.eu

M. Moylesa, P. Nash, I. Girotto: Performance Analysis of Fluid-Structure Interactions using OpenFOAM
PRACE White Paper, available on-line at http://www.prace-ri.eu/IMG/pdf/wp98.pdf

T. Ponweiser, P. Stadelmeyer, and T. Karasek, Fluid-Structure Simulations with OpenFOAM for Aircraft Design
PRACE white paper, http://www.prace-ri.eu/IMG/pdf/wp172.pdf.

A. Duran, M. S. Celabi, S. Piskin and M. Tuncel: Scalability of OpenFOAM for Bio-medical FLow Simulations,
PRACE White Paper, available on-line at
http://www.prace-ri.eu/IMG/pdf/WP162.pdf

Pham Van Phuc et al., Shimizu Corporation, Fujitsu Limited, Riken: Evaluation of MPI Optimization of C++ CFD Code on the K Computer,
SIG Technical Reports Vol. 2015-HPC-151 No. 19 2015/10/01. (in Japanese)

Actual Bottlenecks

Missing for a full enabling on Tier-0 Architecture:

Improve the parallelism paradigm, to be able to scale from the actual order of 1,000 cores to at least one order of magnitude (order of 10,000 or 100,000 procs).

Scalability of the linear solvers

- The linear algebra core libraries are the main communication bottlenecks for the scalability
- Whole bunch of MPI Allreduce stems from an algorithmic constraint and is unavoidable, increasing with the number of cores, . . . unless
- an algorithmic rewrite is proposed.

Generally speaking, the fundamental difficulty is the inability to keep all the processors busy when operating on very coarse grids. Need for communication-friendly agglomeration (geometric) linear multigrid solver.

Actual Bottlenecks

Improve the I/O, which is a bottleneck for big simulation.
For example LES/DNS with hundreds of cores that requires very often saving on disk.

- State of the art: A few million cells is now considered relatively small test case. Cases of this size will not scale usefully beyond 1K cores and there is not much to be done to improve this.
- Where we are looking at is radical scalability =) The real issues are in the scaling of cases of 100's of millions of cell on 10K+ cores.

Suggestions

Tune your application on HPC enviroment

- strong scaling =) how the solution time varies with the number of processors for a fixed total problem size
- The performance results vary depending on different parameters including the nature of the tests, the solver chosen, the number of cells per processors, the class of cluster used, choice of MPI distributions, etc
- Choose the linear system solvers: use the geometrical multi-grid solver (GAMG) for very large problems [1]. The GAMG solver can often be the optimal choice, particularly for solving the pressure equation
- Compile OpenFOAM in SP (Single Precision), if possible for your application.

[1] W. Briggs, V. Henson, and S. McCormick, A Multigrid Tutorial: Second Edition Society for Industrial and Applied Mathematics, 2000.

Suggested future work: CFD4exascale

- We are in the phase of building a consortium to apply for a big H2020 projects to enable OF to be used to the up-coming generation of Tier-0 clusters.
- FET-HPC call. Topic: Transition to Exascale Computing, Dead-line: 26 September 2017
- CINECA will act as HPC core partner during the preparatory phase and will support the co-design, provide the HPC infrastructure and the related competences.

Transition to Exascale Computing

Topic Description:

Specific Challenge: **Take advantage of the full capabilities of exascale computing**, in particular through high-productivity programming environments, system software and management, exascale I/O and storage in the presence of multiple tiers of data storage, supercomputing for extreme data and emerging HPC use modes, **mathematics and algorithms for extreme scale HPC systems** for existing or visionary applications, including data-intensive and extreme data applications in scientific areas such as physics, chemistry, biology, life sciences, materials, climate, geosciences, etc.

e) **Mathematics and algorithms for extreme scale HPC systems and applications working with extreme data**: Specific issues are quantification of uncertainties and noise, multi-scale, multi-physics and extreme data. **Mathematical methods, numerical analysis, algorithms and software engineering for extreme parallelism should be addressed**. **Novel and disruptive algorithmic strategies should be explored to minimize data movement as well as the number of communication and synchronization instances in extreme computing**. Parallel-in-time methods may be investigated to boost parallelism of simulation codes across a wide range of application domains. Taking into account data-related uncertainties is essential for the acceptance of numerical simulation in decision making; a unified European VVUQ (Verification Validation and Uncertainty Quantification) package for Exascale computing should be provided by improving methodologies and solving problems limiting usability for very large computations on many-core configurations; access to the VVUQ techniques for the HPC community should be facilitated by providing software that is ready for deployment on supercomputers.