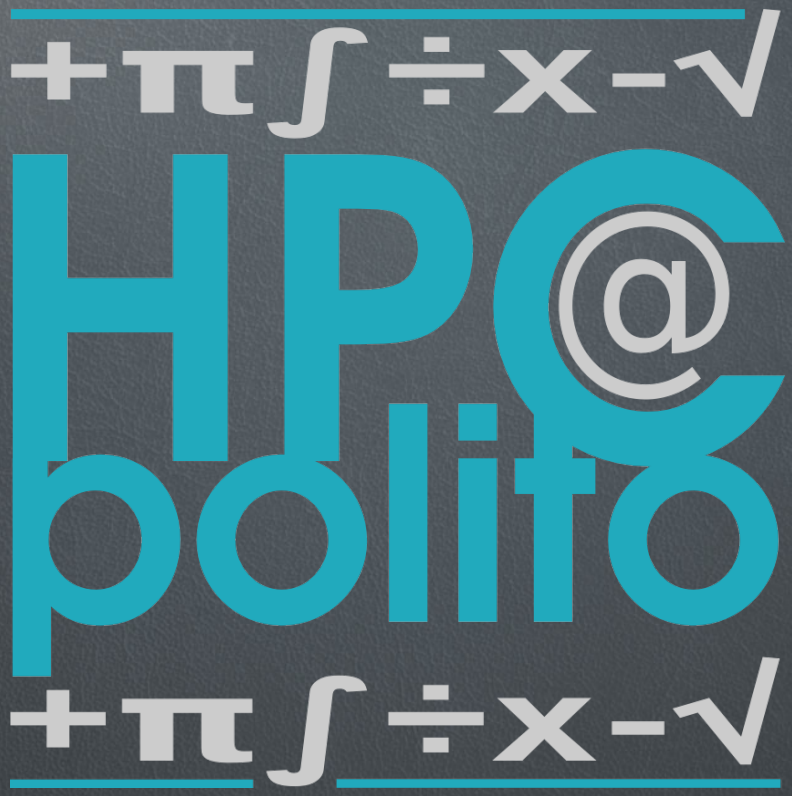
THE CASPER CLUSTER: PRESENT & FUTURE



Nicolò Nepote - Department of Control and Computer Engineering - March 12, 2014

AGENDA

1. CASPER IS...
2. ASK CASPER WHAT HE CAN DO FOR YOU
3. HOW ARE THINGS GOING ON?
4. CASPER IS GOING TO BE...

hpc.polito.it
hpc.dauin@polito.it
@hpc_polito

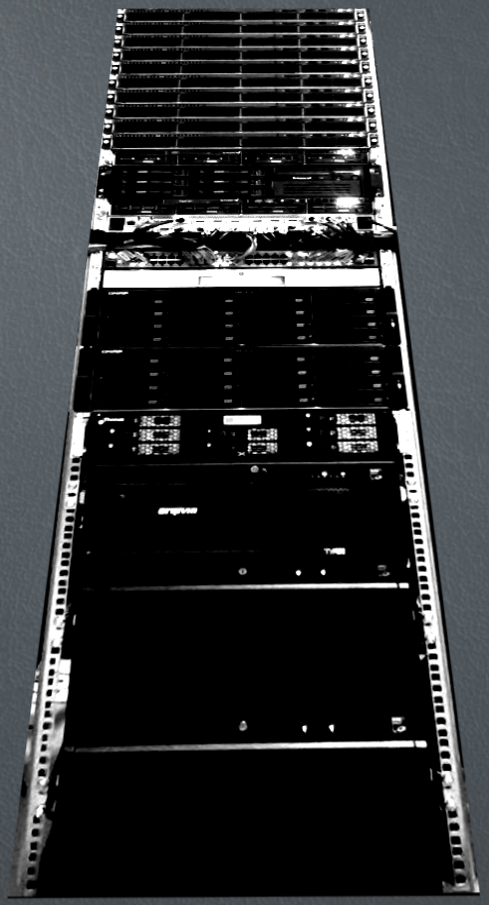
+π∫ ÷x-√ HPC@polito

CASPER IS...

A friendly ghost with my child's face?

It would be fun...

...but it's not the correct answer



A buzzing and noisy kubrickian monolith?

Less fun...

...but far more useful to research

CASPER, BACK IN 2008

1. CASPER is the HPC system available at HPC@POLITO
2. Cluster Appliance for Parallel Execution and Rendering
3. First use case as a Blender 3D render farm
4. Built as a “Beowulf style” cluster
5. Based on Linux and free/open management software
6. Possibly use open standards for high speed interconnection
7. Must provide engineering support
8. Need users to publish their papers and students to learn better: this is our mission
9. Budget = 0

CASPER, BACK IN 2008

- 0.16 TFLOPS
- Athlon XP single core
- 2 GB RAM/core
- 44 nodes
- Gigabit Ethernet
- Local storage
- Huge useless power consumption

not exactly what is called “a supercomputer”

ISSUES & SOLUTIONS 2008-2013

1. Slow network is useless network	InfiniBand
2. Lack of parallelism	Opteron
3. Master node as storage node	Dedicated NAS
4. We need more RAM, a lot more RAM, please!	128 GB/node 4 GB/core
5. Can you give priority to fellow research groups?	Priority queues
6. Help your users, they're your first resource (someone says)	Study, study, study!

Grant from the Board of Governors in 2012

CASPER TODAY

Architecture	Linux InfiniBand MIMD Distributed Shared-Memory Cluster
Node Interconnect	InfiniBand DDR 20 Gb/s on copper
Storage Interconnect	Ethernet 2x 10 Gb/s (bonding 802.3ad)
Service Network	Gigabit Ethernet 2x 1 Gb/s (bonding 802.3ad)
CPU Family	AMD Bulldozer
CPU model	Opteron 6276 2.3 GHz (turbo 3.0 GHz) 16 cores
Sustained performance	~ 3 TFLOPS (recalculating)
Computing Cores	448
Number of Nodes	14 (dual socket)
Total RAM Memory	1.8 TB DDR3 Registered ECC
Working Storage	47 TB on RAID 6, throughput near 800 MB/s
OS/Scheduler	ROCKS Clusters + GridEngine

CASPER TODAY

Compute nodes dedicated to fellow research groups

Standard compute nodes

Master node and Login node

InfiniBand DDR switching fabric

Ethernet switch

60 TB NFS-shared Storage via dual 10 Gbe

12 TB backup storage via Gbe

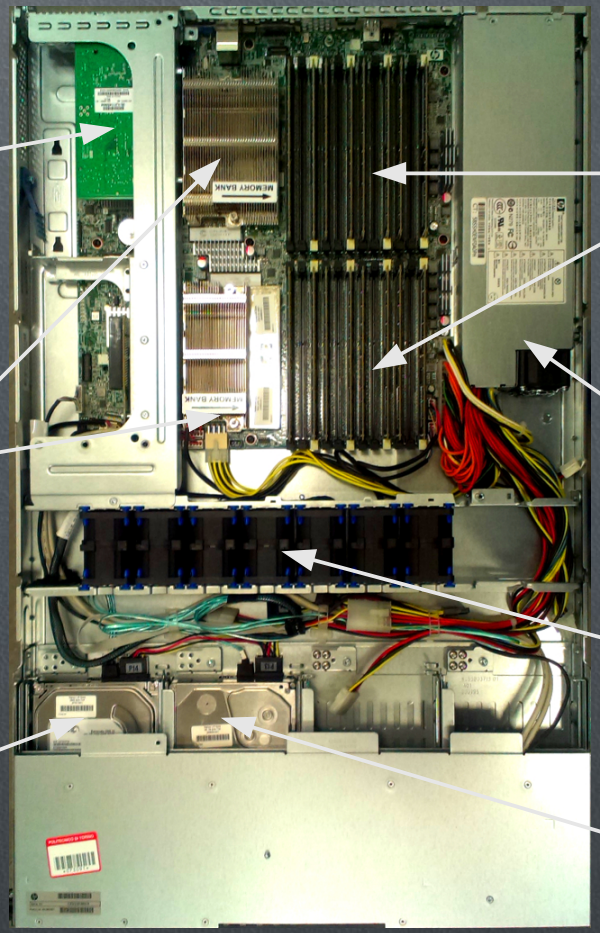
Compute nodes dedicated to fellow research groups

COMPUTE NODE HW CONFIGURATION

InfiniBand DDR/QDR HCA dual port with CX4/QSFP connectors

Dual socket AMD Opteron 6276

bootstrap HD (OS + temporary folders)



128 GB DDR3 Registered ECC

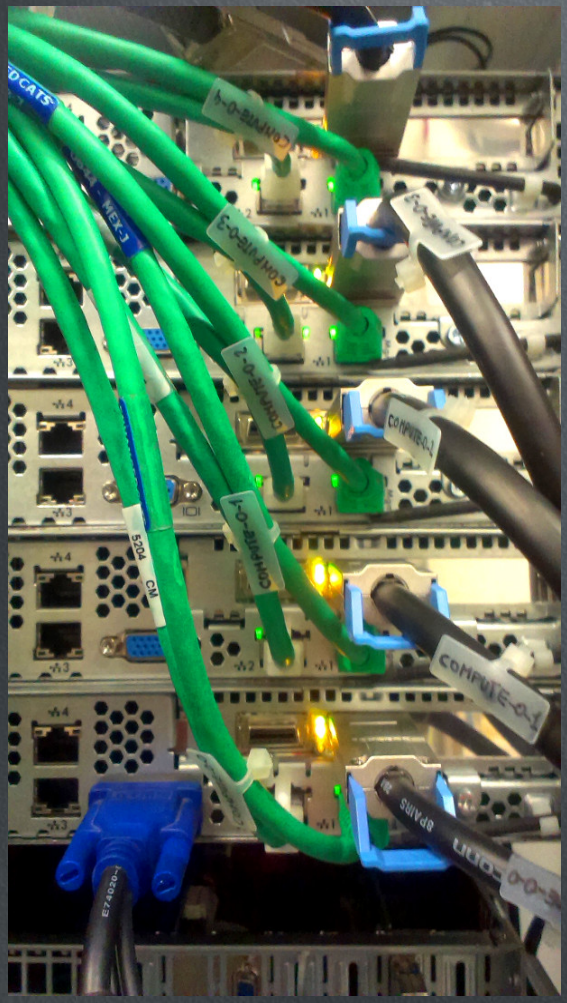
Redundant PSU

Cooling

1TB SATA 3 HD (dedicated to HDFS)

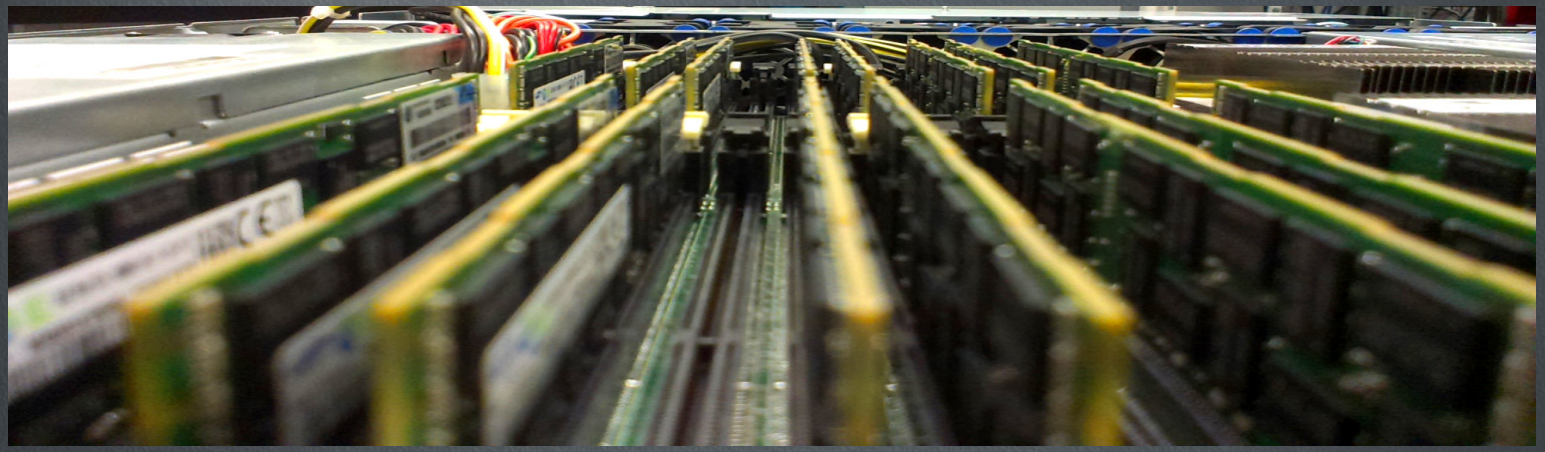
RACK FRONT

THE ADDED VALUE OF InfiniBand



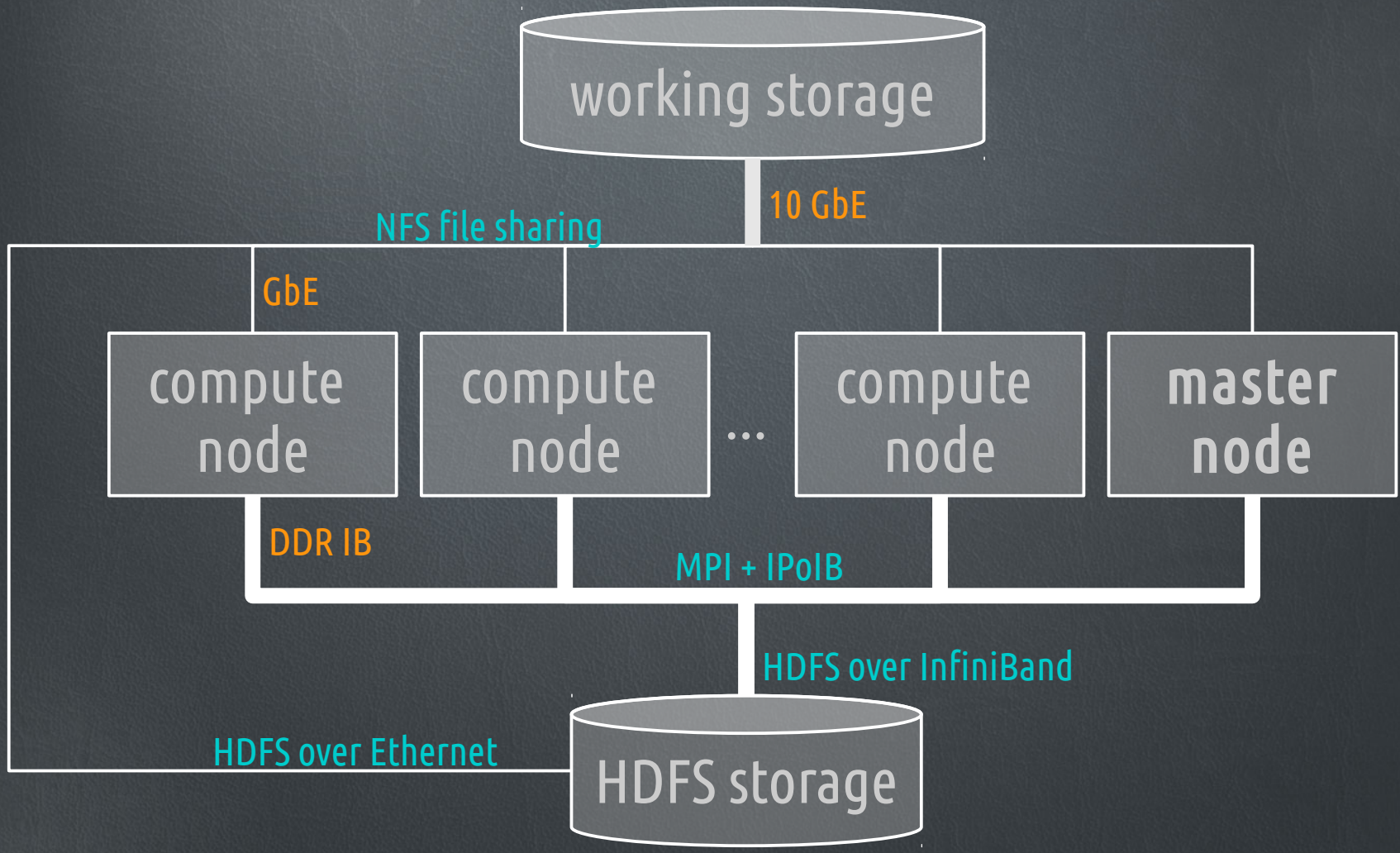
1. Layer 1-2 open (and expensive) high speed network technology
2. Industry standard in HPC
3. Delivers large bandwidth and low latency
4. Carries MPI traffic as well as IP (IPoIB)
5. Can really boost your computation

THE ADDED VALUE OF FAT NODES



1. CASPER (3 TFLOPS): 4 GB/core, 128 GB/node, 32 cores/node
2. CINECA's Fermi (2 PFLOPS) : 1 GB/core, 16 GB/node, 16 cores/node
3. Tyane 2 (33 PFLOPS): 320 MB/core, 64 GB/node, 24 cores/node (plus 2 Xeon Phi accelerators)
4. Fat nodes provide the user with high parallelism and large memory available for non-MPI, multi-threaded or even sequencial programs

CASPER LOGICAL ARCHITECTURE



ASK CASPER WHAT HE CAN DO FOR YOU

+π∫ ÷x-√ HPC@polito

SOME USE CASES AND STARTEGIES

MPI parallel programming

In-house code (C, Fortran) using OpenMPI, MPICH, etc.

Truly parallel through InfiniBand

MPI-capable 3rd party software suites

Star-CCM+, OpenFOAM, Quantum-Espresso, Gromacs...

Huston, we need a license!

Rendering of movies and hi-res static images

Blender 3D + Cycles

CASPER is a render farm again

Domain splitting simulations

Parallelize the program by dividing the domain into (mostly) independent regions

Can exploit MPI and InfiniBand

Multiple instances like level-0 parallelism

Same non-MPI program runs on different inputs under different conditions inside thousands of jobs

Often memory consuming

Easy to deploy and very effective

Map-reduce applications

Hadoop + HDFS

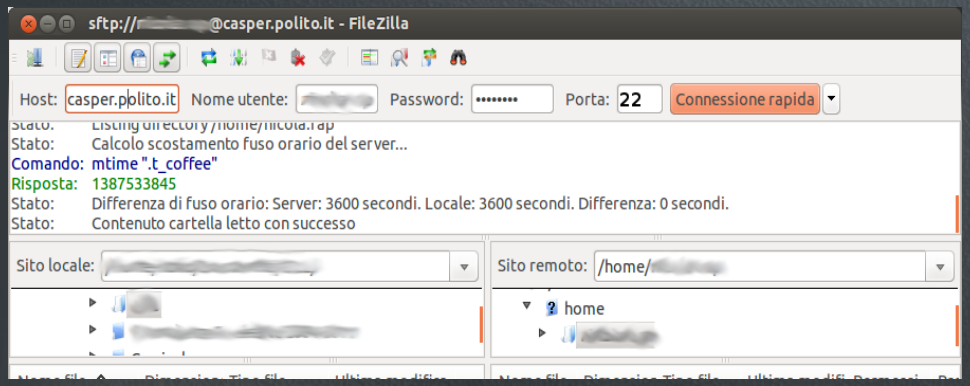
This is HPC for Big Data

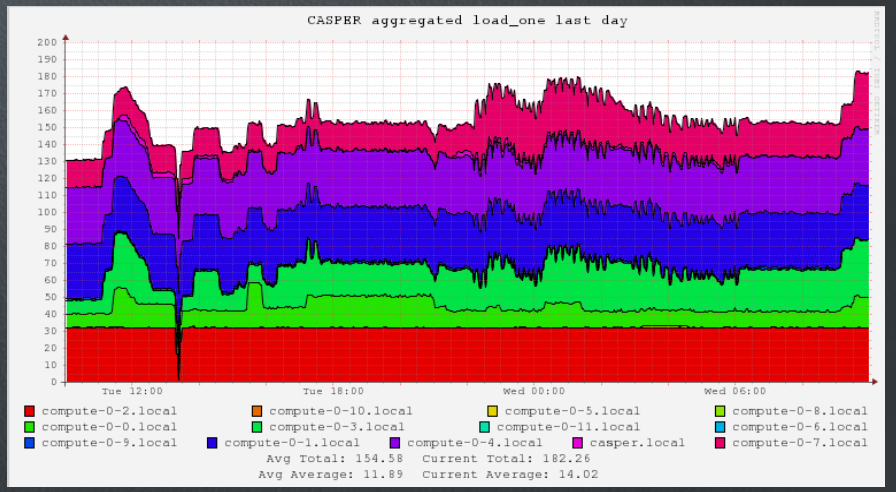
HDFS runs fast on IPoIB

USER EXPERIENCE

1. Modeling & development on your workstation
2. SFTP file transfer
3. Secure Shell connection
4. Job definition, in-queue submission and monitoring on CASPER
5. email notification for jobs
6. On-line statistics (ganglia)

```
[~]gestione@casper$ qstat -u \*
job-ID  prior   name       user         state submit/start at     queue                          slots ja-task-ID
-----------------------------------------------------------------------------------------------------------------
 220095 0.60500 zeolite-LJ small-group   r     03/06/2014 14:04:04 public.q@compute-0-2.local      128
 220343 0.50500 Test       cantini      r     03/06/2014 14:48:19 all.q@compute-0-4.local           1
 220936 0.52783 QLOGIN     giulia       r     03/10/2014 17:45:14 all.q@compute-0-9.local          30
 220959 0.51681 QLOGIN     srg          r     03/11/2014 11:01:34 public.q@compute-0-3.local       16
 220979 0.51681 QLOGIN     srg          r     03/11/2014 15:17:57 public.q@compute-0-0.local       16
 221035 0.50579 me_pos0    francescob   r     03/11/2014 23:29:52 bioeda.q@compute-0-7.local        2
 221039 0.50579 me_pos0    francescob   r     03/11/2014 23:59:07 bioeda.q@compute-0-7.local        2
 221071 0.50500 b_sa_all   francescob   r     03/12/2014 09:06:07 public.q@compute-0-0.local        1
 221072 0.50500 b_sa_all   francescob   r     03/12/2014 09:06:07 public.q@compute-0-0.local        1
 221073 0.50500 b_sa_all   francescob   r     03/12/2014 09:07:07 public.q@compute-0-3.local        1
 221074 0.50500 b_sa_all   francescob   r     03/12/2014 09:07:07 public.q@compute-0-3.local        1
 221075 0.50500 b_sa_all   francescob   r     03/12/2014 09:07:07 public.q@compute-0-3.local        1
 221076 0.50500 b_sa_all   francescob   r     03/12/2014 09:07:07 public.q@compute-0-3.local        1
 221077 0.50500 b_sa_all   francescob   r     03/12/2014 09:07:07 public.q@compute-0-3.local        1
```

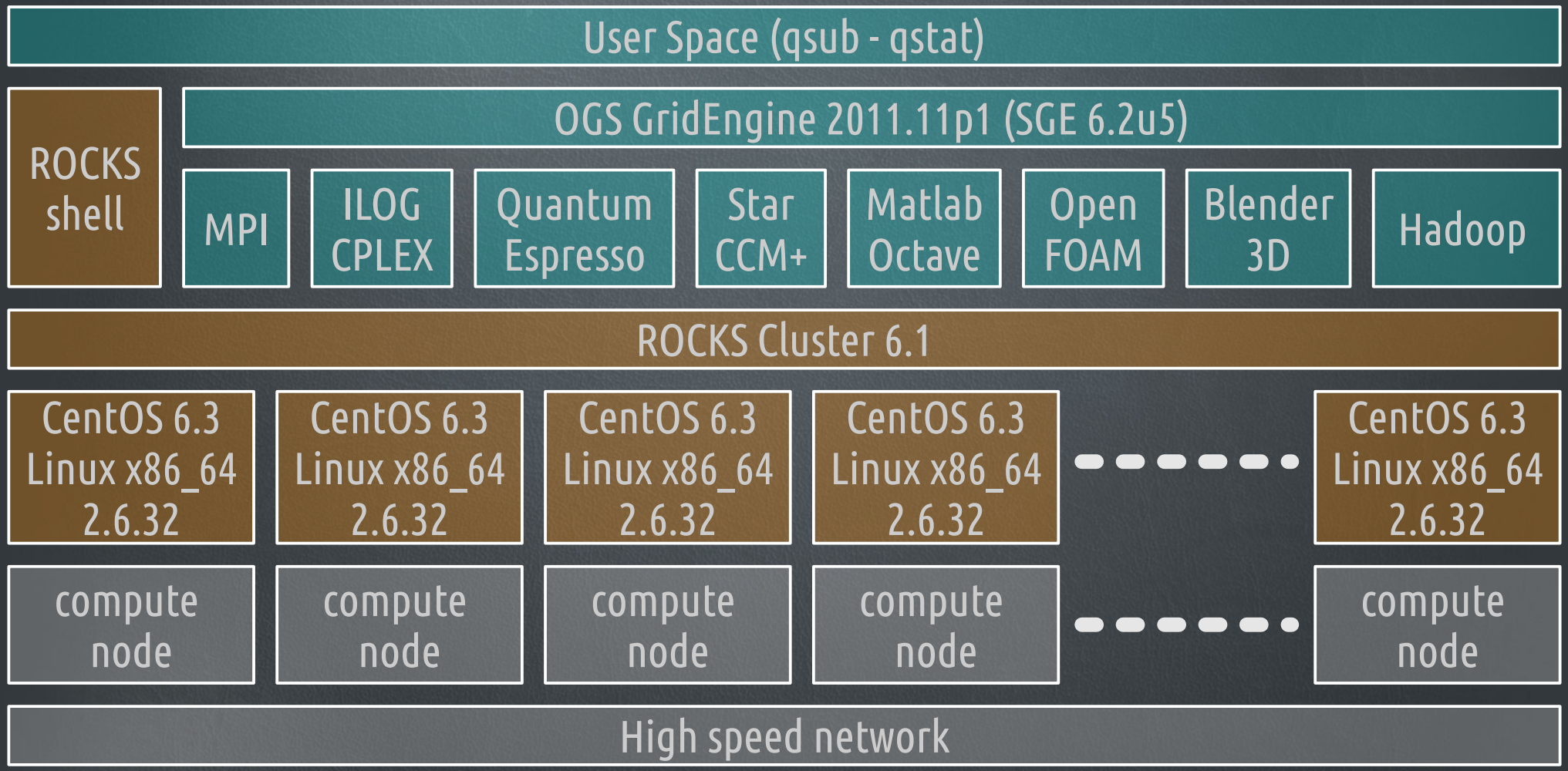




(some) ALREADY AVAILABLE SOFTWARE

1. All ROCKS 6.1 included software - Programs and libraries for technical and high performance computing in bondle with the ROCKS cluster distribution
2. AMD Open64 - AMD compilers optimized for Opteron Bulldozer architecture
3. AMD Core Math Library - Opteron optimized BLAS, LAPACK, FFTs and random number generators
4. Blender 3D 2.59 and 2.69 - 3D creation for everyone, free to use for any purpose (cit.)
5. IBM ILOG CPLEX 12.6.0.0- High-performance mathematical programming solver
6. GotoBLAS2 & ATLAS - very fast multi-threaded BLAS implementations
7. Gromacs 4.6.5 and 5.0-beta2 - A versatile tool for molecular dynamics
8. OpenFOAM 2.1.1 and 2.2.x - Free, open source Computational Fluid Dynamics software package, plus 3rd party SW
9. Quantum Espresso 5.0.3 - Integrated suite of Open-Source computer codes for electronic-structure calculations and materials modeling at the nano-scale
10. Star CCM+ 8.06.005 - Simulation of Turbulent flow in Arbitrary Regions and Computational Continuum Mechanics
11. Octave 3.6.4 and 3.8.0 - High-level interpreted language, primarily intended for numerical computations
12. Matlab 8.1 R2013a - The language of technical computing (cit.)

CASPER SOFTWARE ARCHITECTURE



GridEngine IS YOUR FRIEND

1. Remember that clusters are batch systems with different queues mastered by some scheduler

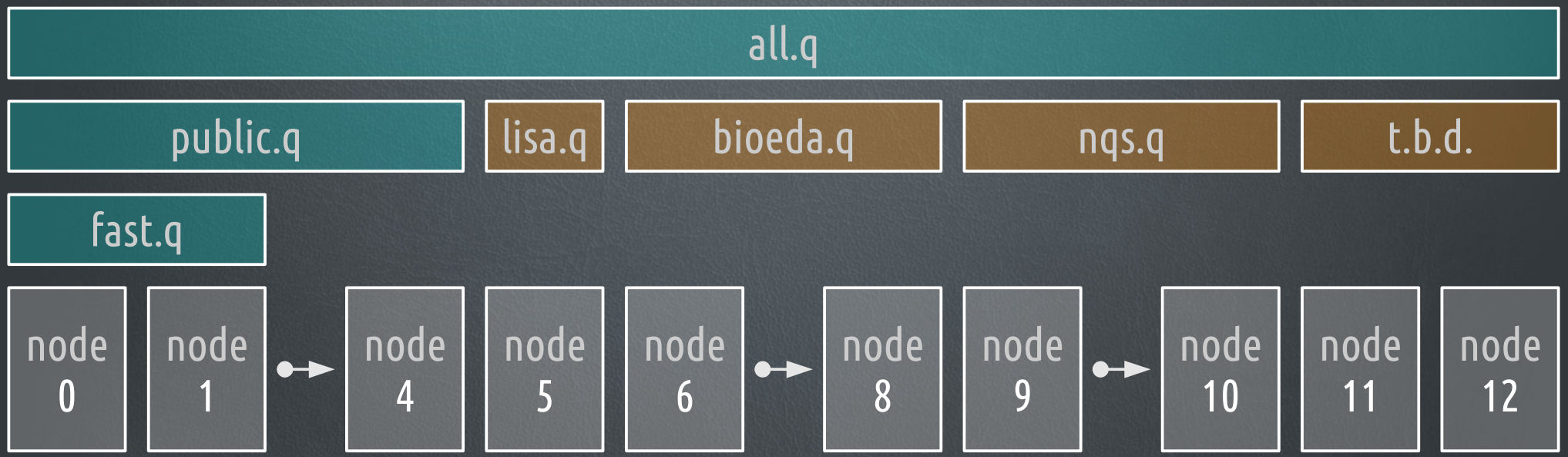
2. Know your software needs and behavior

3. Know the cluster and watch the situation before submitting

4. Choose the queue that better fits your expectations and ask what you need to the scheduler

QUEUES AVAILABLE ON CASPER

Job submission via qsub command



Fine tuning in GE: load thresholds, queue subordination, core over subscription, job priorities

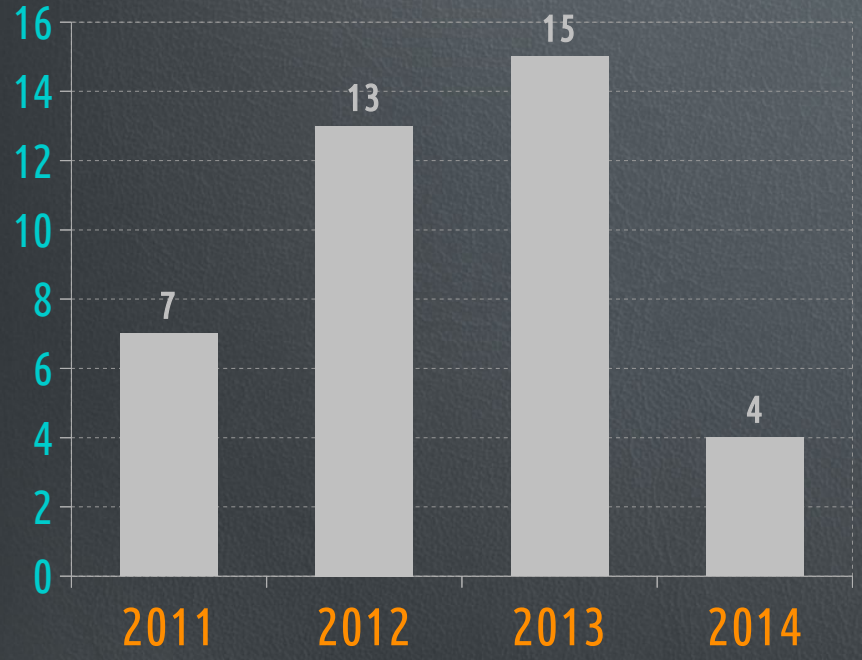
HOW ARE THINGS GOING ON?

SOME FACTS ABOUT HPC@POLITO

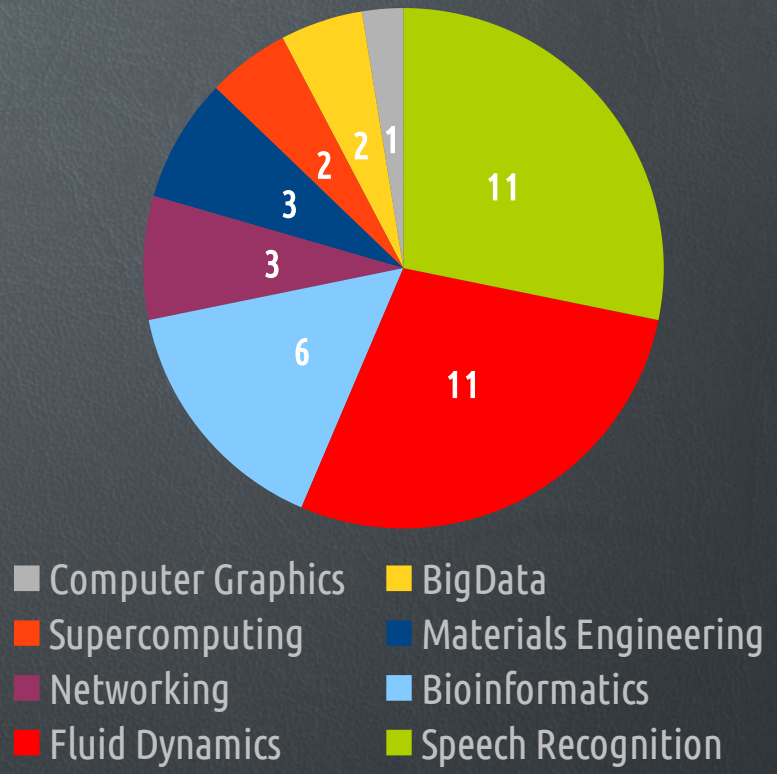
1. 47 research projects have been hosted and more are coming
2. 17 projects are still active on CASPER
3. 26 research groups from 7 departments were involved
4. 39 papers were published
5. CASPER received 3 hardware upgrades in 5 years
6. HPC@POLITO was accepted as a member of the “HPC Advisory Council” in 2013
7. HPC@POLITO has no budget for human resources

PAPERS

Papers published per year



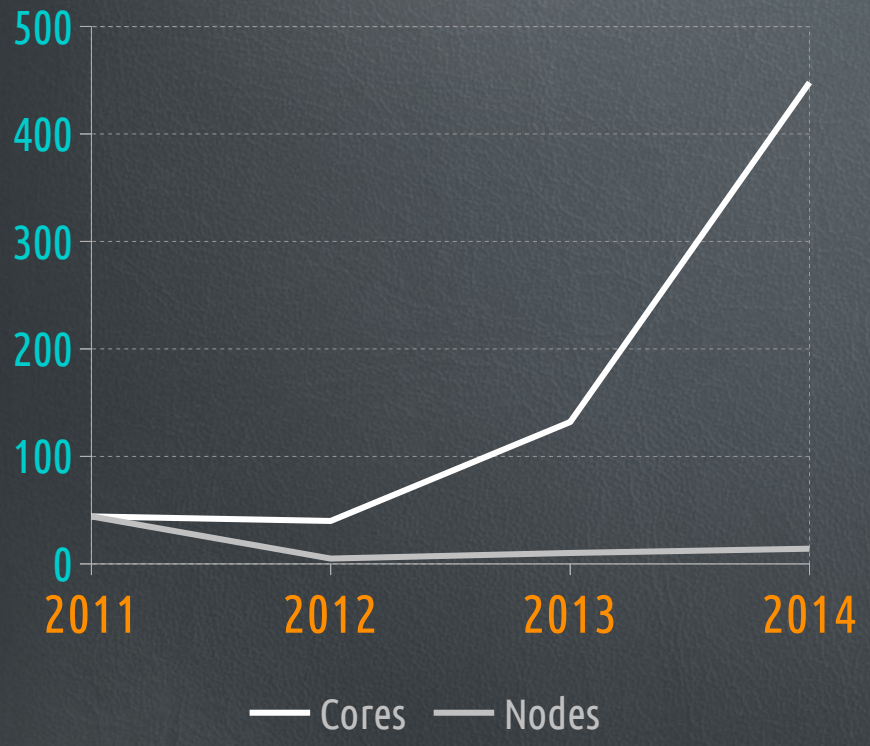
Papers published per area



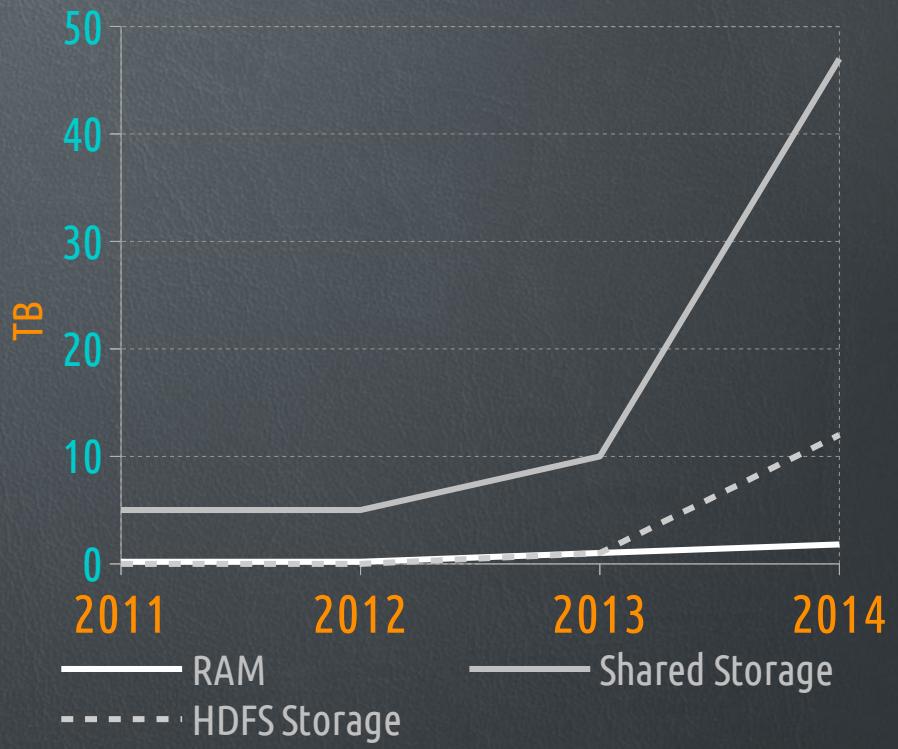
Data source: *hpc.polito.it*

HARDWARE EVOLUTION

Total Cores and Nodes



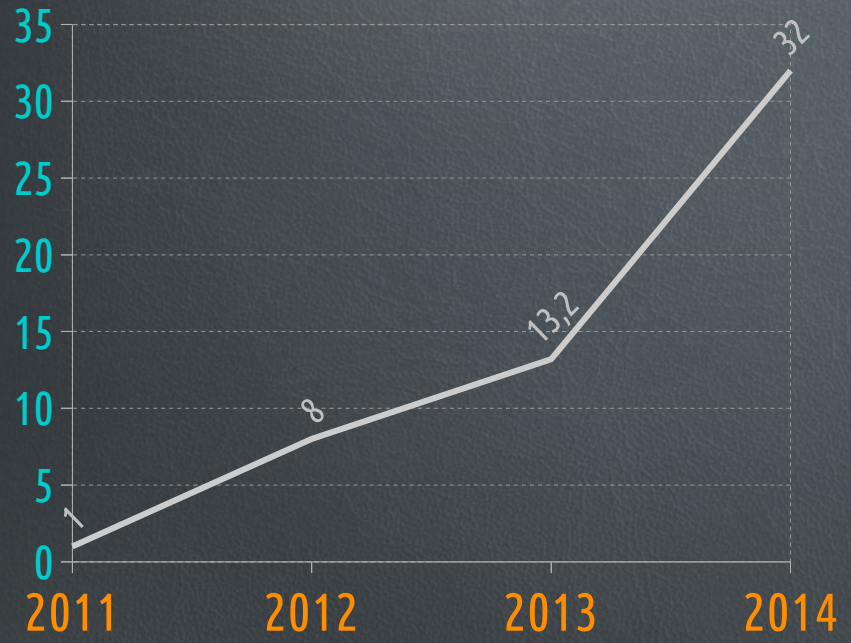
Total Memory and Storage



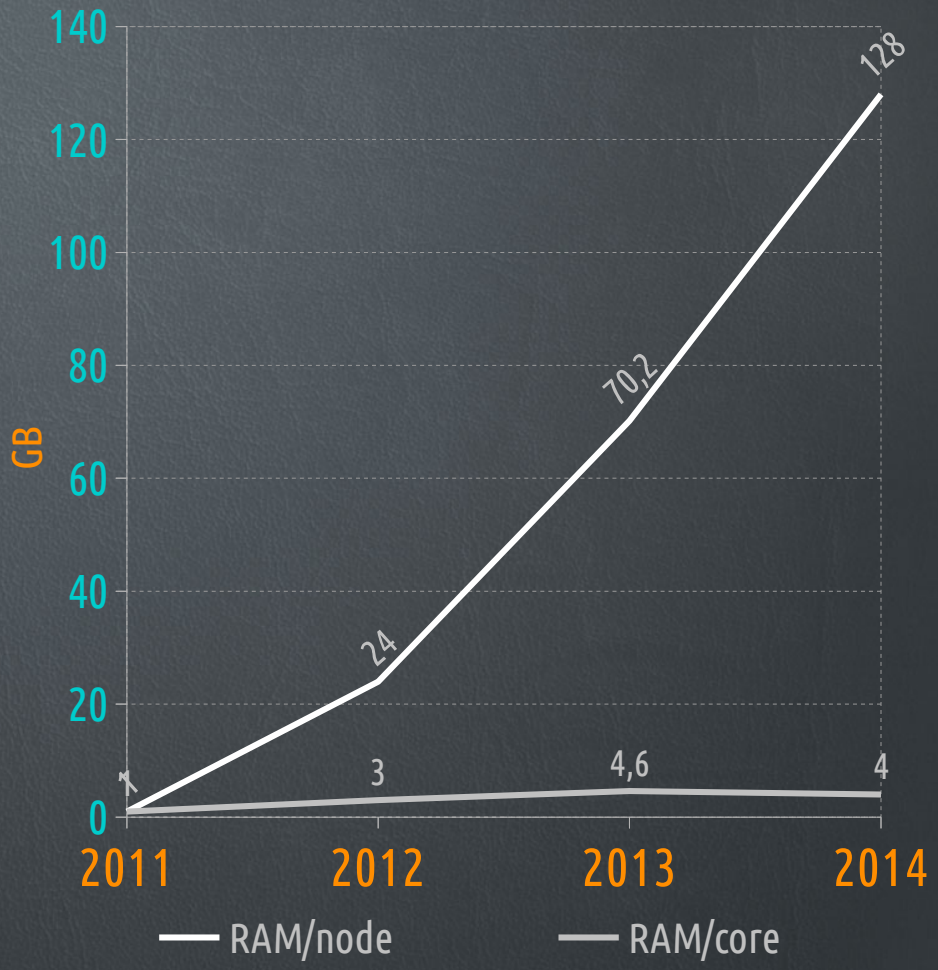
Data source: *hpc.polito.it* and P. Margara, N. Nepote, E. Piccolo, C. G. Demartini, P. Montuschi, *"Thinking BigData: Motivation, Results and a Few Recipes for a Balanced Growth of HPC in Academia"*, *2013*

HARDWARE EVOLUTION

Cores per Node ratio



Memory ratios



Data source: *hpc.polito.it* and P. Margara, N. Nepote, E. Piccolo, C. G. Demartini, P. Montuschi, *"Thinking BigData: Motivation, Results and a Few Recipes for a Balanced Growth of HPC in Academia" , 2013*

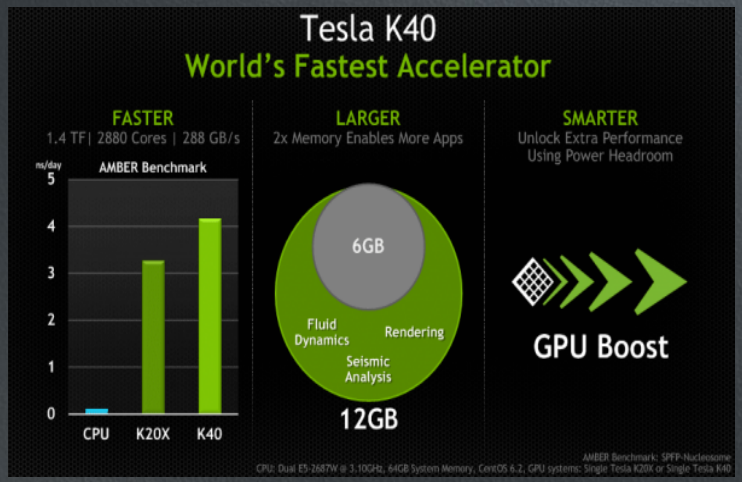
CASPER IS GOING TO BE...

+π∫ ÷x-√ HPC@polito

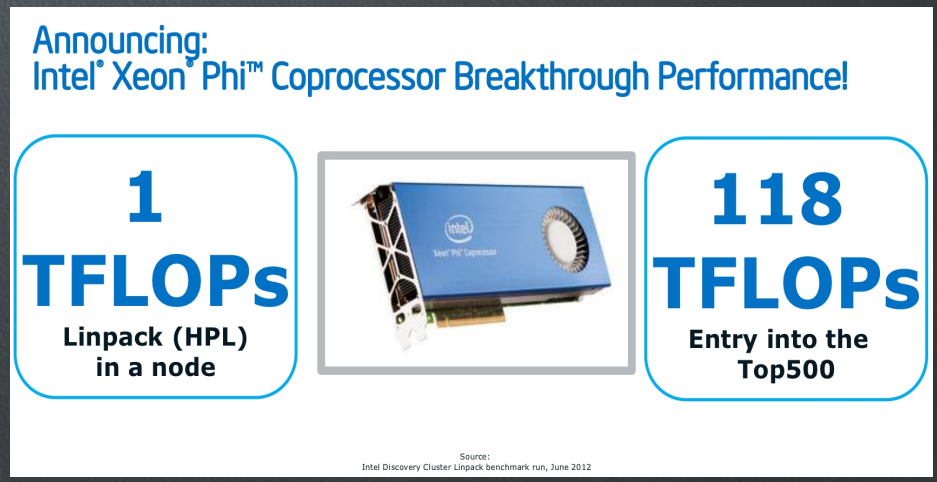
HARDWARE EVOLUTION

1. Goal: more rough power → 12-core Intel Xeon E5 2695 v2 dual socket
2. Goal: faster storage → Lustre or pNFS over InfiniBand or Fiber Channel
3. Goal: reach 10 TFLOPS → InfiniBand FDR 56 Gb/s on fiber, ~ 20 nodes
4. Goal: use less power → multi-node highly-packed chassis (like HP s6500)

Why not general purpose GPUs?



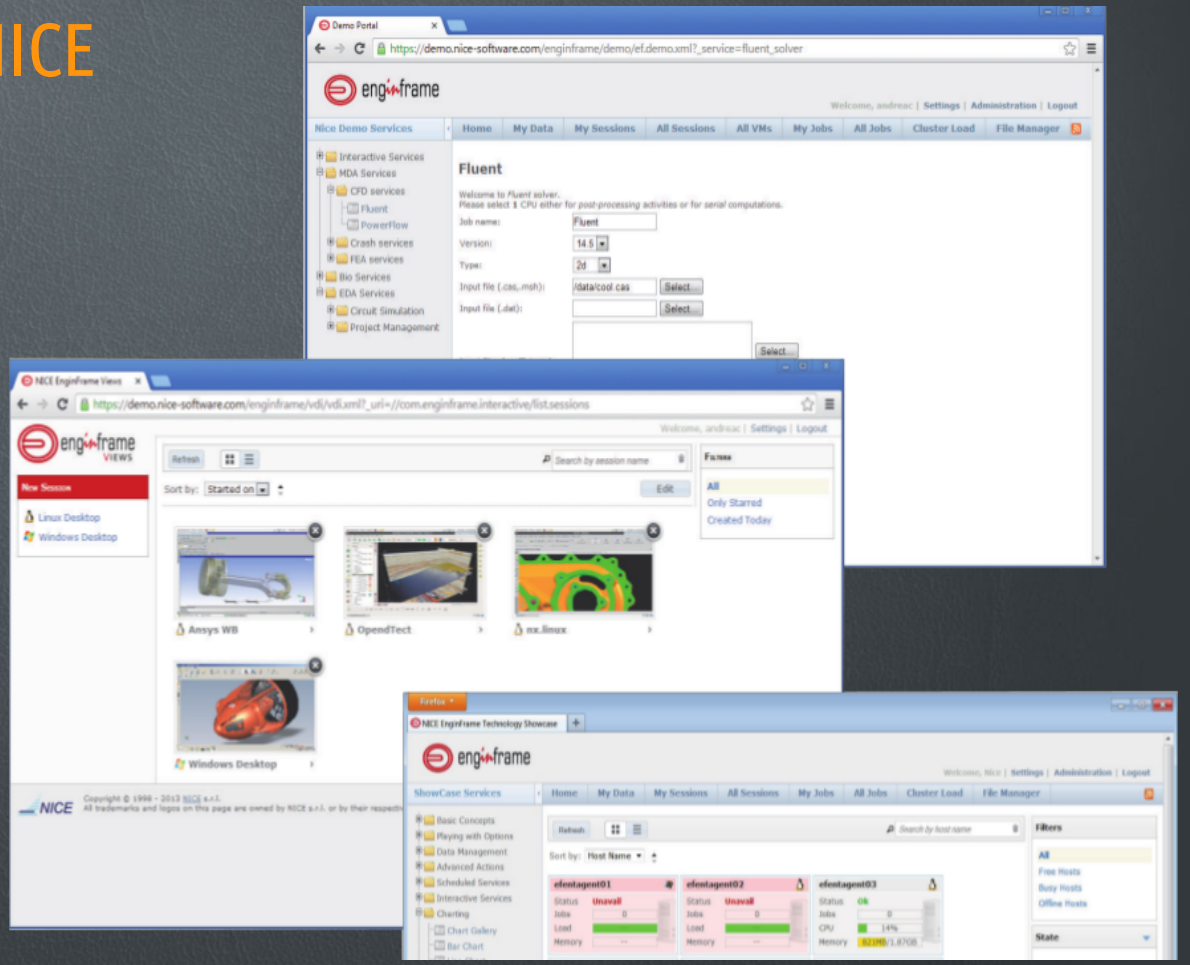
Why not HPC coprocessors?



BE MORE USER FRIENDLY

1. Deploy already available NICE EnginFrame portal:

2. Develop a tile rendering system for Blender, but tightly integrated with GridEngine...



TO WHOM IT MAY CONCERN

P. Garza, P. Margara, N. Nepote, L. Grimaudo, E. Piccolo "Hadoop on a Low-Budget General Purpose HPC Cluster in Academia" Springler's Advances in Intelligent Systems and Computing - New Trends in Databases and Information Systems, vol. 241, 2014, pp. 187-192

P. Margara, N. Nepote, E. Piccolo, C. G. Demartini, P. Montuschi "Thinking BigData: Motivation, Results and a Few Recipes for a Balanced Growth of HPC in Academia" 50th AICA Conference on Digital Frontiers 2013, Salerno

N. Nepote, E. Piccolo, C. G. Demartini, P. Montuschi "Why and How Using HPC in University Teaching? A Case Study at PoliTo" 27th DIDAMATICA Conference on Technologies and Methods for future Teaching 2013, Pisa, pp. 2019-2028

F. Della Croce, N. Nepote, E. Piccolo "A Terascale Cost-Effective Open Solution for Academic Computing: Early Experience of the Dauin HPC Initiative" 49th AICA Conference on Smart Tech and Smart Innovation 2011, Turin, pp. 56-65

A bibliography about HPC@POLITO

+π∫ ÷x-√ HPC@polito

THANKS!

hpc.polito.it
hpc.dauin@polito.it
@hpc_polito

+π∫ ÷x-√ HPC@polito