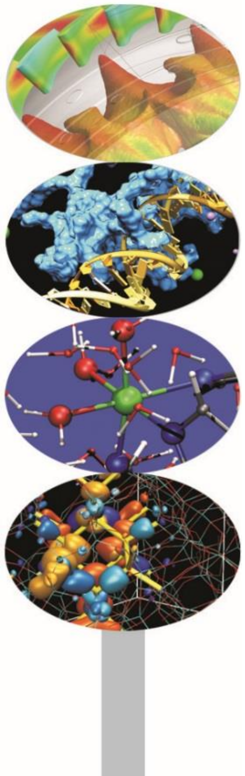
HPC-CINECA infrastructure: The New Marconi System



HPC methods for Computational Fluid Dynamics and Astrophysics

Giorgio Amati, g.amati@cineca.it

Agenda

1. New Marconi system
   - ✓ Roadmap
   - ✓ Some performance info
   - ✓ Caveats
2. Cineca HPC environment
3. Personal opinions about performance & HPC

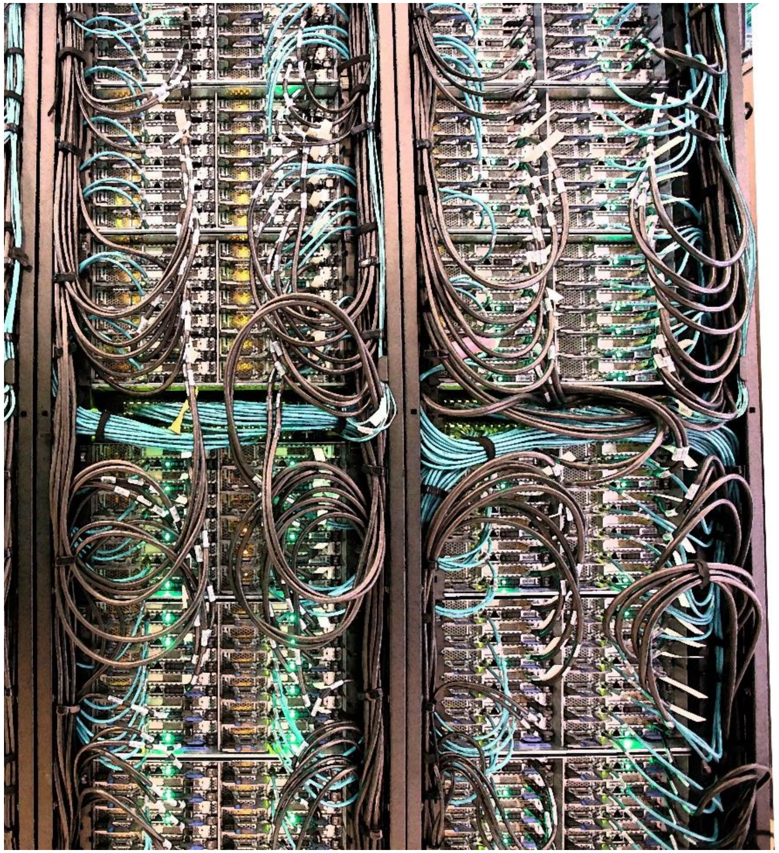
Marconi Road Map

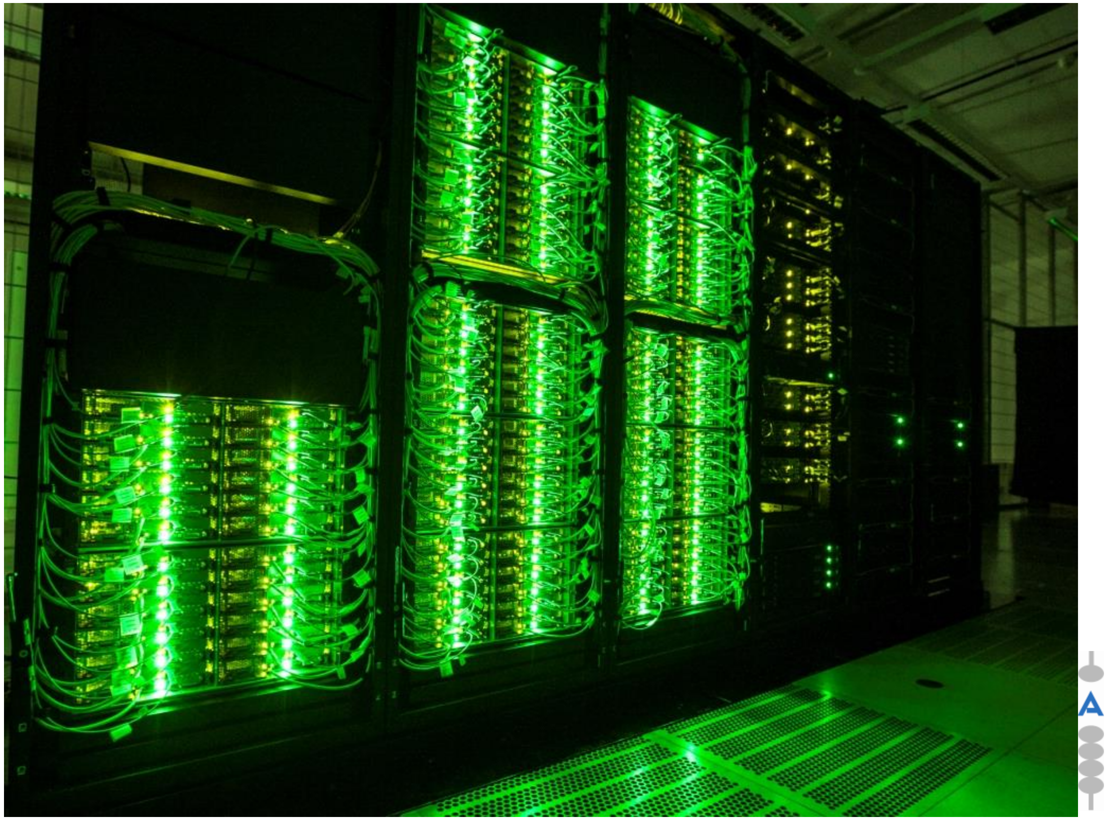
Marconi is the new Tier-0 HPC machine from LENOVO.

In its final version (A3) it will be a 18 PFlops machine (Peak value)

- A1 (06/2016):
  - ✓ 2 PFlops; based on classical x68 Intel CPU (Broadwell),
- A2 (11/2016)
  - ✓ +11 PFlops; based on intel Knights Landing (KNL)
- A3 (07/2017)
  - ✓ +5 PFlops; based on new x86 Intel CPU (Sky Lakes, 20 cores per CPU)
- Planning to increase A3 phase with 2 more PFlops

Marconi (A1)





Marconi status (A1)

- 1512 compute nodes, each with 2 CPU:
  - ✓ RAM = 128 GB
  - ✓ Intel(R) Xeon(R) CPU E5-2697 v4 @2.30GHz, 18 cores
  - ✓ cache size: 46080 KB
- Switch Intel OmniPath: world biggest OPA installation in the Ranked 46th in 06/2016 Top500 list.
- Configuration (at 10/2016):
  - ✓ S.O: **`Linux r037c06s02 3.10.0-327.36.1.el7.x86_64`**
  - ✓ OPA stack: **`rel. 10.2.0.0.158`**
  - ✓ **`MTU = 10KB`**
  - ✓ No turbo mode (max clock = 2.3 Ghz)
  - ✓ No hypertrhreading

Marconi status (A2)

Marconi Machine:

- ✓ 3600 KNL compute nodes
- ✓ Stand-alone version

Configuration (at 10/2016):

- ✓ All rack and nodes installed
- ✓ Under testing by LENOVO guys
- ✓ Working on LINPACK for Top500 list of 11/2016

Performance figures

Some Benchmark figures: baseline values

Single node performance

- Stream: 110 GB/s (Copy operation)
- Linpack: 1100 GFlops
- Hpcg: 21.6 GFlops

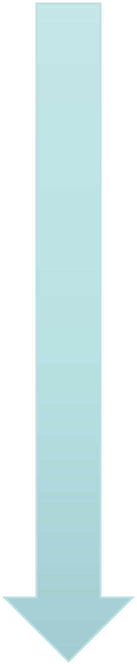
Cluster performance: linpack

Task	Nodes	size	GFLOPs
2	2	100000	2337
8	8	200000	9281
32	32	400000	36821
128	128	800000	145522

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



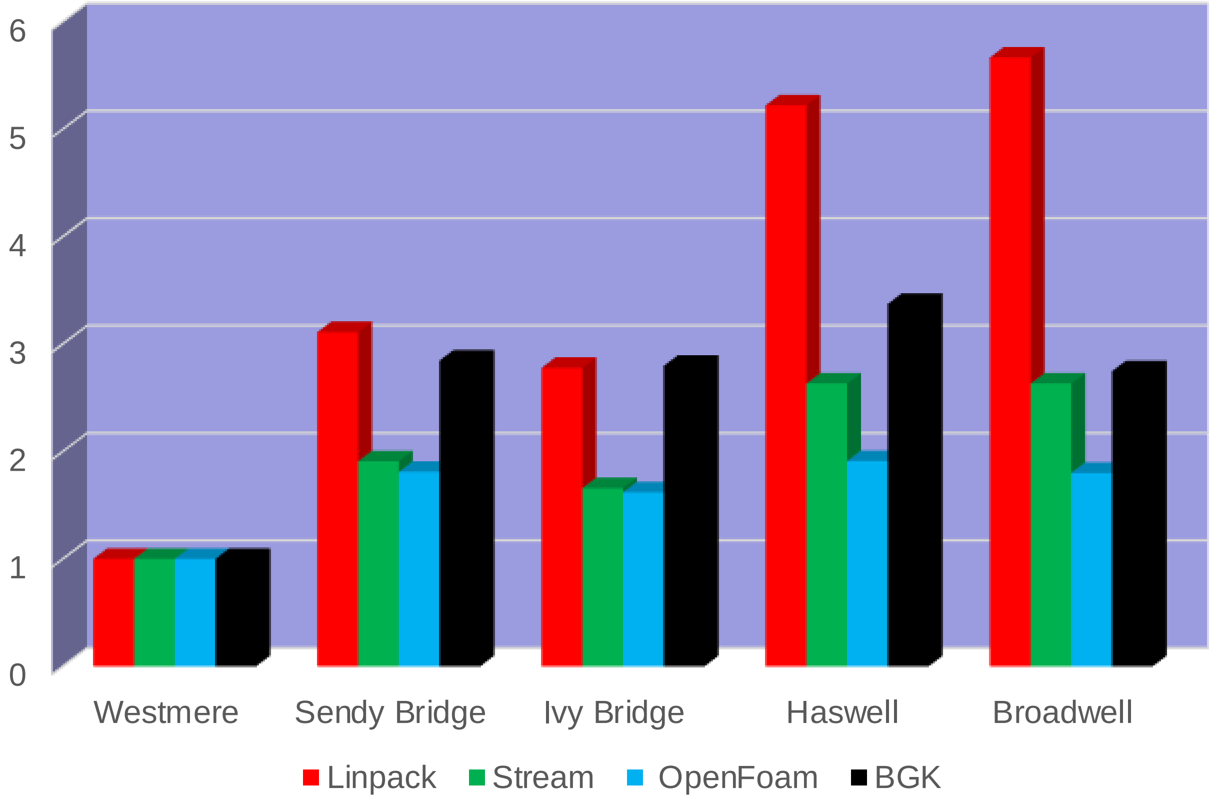
Increasing core

Same clock

Performance Evolution

About 6 year CPU evolution

- ✓ Linpack (Floating point Benchmark)
- ✓ Stream (Memory BW benchmark)
- ✓ OpenFoam (3D lid driven cavity, 80^3)
- ✓ BGK3d (3D Channel flow)



← **5.5x Linpack**

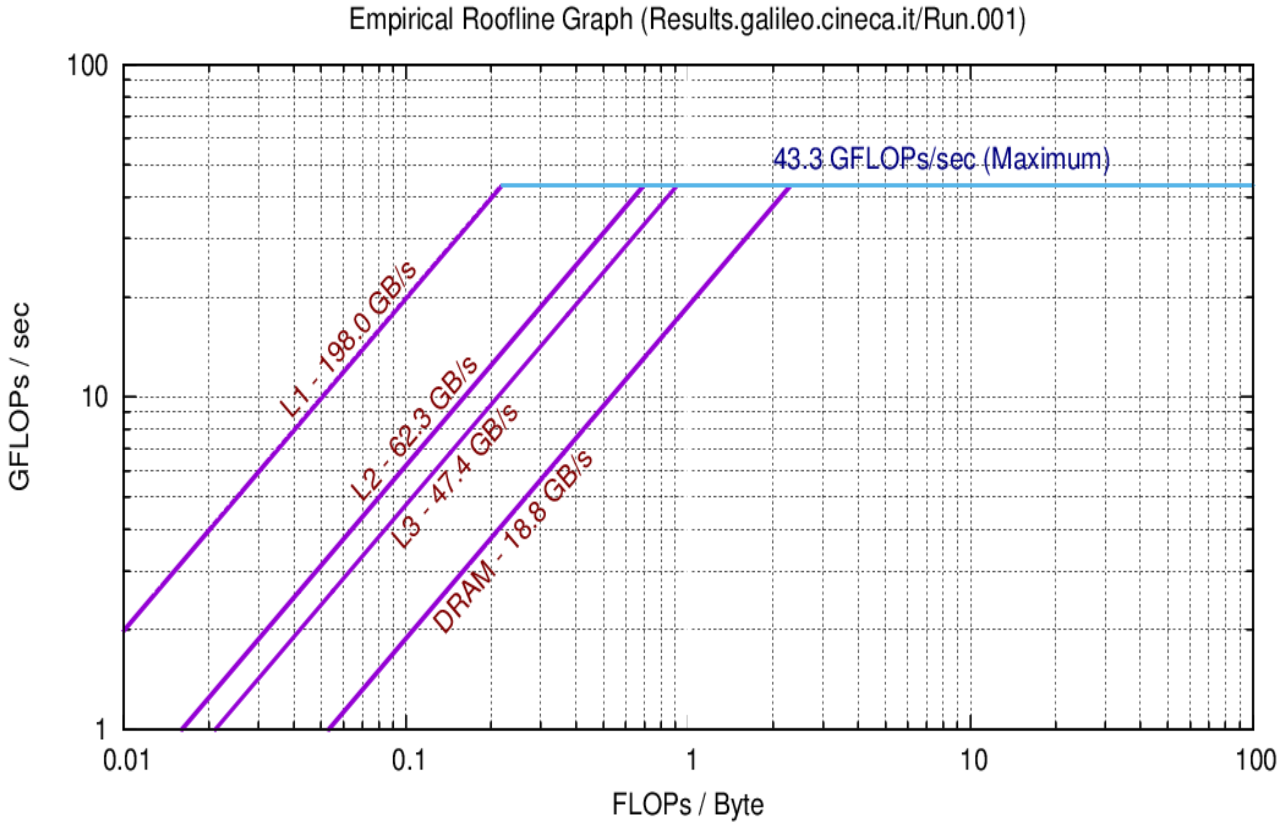
← **2.5x Stream**

← **1.8x OpenFoam**

← **2.7x BGK3d**

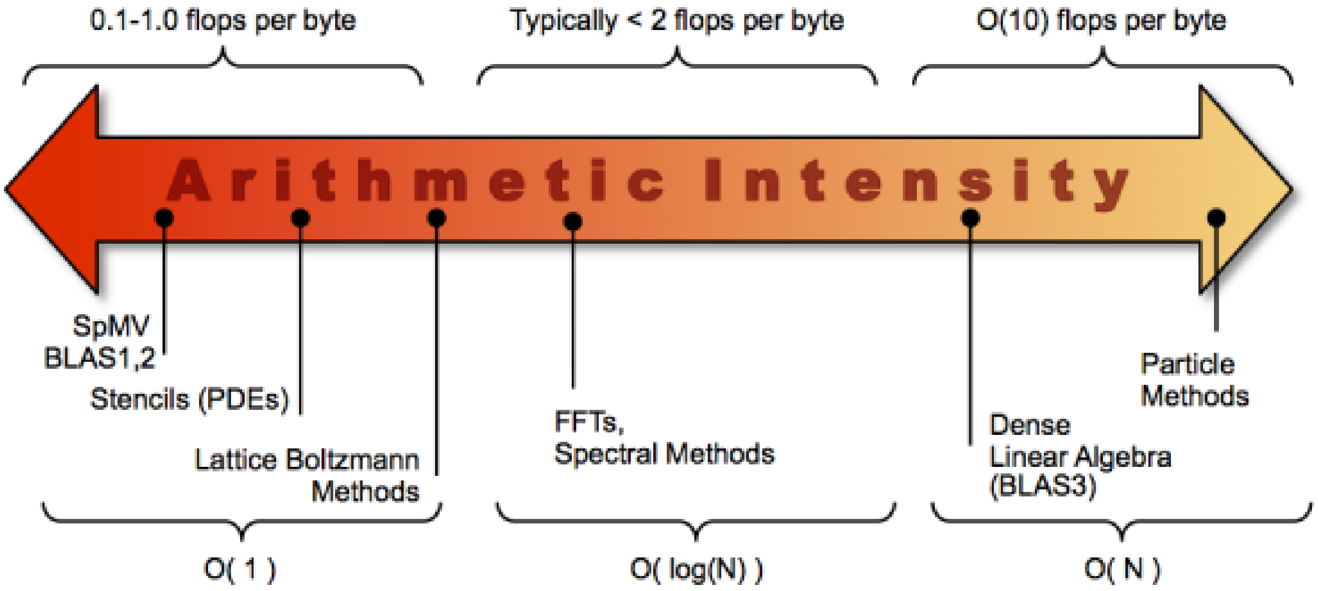
Boring performance issues/1

- Performance ordered according to arithmetic intensity (i.e. GFLOPs/Byte)
- http://crd.lbl.gov/departments/computer-science/PAR/research/roofline/



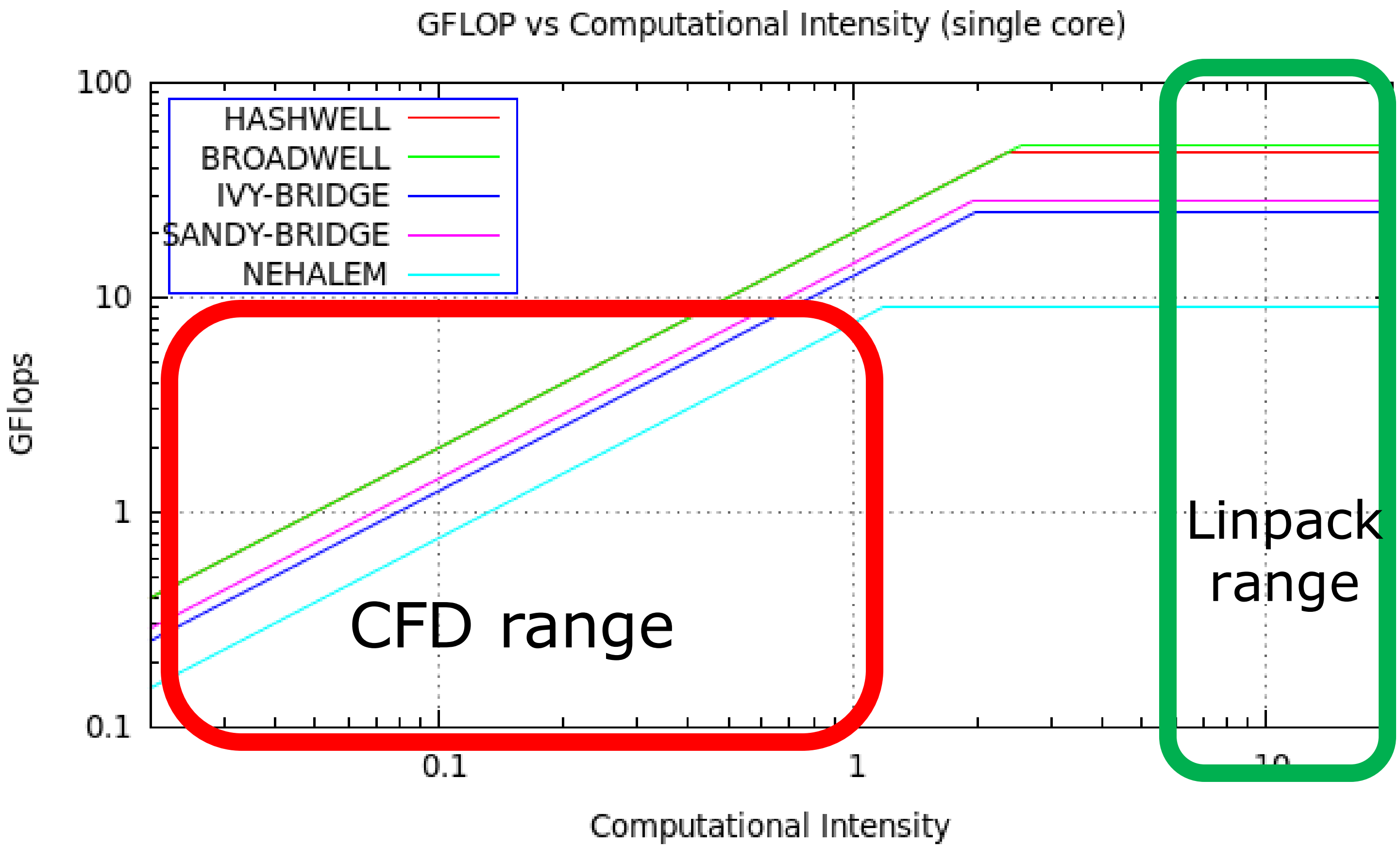
Boring performance issues/2

- The roofline model gives you an upper limit (BW or Floating point) according the arithmetic intensity of your code



Boring performance issues/3

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



Caveat

- Each Tier-0 is "one of his kind"
- Always an "experimental" machine
  - Intel OPA presented serious performance issues, fixed by a firmware upgrade
  - IntelMPI 2017 it the mpi library only OPA aware
  - Message transaction are performed by the host in OPA architecture
- Some "strange behavior" has been found
  - Using OpenFoam: task < cores
- Still to learn how increase performance
- Previous Intel Phi (KNC) not so successful

OpenFoam performance

Hint 1: internode scaling is no good (12 over 36)

Hint 2: look for the best ratio core/tasks

Task	Nodes	Time
16	**32**	51’’
20	**32**	61’’
26	**32**	70’’
28	**32**	74’’
30	**32**	75’’
32	**32**	77’’
34	**32**	93’’
36	**32**	359’’

Task	Nodes	Time
16	**16**	78’’
20	**16**	73’’
26	**16**	57’’
28	**16**	55’’
30	**16**	81’’
32	**16**	79’’
34	**16**	92’’
36	**16**	229’’

Task	Nodes	Time
16	**8**	262’’
20	**8**	120’’
26	**8**	228’’
28	**8**	235’’
30	**8**	110’’
32	**8**	108’’
34	**8**	116’’
36	**8**	186’’

Agenda

1. New Marconi system
2. Cineca HPC Environment
   - ✓ Machines & Storage
   - ✓ Access to HPC
   - ✓ Support to researcher
3. Personal opinions about performance & HPC

Machines & Storage

1. Computing facilities
   1. Marconi (tier0), 20PB local storage
      - ✓ CPU
      - ✓ intel KNL
   2. Galileo (tier1), 1.8 PB local storage
      - ✓ CPU
      - ✓ Intel KNC
      - ✓ Nvidia GPU K80
   3. Pico (Big Data), 0.6PB local storage
2. Shared Storage
   1. 4 Petabyte
   2. 12 PB tape LFTS storage system
3. User support (1st/2nd Level)

Access to CINECA HPC

1. via Agreement (e.g. EUROfusion, INFN)
2. via (peer-reviewed)
   - ISCRA: national project:
     - ✓ Class B: up to 2’000’000 core hours (twice a year)
     - ✓ Class C: up to 200’000 core hours (every month)
   - LISA: regional project
     - ✓ Supported by Regione Lombardia
   - PRACE: European project
     - ✓ Call 14, deadline 21/11 2017

Reference

- http://www.hpc.cineca.it/services/iscra
- http://www.hpc.cineca.it/services/lisa
- http://www.prace-ri.eu/

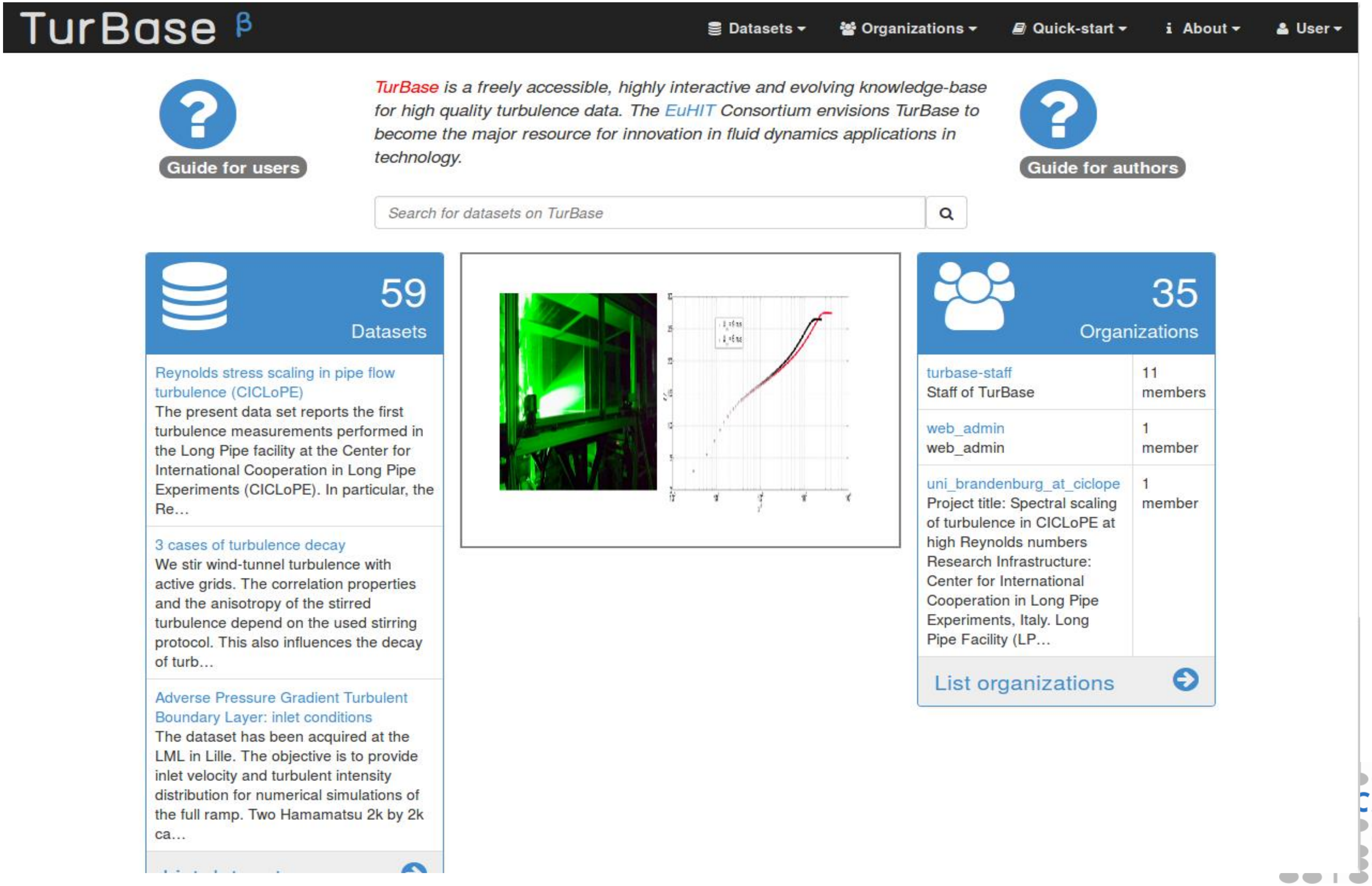
EuHIT: High-Performance infrastructure in turbulence

- ✓ Digital Library of Turbulence Data: iRODS storage at Cineca
- ✓ TurBase web-portal: freely accessible, highly interactive and evolving knowledge-base for high quality turbulence data
- ✓ Currently 59 datasets hosted, around 100 TB of online data
- ✓ High-Performance data exchange possible via GridFTP mechanism
- ✓ Online data inspection and previewing available
- ✓ Further info: f.salvadore@cineca.it

References

- https://www.euhit.org/
- Data Portal: http://turbase.cineca.it
- Online data inspection: https://turbaseservice.cineca.it

European Projects for data/2



TurBase is a freely accessible, highly interactive and evolving knowledge-base for high quality turbulence data. The EuHIT Consortium envisions TurBase to become the major resource for innovation in fluid dynamics applications in technology.

Reynolds stress scaling in pipe flow turbulence (CICLoPE)
The present data set reports the first turbulence measurements performed in the Long Pipe facility at the Center for International Cooperation in Long Pipe Experiments (CICLoPE). In particular, the Re…

3 cases of turbulence decay
We stir wind-tunnel turbulence with active grids. The correlation properties and the anisotropy of the stirred turbulence depend on the used stirring protocol. This also influences the decay of turb…

Adverse Pressure Gradient Turbulent Boundary Layer: inlet conditions
The dataset has been acquired at the LML in Lille. The objective is to provide inlet velocity and turbulent intensity distribution for numerical simulations of the full ramp. Two Hamamatsu 2k by 2k ca…

turbase-staff Staff of TurBase	11 members
web_admin web_admin	1 member
uni_brandenburg_at_ciclope Project title: Spectral scaling of turbulence in CICLoPE at high Reynolds numbers Research Infrastructure: Center for International Cooperation in Long Pipe Experiments, Italy. Long Pipe Facility (LP…	1 member

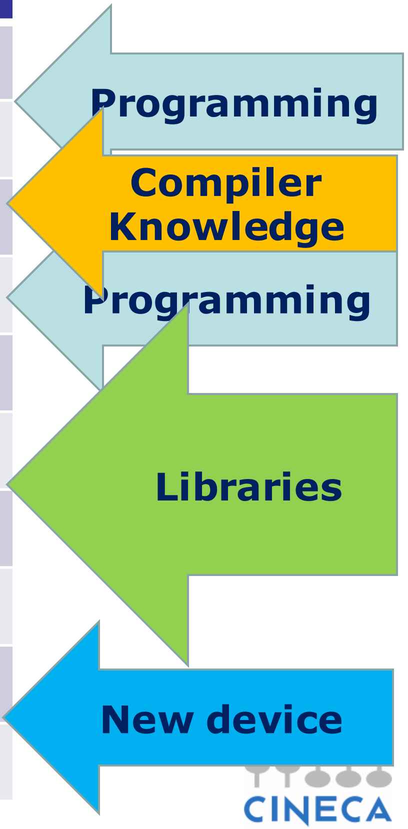
Agenda

1. New Marconi system
2. Cineca HPC environment
3. Personal ideas about performance & HPC
   - ✓ What to know to exploit performance
   - ✓ 20 years oh HW

Example 1

- Few correct way to coding, many wrong (for performance)
- Matrix-Matrix multiplication (time in seconds)

	Single prec.	Double prec.
Cache un-friendly loop	**7500''**	**7300''**
Cache friendly loop	**206''**	**246''**
Compiler Optimization	**84''**	**181''**
Handmade Optimization	**23''**	**44''**
Optimized library (serial)	**6.7''**	**13.2''**
Optimized library (OMP, 2 threads)	**3.3''**	**6.7''**
Optimized library (OMP, 4 threads)	**1.7''**	**3.5''**
Optimized library (OMP, 8 threads)	**0.9''**	**1.8''**
PGI accelerator (GPU)	**3''**	**5''**
CUBLAs (GPU)	**1.6''**	**3.2''**



Programming

Compiler Knowledge

Programming

Libraries

New device

20 Years of HW evolution

In 20 years many architecture/CPU are been used

- IBM 3090 (vector machine)
- APE Quadrics (SIMD machine)
- DEC/Compaq/HP EV4,EV5, EV6, EV68, EV7…
- Sun UltrasparcII (SMP, 8/14 CPU)
- IBM Power3/4/5
- NEC SX6 (vector machine)
- Intel Itanium

All dead
R.I.P.

- AMD Opteron
- Intel Xeon (Woodcrest, Clowertown, Nehalem,….. Broadwell)
- Nvidia GPU (Fermi, Tesla, Pascal)
- Intel Phi (KNC, KNL)

Some figure/1

Single core-cpu performance, BGK3D, double precision

- MLUPS: Mega lattice update per second (Higher is better)

Machine	MLUPS	Notes
APE 100	77	Using 512 core, SP, 1995/98
IBM Power3, 375MHz	1,4	2002
IBM Power4, 1300 MHz	3.5	2004
IBM Power5, 1900 MHz	5.8	2005
HP EV68, 1250 MHz	5.6	2004
HP EV7, 1100 MHz	6.0	2005
Intel Itanium2, 1500 MHz	8.3	2004
Intel Xeon, 2800 MHz	3.0	2004
AMD MP, 1533 MHz	3.3	2004
NEC SX6, 565 MHz	28.5	2004
IBM PowerPC, 2000 MHz	4.1	2004
Intel Core2	4.8	2007
AMD Opteron	10.1	2009

Personal ideas about performance

- Hardware evolution depends on economic & technological issues
- Researcher & Scientists reasons/desires are not considered at all! ☹
- Researcher & Scientists has to follow HW evolution, this means:
  - ✓ Numerical schemes used can be good/wrong according the used HW
  - ✓ Software must be upgraded always to keep pace with HW
  - ✓ Basic Knowledge of HW is mandatory
  - ✓ Parallel paradigm can vary over time
    - ✓ Pure MPI
    - ✓ Hybrid
    - ✓ What else (OpenCL, CUDA?)

…Even if it isn’t Marconi compliant

- Sauro Succi 2017 winner of “**Aneesur Rahman Prize for Computational Physics”**: "*For ground-breaking contributions to the development and application of the lattice Boltzmann method.*"

- https://www.aps.org/programs/honors/prizes/rahman.cfm

- Previous Italian winner were Carr & Parrinello (1995)

Thanks for patience…

- Any Questions?