FHPC/P1.2 course:

Lecture 1: Introduction to HPC (second part)

Stefano Cozzini

CNR/IOM and eXact-lab srl

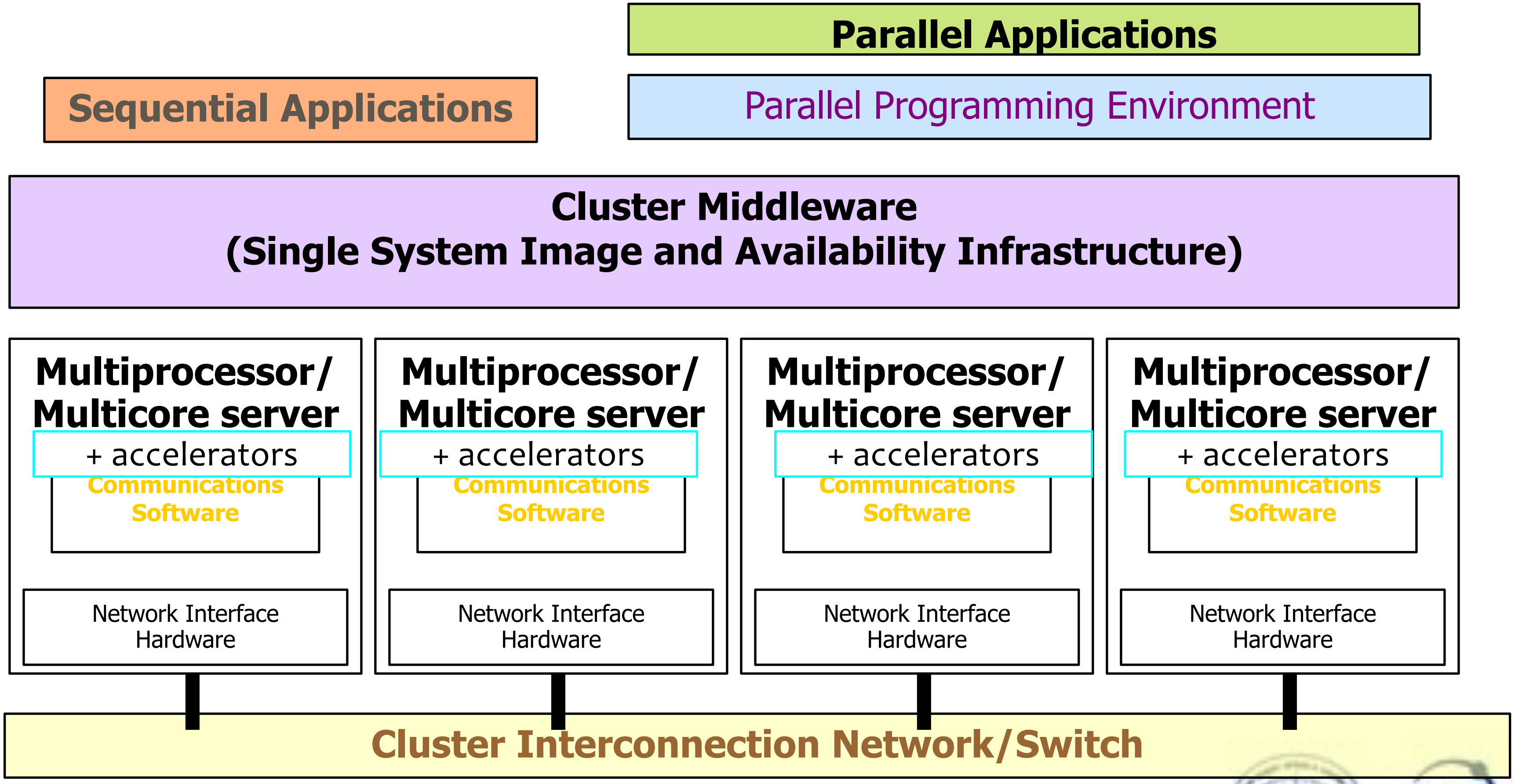
Agenda of the previous lecture

- Prologue: why and where HPC ?
- What is HPC ?
  - Definitions&metrics
- Component of a HPC infrastructure
- HPC Concepts
  - Parallel programming paradigms
  - Evolution of paradigms
  - Ahmdal law / Gustafson law
  - Strong/weak scalability
- HOMEWORK&LABS

Agenda: for today

- What is HPC infrastructure ?
  - Supercomputers & HPC Cluster
  - CPUs and Accelerators
  - Network/storage
- Software stack for HPC
  - Middleware: queue systems
  - Libraries/ Compiler/ performance Tools

HPC Cluster Computer Architecture



Parallel Applications

Sequential Applications

Parallel Programming Environment

Cluster Middleware
(Single System Image and Availability Infrastructure)

Multiprocessor/ Multicore server
+ accelerators
Communications Software
Network Interface Hardware

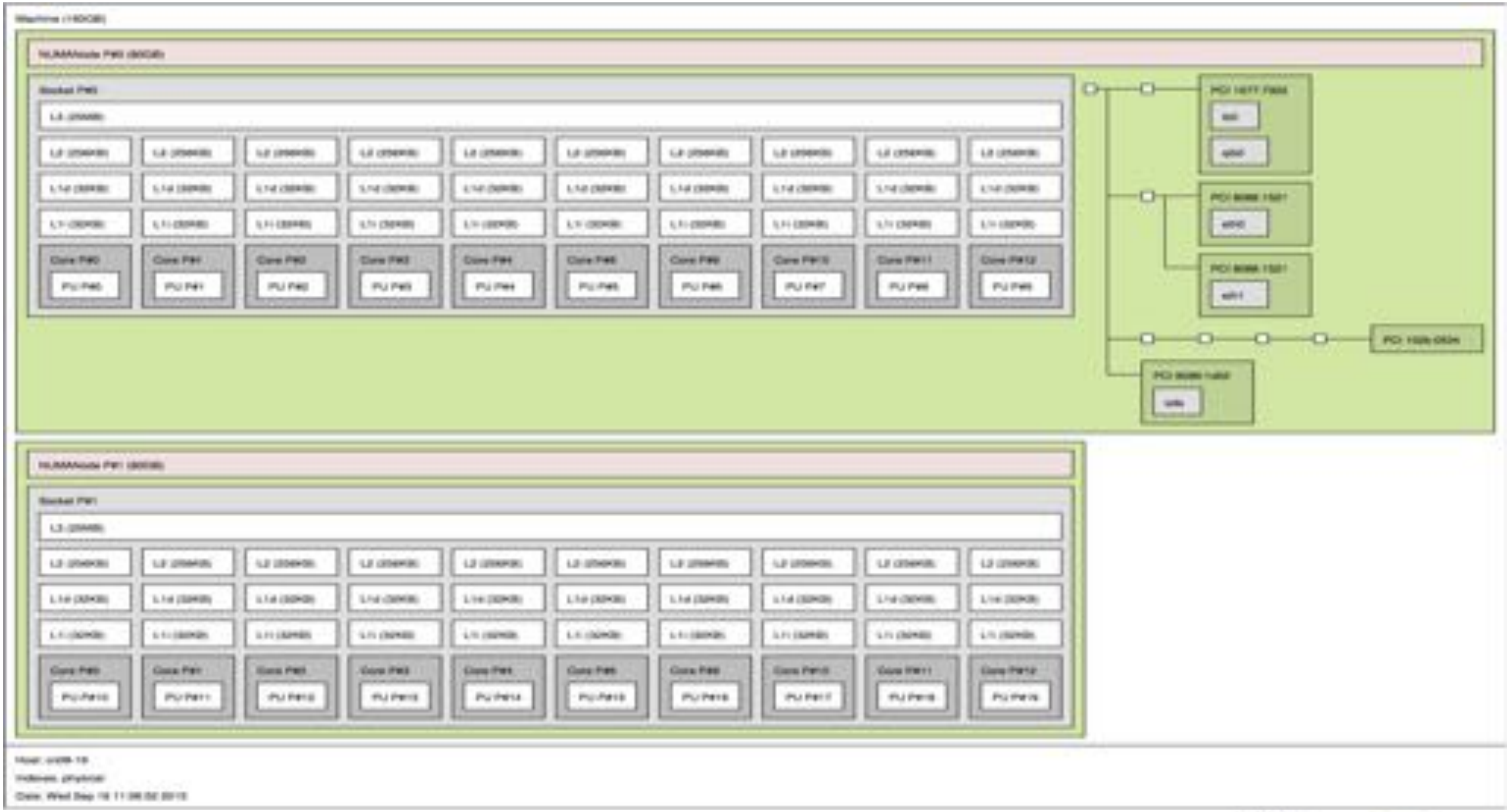
Multiprocessor/ Multicore server
+ accelerators
Communications Software
Network Interface Hardware

Multiprocessor/ Multicore server
+ accelerators
Communications Software
Network Interface Hardware

Multiprocessor/ Multicore server
+ accelerators
Communications Software
Network Interface Hardware

Cluster Interconnection Network/Switch

A modern node picture (Ullysses node)



About HPC jargon

- Multiprocessor = server with more than 1 CPU
- Multicore= a CPU with more than 1 core

Processor = CPU =socket

BUT SOMETIME:

Processor= core

a process for each processor (i.e. each core)

Elements of the clusters

- Several computers, nodes, often in special cases for easy mounting in a rack
- One or more networks (interconnects) to hook the nodes together
- Storage facilities.

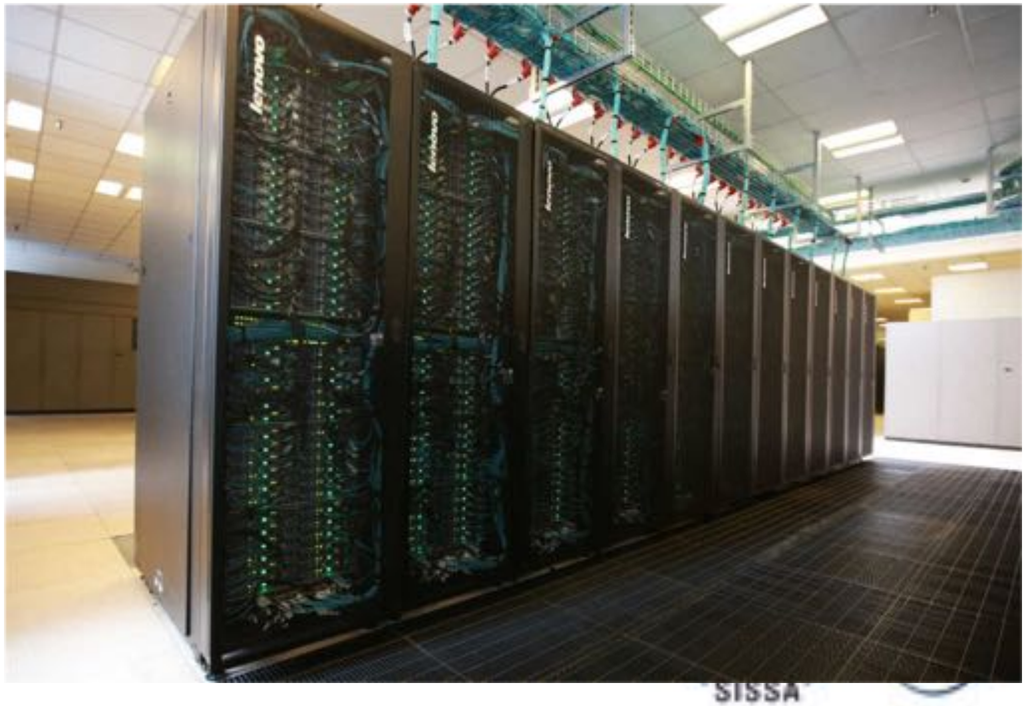
A node of modern HPC cluster

1U box

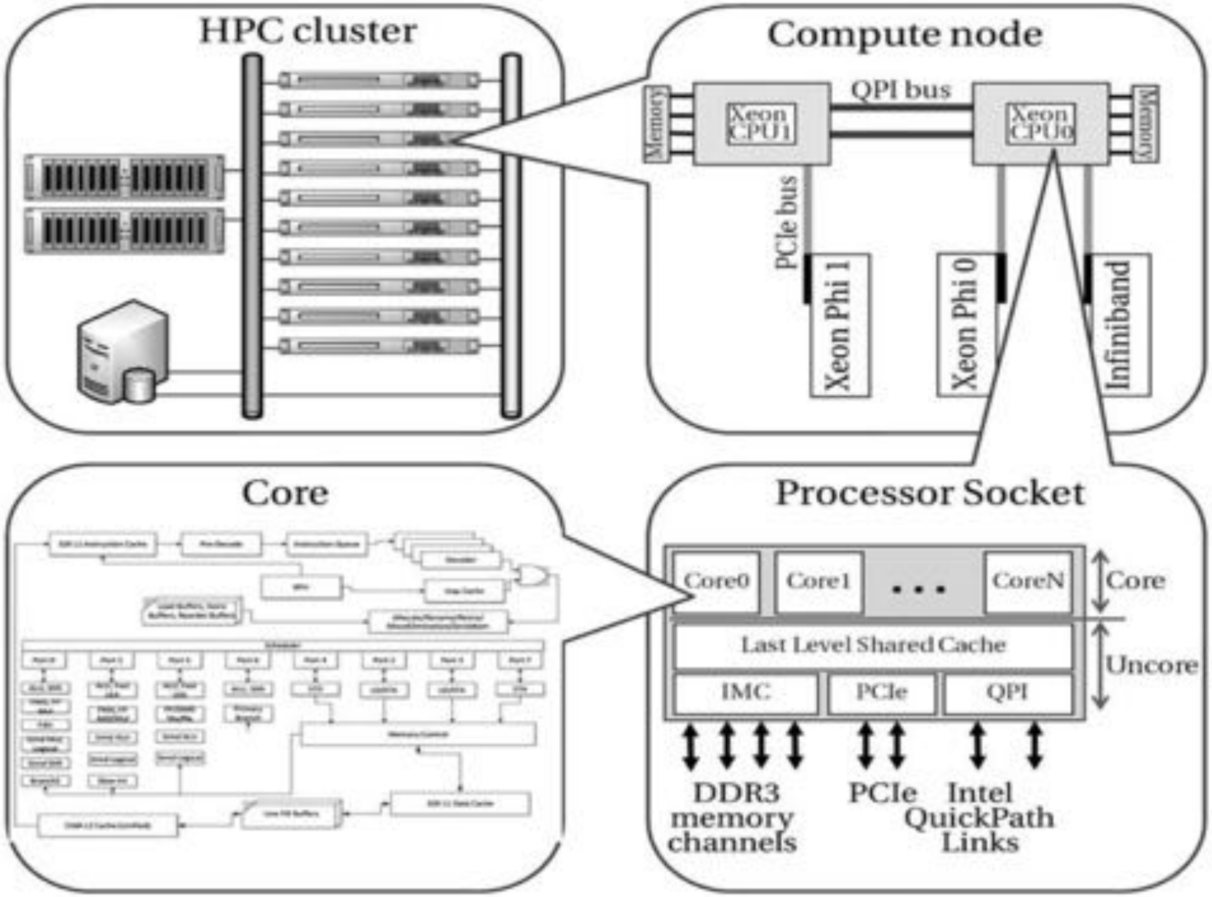
1 or 2 accelerators

A shared memory machine (SMP or NUMA)

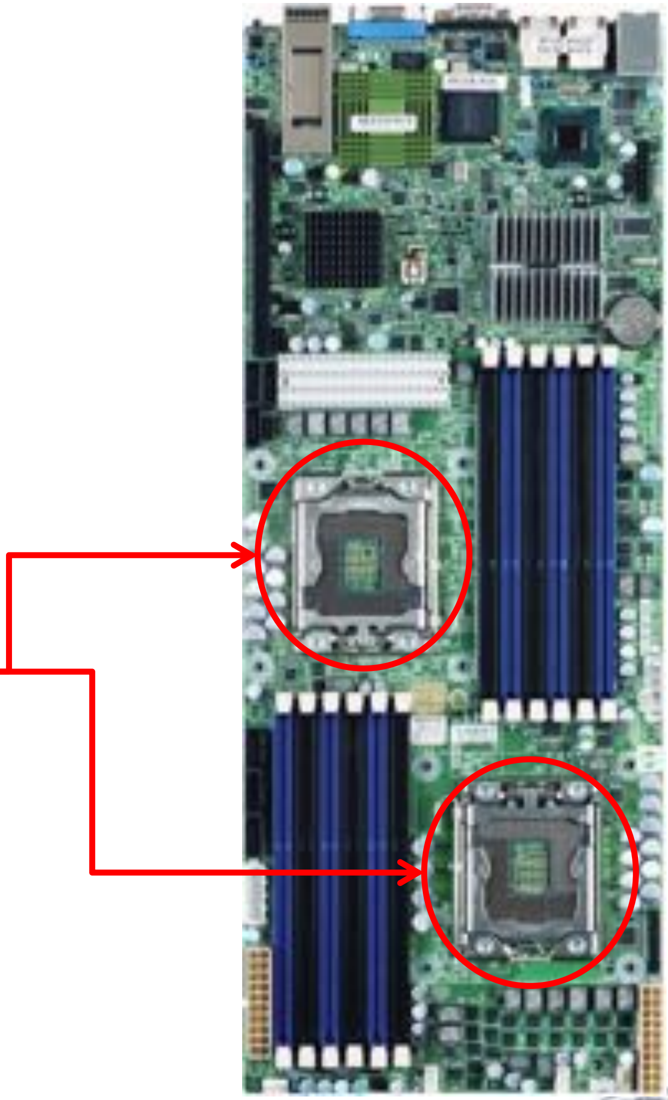
Some times also blades.. In racks



Building blocks of a cluster

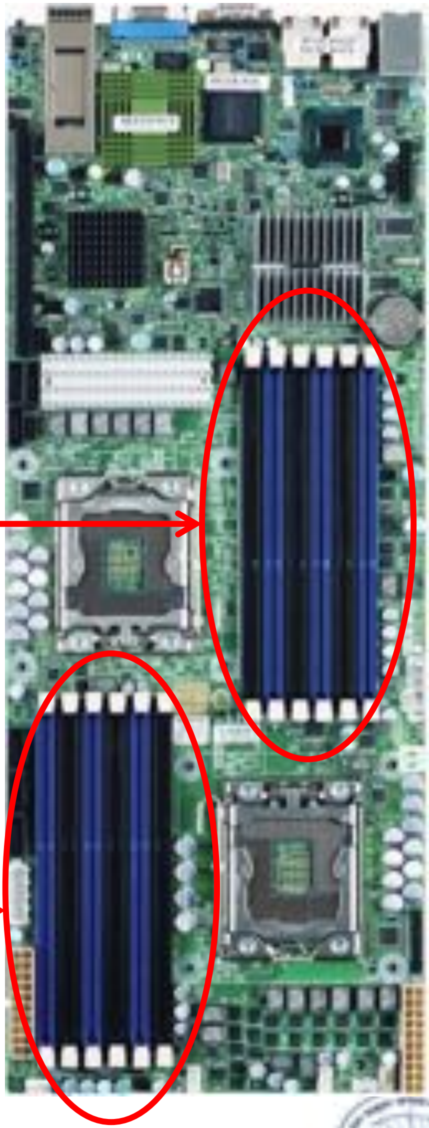


The motherboard components



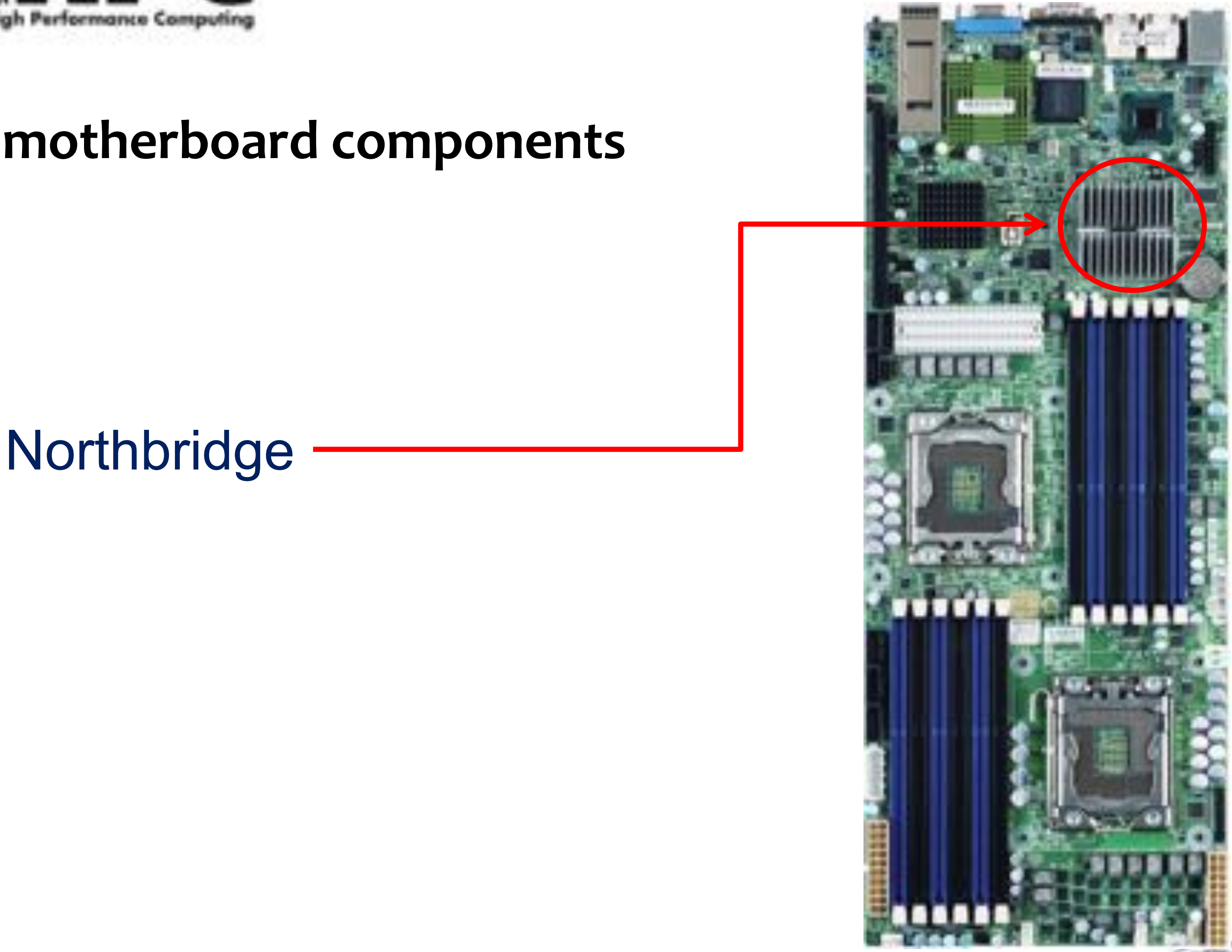
Dual socket CPU

The motherboard components



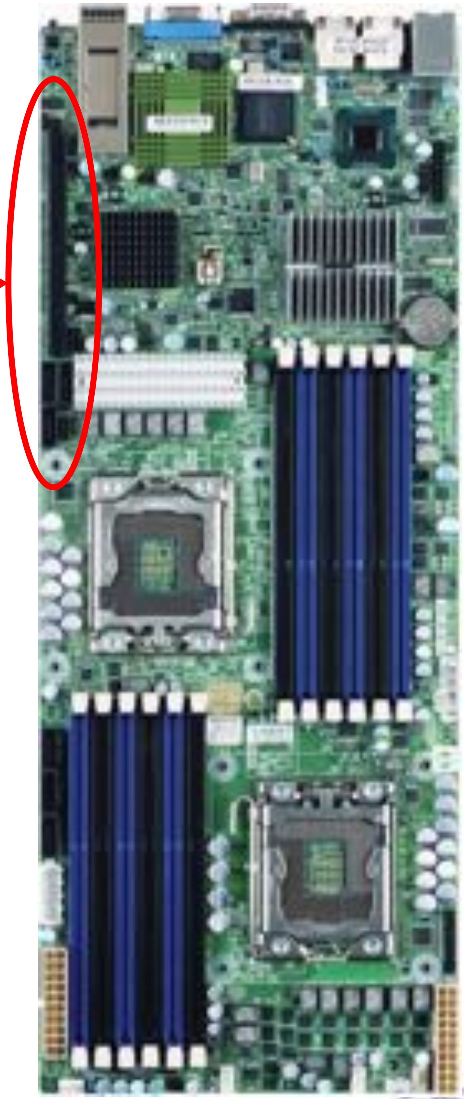
DDR3 Ram 12 slots

The motherboard components



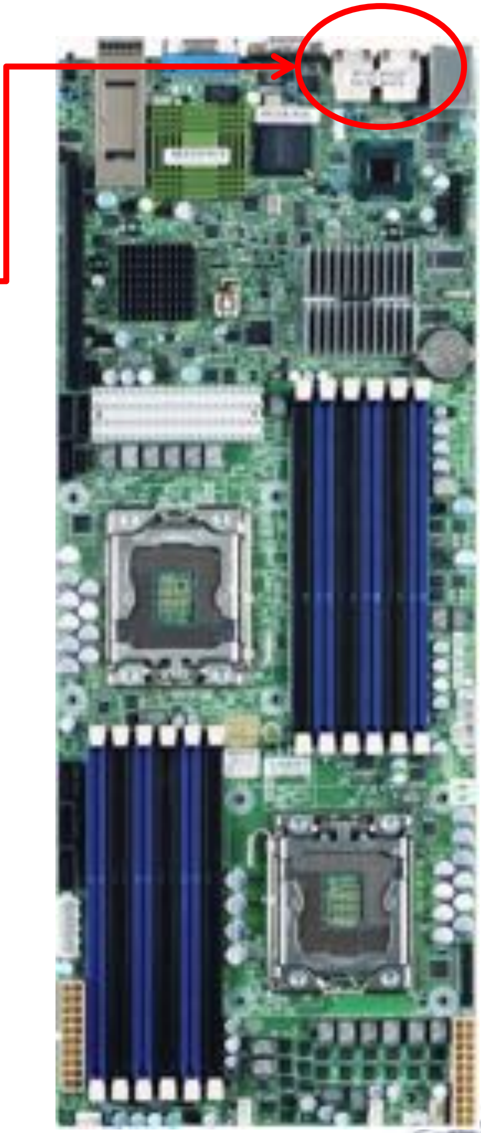
Northbridge

The motherboard components



PCI 8x 1 slot

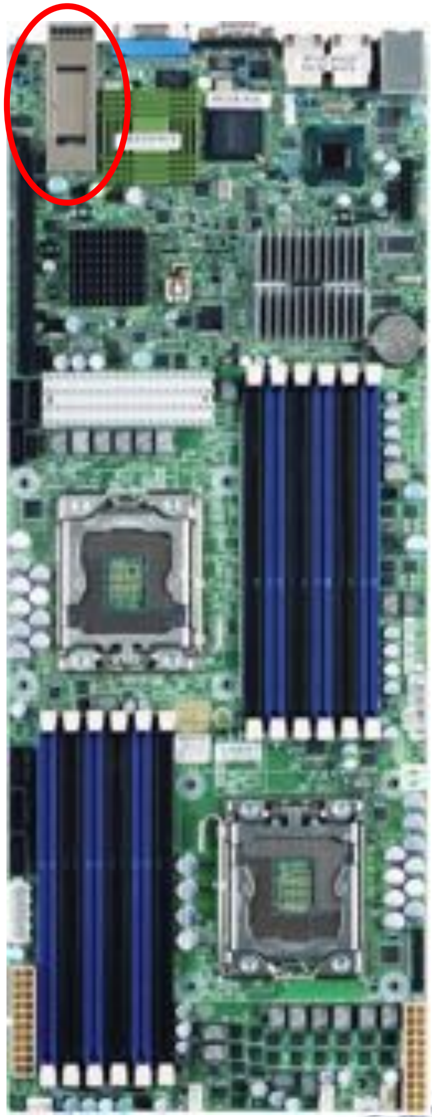
The motherboard components



Gb Ethernet dual port

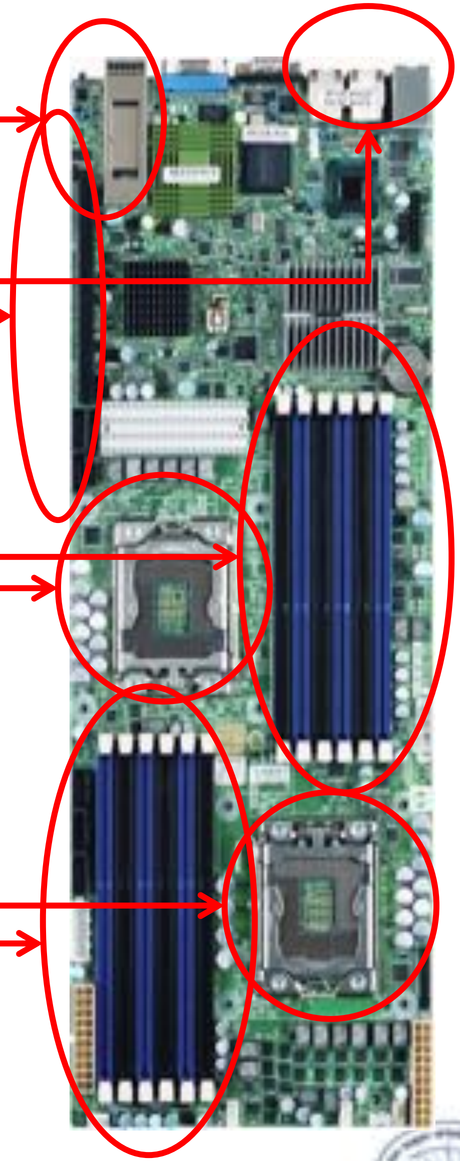
The motherboard components

Infiniband 4x QDR



Motherboard components

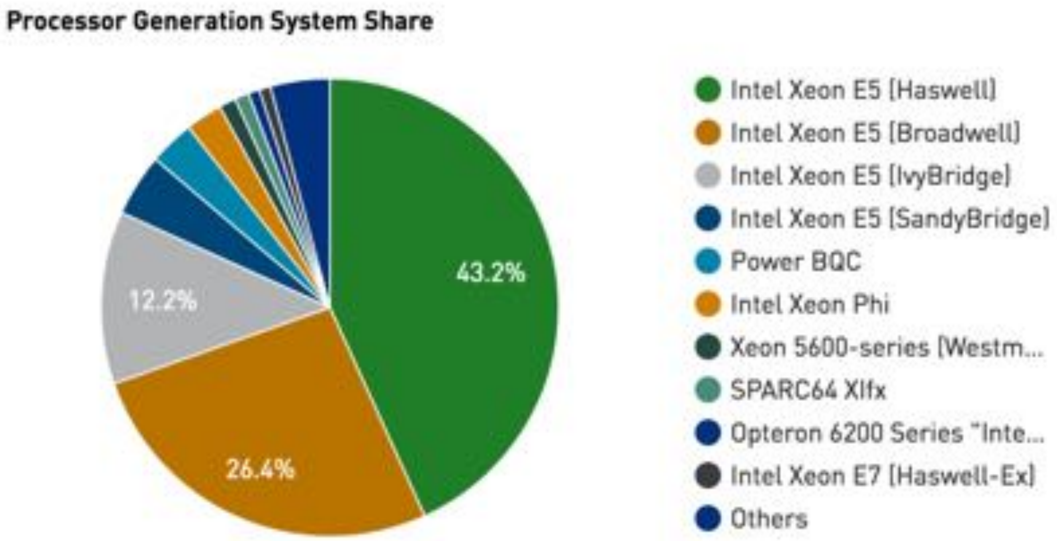
- Gb Ethernet dual port
- PCI 8x 1 slot
- Infiniband 4x QDR
- Dual socket CPU
- DDR3 Ram 12 slots

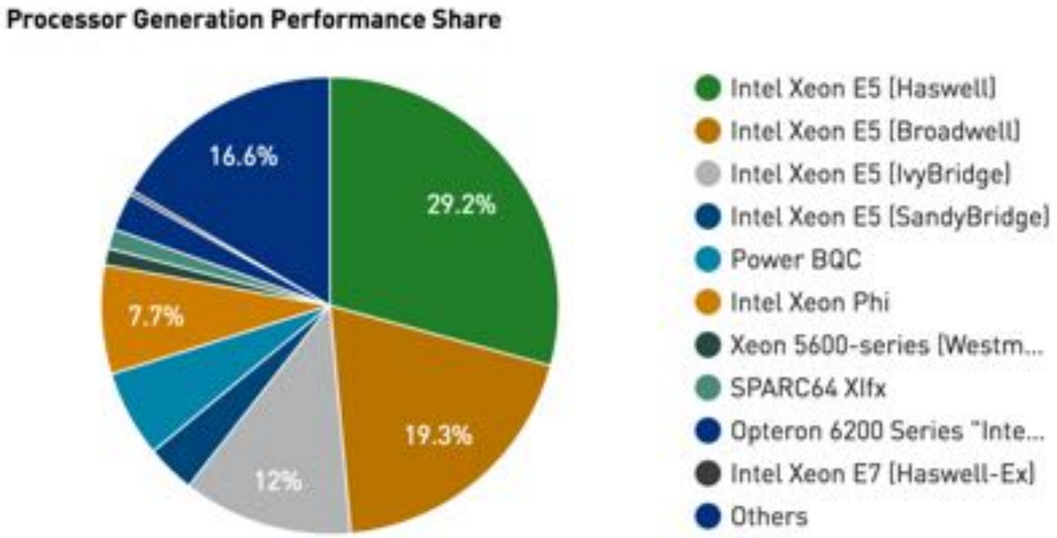


Slides from P.Altoe E4company

MHPC
Master in High Performance Computing

Which CPUs on TOP500 system ?

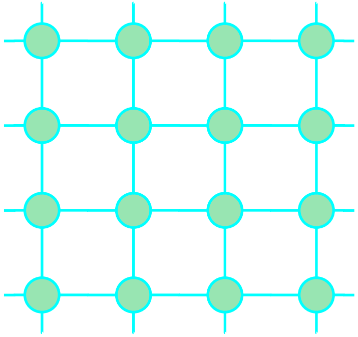

Processor Generation System Share

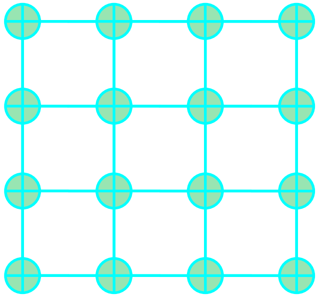

Processor Generation Performance Share

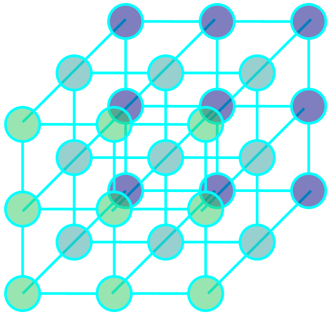
About network for cluster

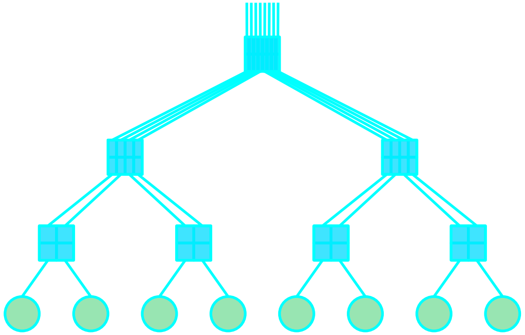
- The performance of the network cannot be ignored
  - Latency: Initialization time before data can be sent
  - Per-link Peak Bandwidth: Maximum data transmission rate (varies with packet size)
  - Topology: how the network is done.

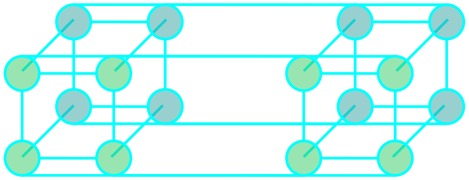
Clustering topology


2D Mesh


2D Torus


3D Mesh


FAT TREE


Hypercube (4-cube)

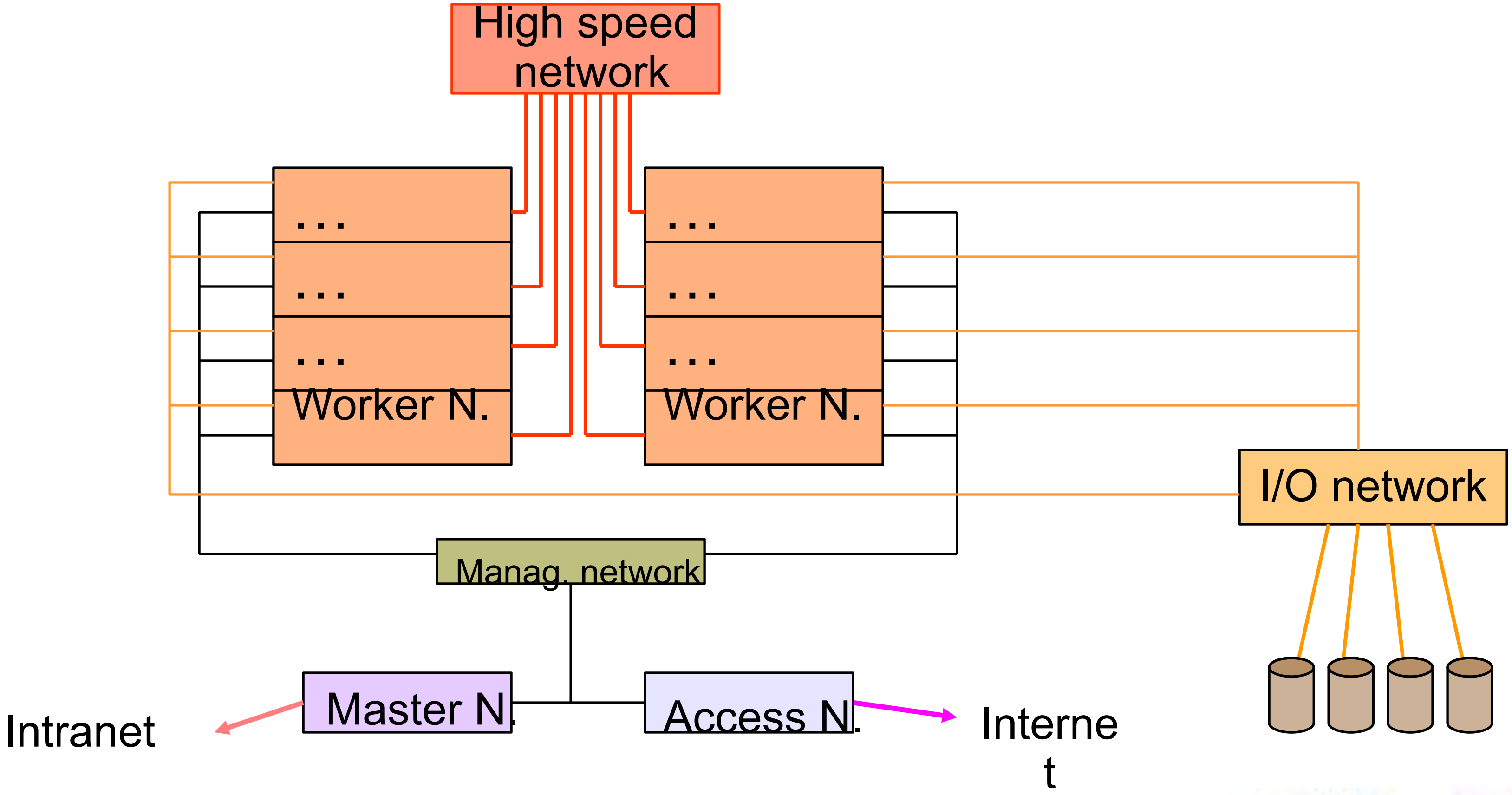
Latency&bandwidth

NETWORK	Latency	Bandwidth (GB/sec)
Gigabit	70-40	~ 0.125
10G	<5	~1.250
Infiniband 4DDR	~1.5/1.9	~ 3.2
Infiniband FDR	<1.0	~ 5

What is the UNIT OF MEASURE OF LATENCY ?

Microseconds: 3 order of magnitude larger than unit of measure of FP operations

HPC cluster logical structure..

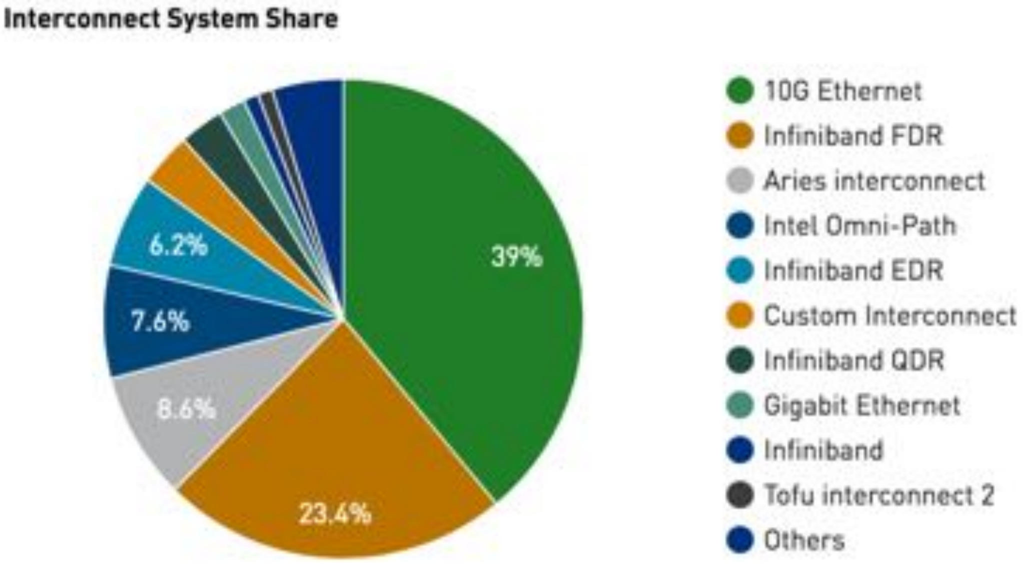


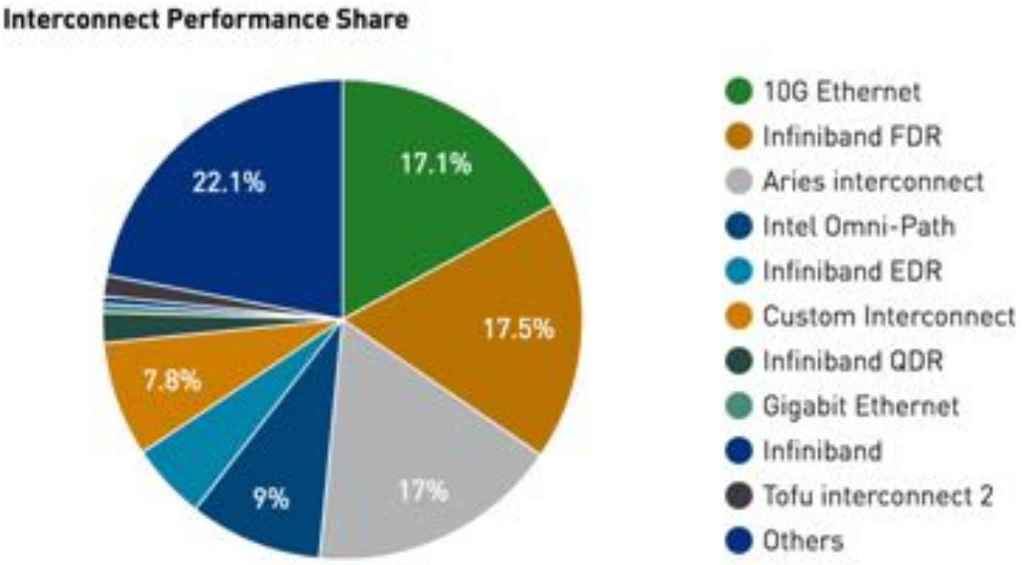
HPC cluster: 3 kind of network

- HIGH SPEED NETWORK
  - parallel computation
  - low latency /high bandwidth
  - Usual choices: Infiniband…
- I/O NETWORK
  - I/O requests (NFS and/or parallel FS)
  - latency not fundamental/ good bandwidth
  - GIGABIT is ok
- Management network
  - management traffic
  - any standard network

MHPC
Master in High Performance Computing

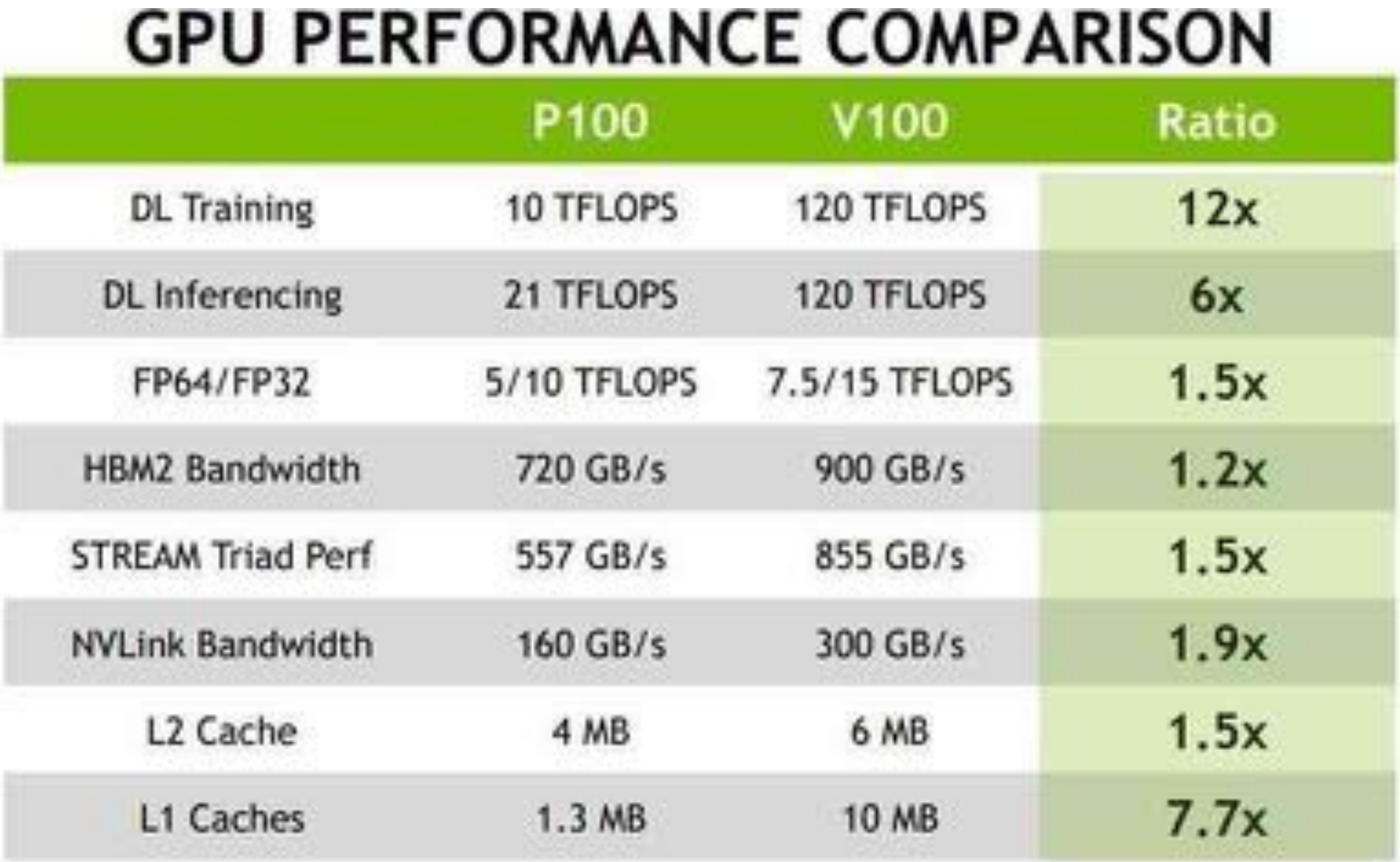
Network in Top500





Accelerators: GPU

- Co-processors or accelerators have been around for a while
- Big burst in its adoption in HPC when Nvidia released CUDA (2006).
- GPGPUs or simply GPUs work in a different way to conventional CPUs. Emphasis on stream processing.
- Acceleration can be significant but depends on application.
- Nvidia market leader with astonishing performance..

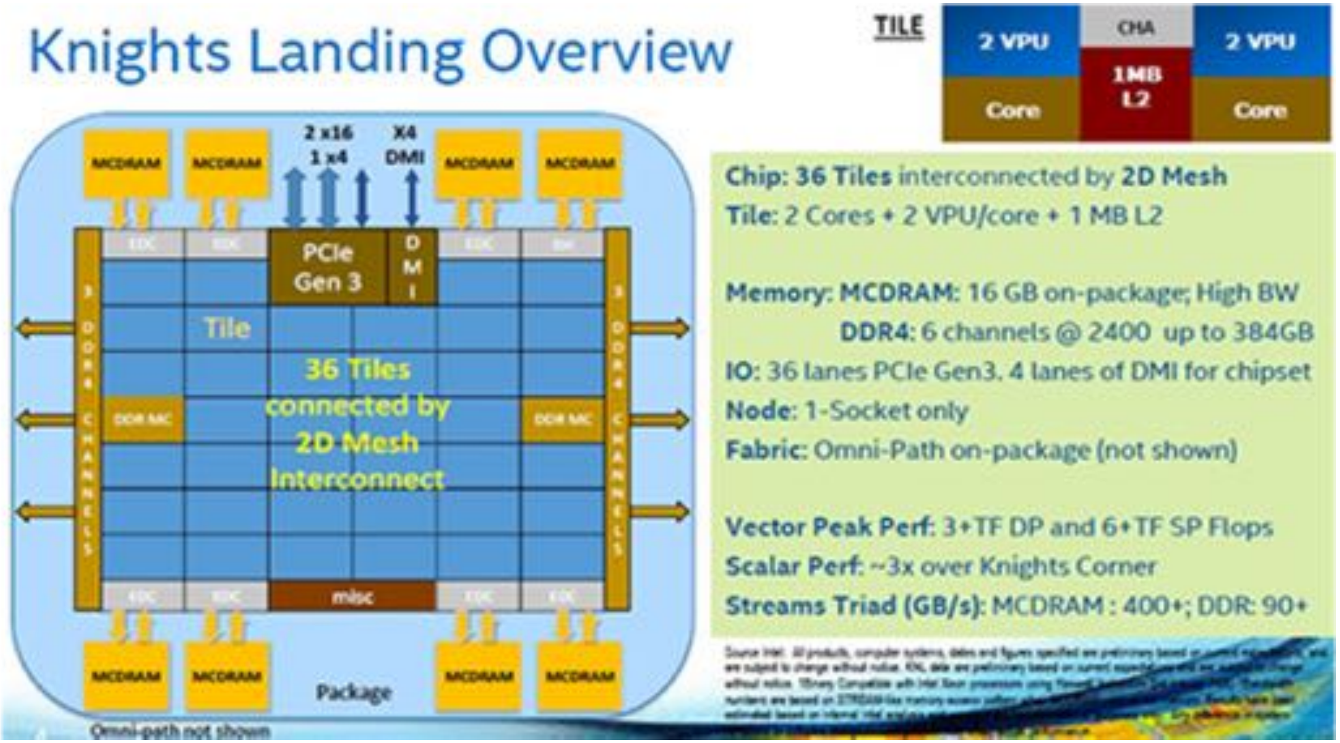


GPU PERFORMANCE COMPARISON

	P100	V100	Ratio
DL Training	10 TFLOPS	120 TFLOPS	12x
DL Inferencing	21 TFLOPS	120 TFLOPS	6x
FP64/FP32	5/10 TFLOPS	7.5/15 TFLOPS	1.5x
HBM2 Bandwidth	720 GB/s	900 GB/s	1.2x
STREAM Triad Perf	557 GB/s	855 GB/s	1.5x
NVLink Bandwidth	160 GB/s	300 GB/s	1.9x
L2 Cache	4 MB	6 MB	1.5x
L1 Caches	1.3 MB	10 MB	7.7x

Accelerators: Intel PHI (MIC)

- Also an accelerator but more similar to a conventional multicore CPU.
- • Cores connected in a ring topology.
- • No need to write CUDA or OpenCL as Intel compilers will compile Fortran or C code for the MIC.



Knights Landing Overview

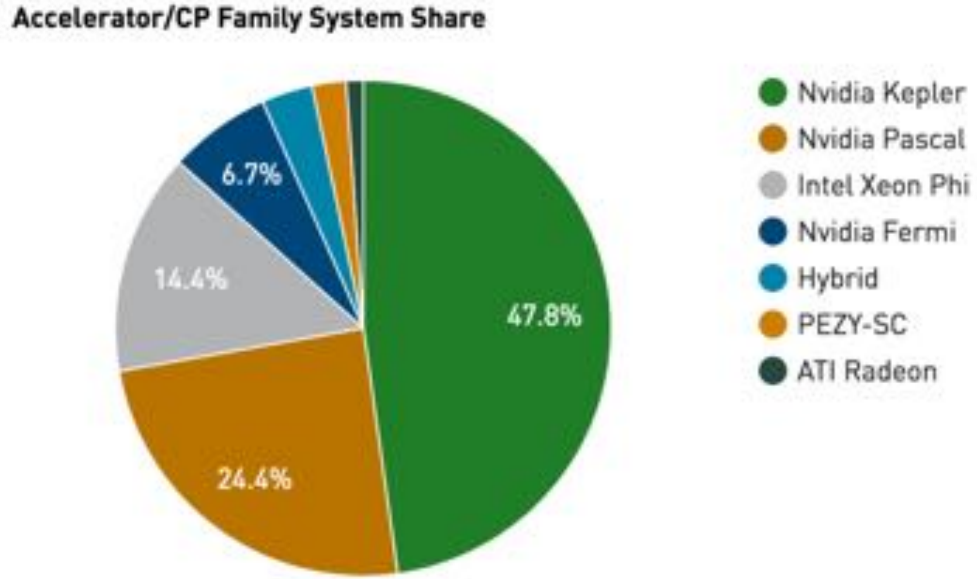
Chip: 36 Tiles interconnected by **2D Mesh**
Tile: 2 Cores + 2 VPU/core + 1 MB L2

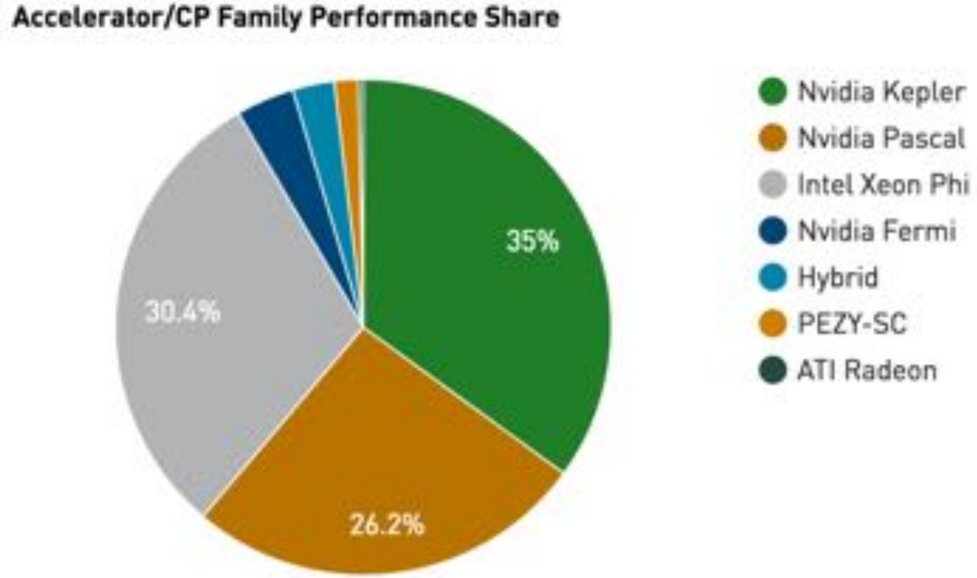
Memory: MCDRAM: 16 GB on-package; High BW
DDR4: 6 channels @ 2400 up to 384GB
IO: 36 lanes PCIe Gen3. 4 lanes of DMI for chipset
Node: 1-Socket only
Fabric: Omni-Path on-package (not shown)

Vector Peak Perf: 3+TF DP and 6+TF SP Flops
Scalar Perf: ~3x over Knights Corner
Streams Triad (GB/s): MCDRAM : 400+; DDR: 90+

Source Intel. All products, computer systems, dates and figures specified are preliminary based on [illegible]

Accelerators in Top500


Accelerator/CP Family System Share


Accelerator/CP Family Performance Share

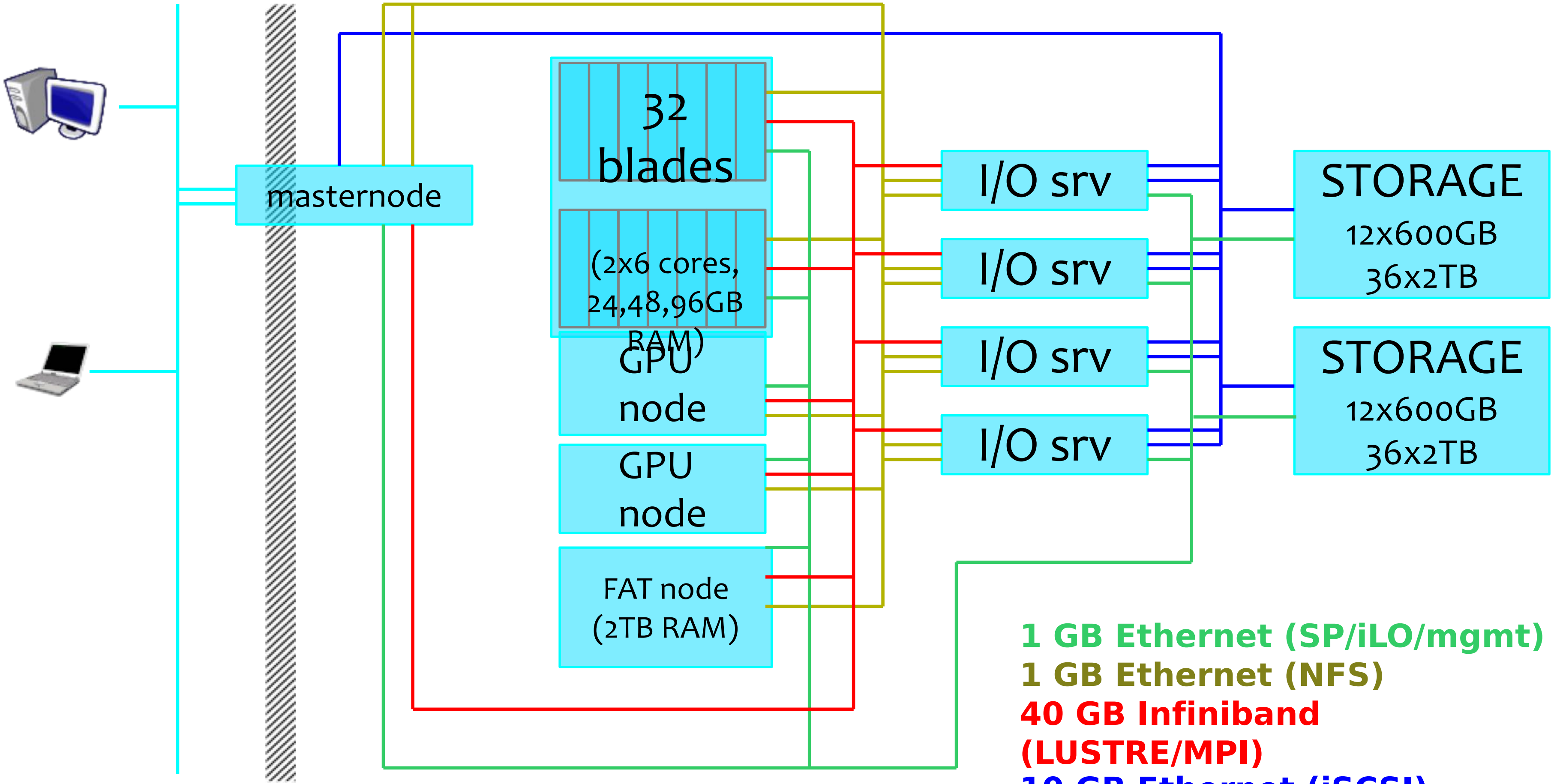
Why accelerators ?

- GPUs and MIC mandatory in HPC because of high performance and efficiency (i.e. Flops/watt).
- they are mainly to be attached to host CPUs via the PCIe bus (a standard PC-like connection).
- Both device families have limitations:
  - low device memory
  - slow transfer rate via PCIe link
  - difficulty in programming (particularly CUDA).
  - speedup is highly application and data dependent.
- New model are standalone models (e.g Knight’s Landing) and/or and with faster connections (Nvlink).

Last but not least: Storage

- High Speed Storage is required for HPC
  - Parallel Filesystem is mandatory:
    - Lustre/GPFS/BeeGFS etc..
- Hierarchical storage is also a solution:
  - Hierarchical storage management (HSM) is a data storage technique, which automatically moves data between high-cost and low-cost storage media.
    - First layer: SSD
    - Second layer : parallel FS
    - Third layer: SAN
    - Fourth layer: Tapes

Cluster example



1 GB Ethernet (SP/iLO/mgmt)
1 GB Ethernet (NFS)
40 GB Infiniband (LUSTRE/MPI)
10 GB Ethernet (iSCSI)
1 GB (LAN)

System stack of a supercomputing [from ref 1]

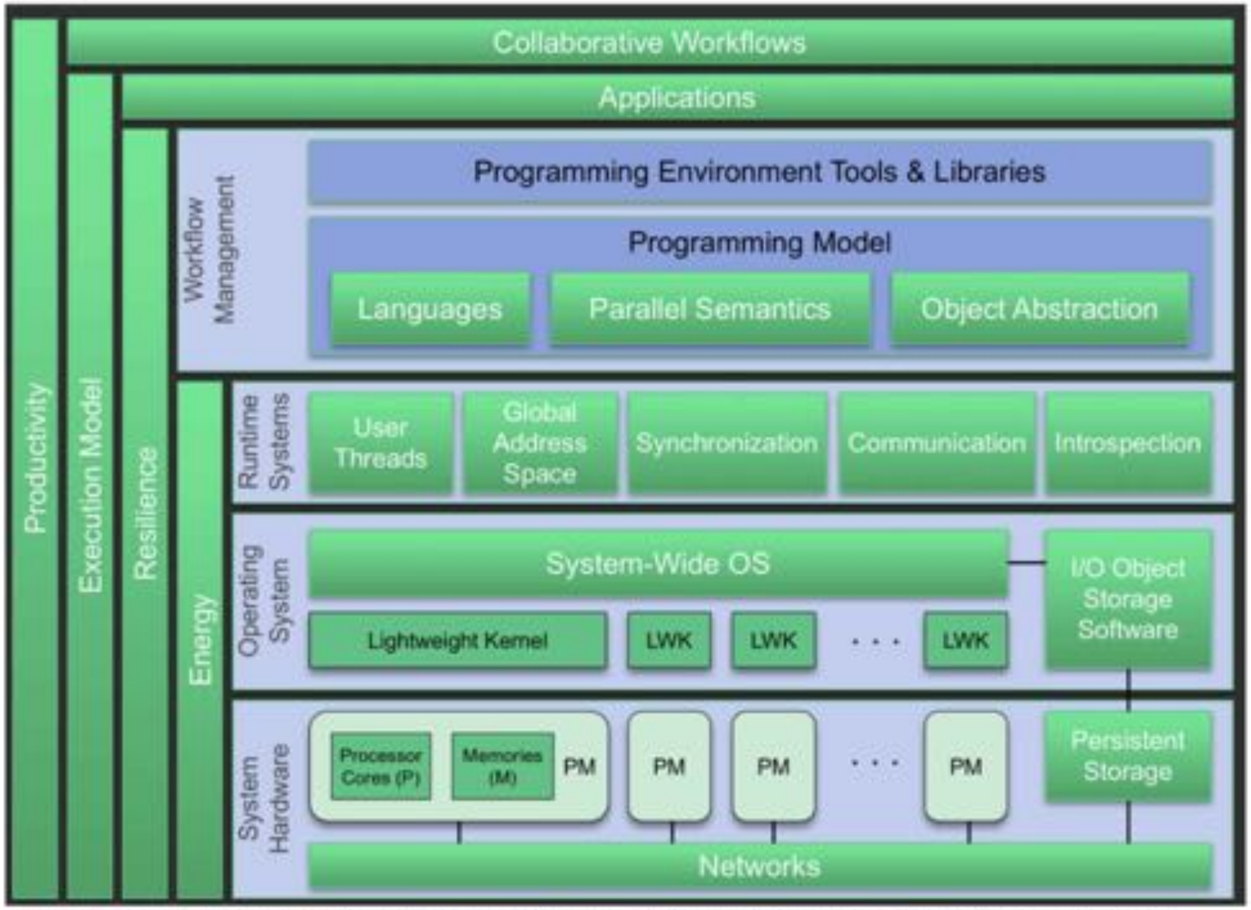
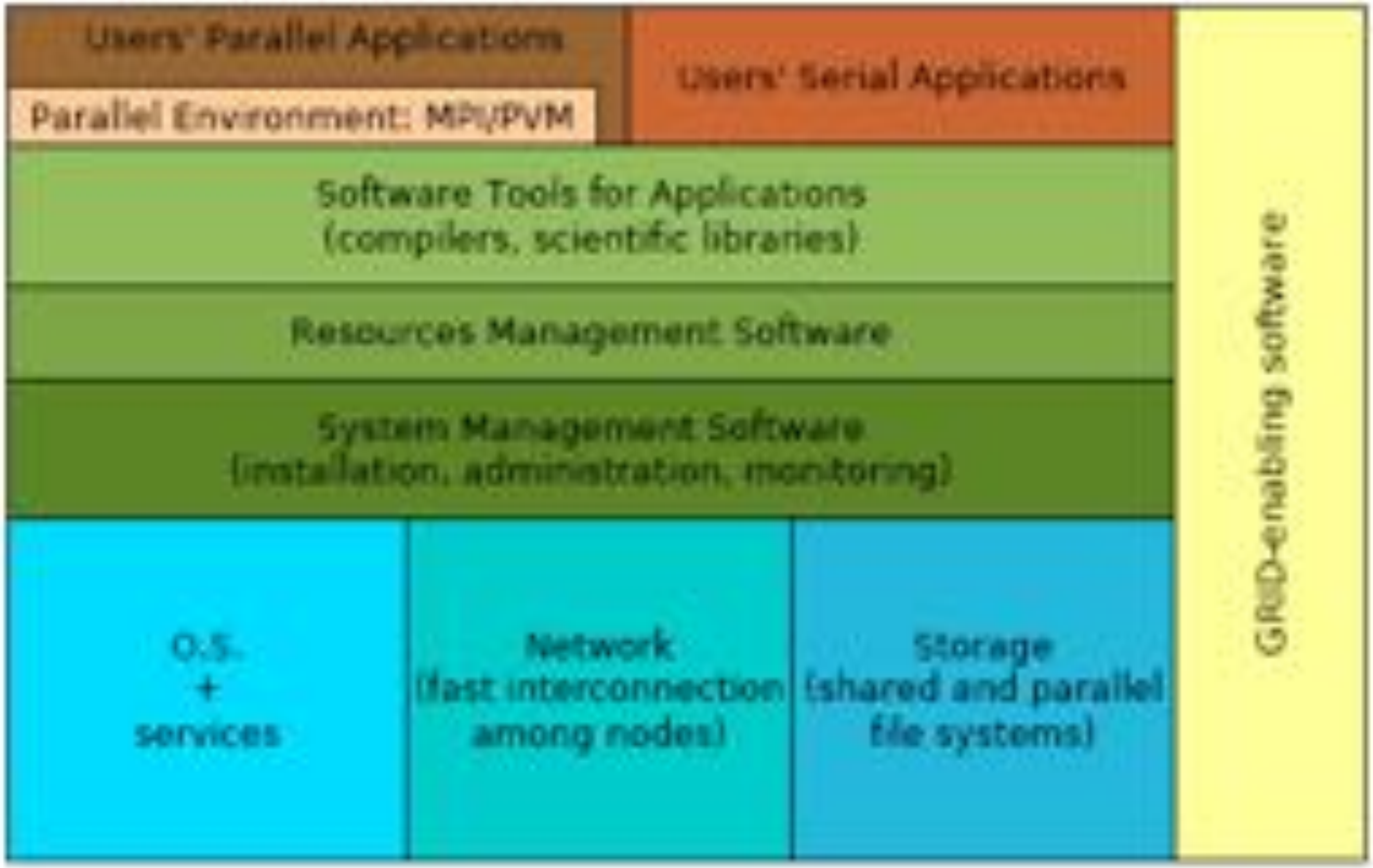


Figure 1.9 The system stack of a general supercomputer consists of a system hardware layer and several software layers. The first software layer is the operating system, encompassing both resource management and middleware to access input/output (I/O) channels. Higher software layers include runtime systems and workflow management.

HPC platform: the software stacks



Users' Parallel Applications

Parallel Environment: MPI/PVM

Users' Serial Applications

Software Tools for Applications (compilers, scientific libraries)

Resources Management Software

System Management Software (installation, administration, monitoring)

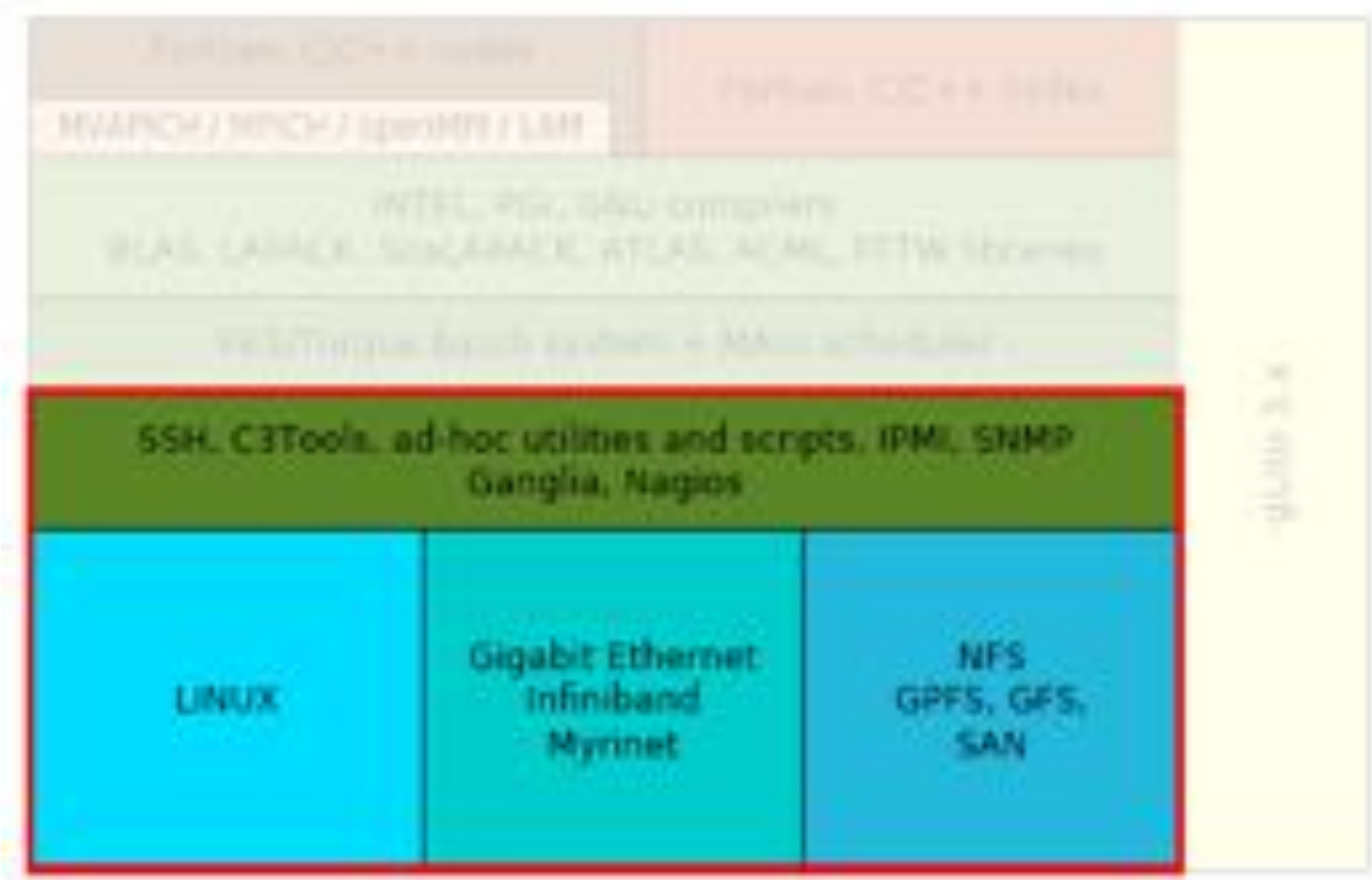
O.S. + services

Network (fast interconnection among nodes)

Storage (shared and parallel file systems)

GRID-enabling software

the software stacks: sys. adm...



Fortran, C/C++ codes

MVAPICH / MPICH / openMPI / LAM

Fortran, C/C++ codes

INTEL, PGI, GNU compilers
BLAS, LAPACK, ScaLAPACK, ATLAS, ACML, FFTW libraries

PBS/Torque batch system + MAUI scheduler

SSH, C3Tools, ad-hoc utilities and scripts, IPMI, SNMP
Ganglia, Nagios

LINUX

Gigabit Ethernet
Infiniband
Myrinet

NFS
GPFS, GFS,
SAN

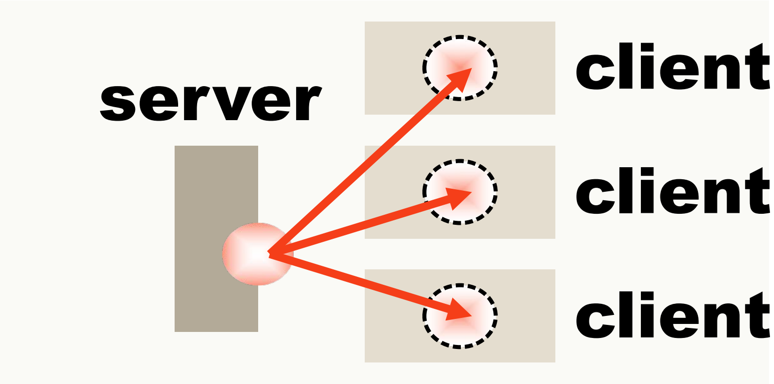
Cluster middleware: Middleware Design Goals

- Complete Transparency (Manageability):
  - Lets us see a single cluster system..
    - Single entry point, ftp, ssh, software loading...
- Scalable Performance:
  - Easy growth of cluster
    - no change of API & automatic load distribution.
- Enhanced Availability:
  - Automatic Recovery from failures
    - Employ checkpointing & fault tolerant technologies

Cluster middleware

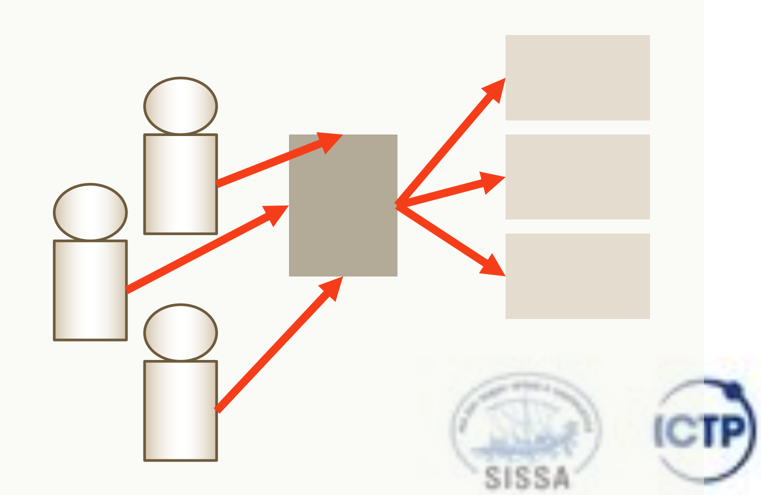
Administration software:

- user accounts
- NTP/NFS/ etc…



Resource management and scheduling software (LRMS)

- Process distribution
- Load balance
- Job scheduling of multiple tasks



How much does it cost a computational infrastructure ?

- It is not just a matter of HW…
- Total Cost of Ownership is the right way to calculate the budget for an HPC infrastructure..

Total Cost of Ownership

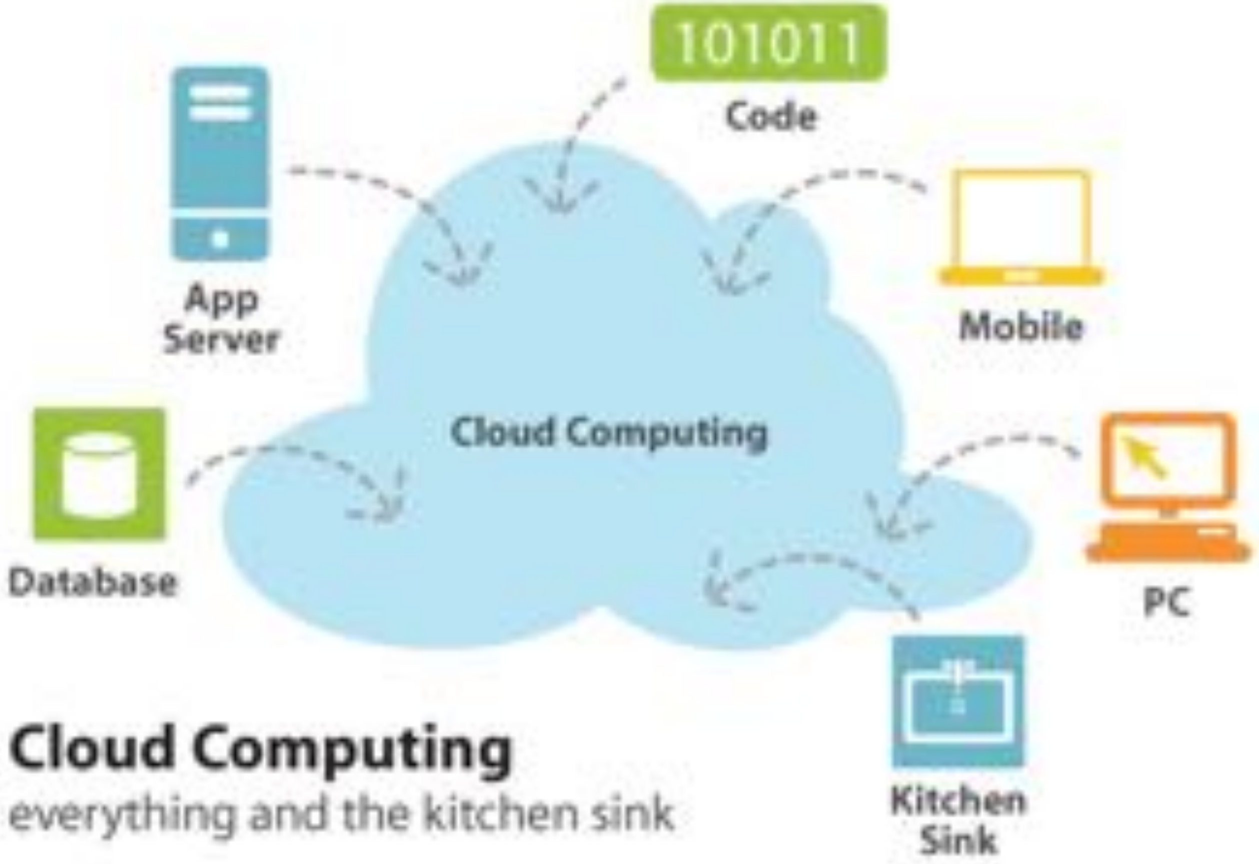
- It is the sum of all of the costs that a customer incurs during the lifetime of a technology solution.
- In the High Performance Computing (HPC) field, the Total Cost of Ownership is normally referred to the data center costs.
- Cost to the owner to build, operate and maintain the data center.
- Cost of Services delivered should be computed taking into account TCO.

What should be included in the TCO for HPC ?

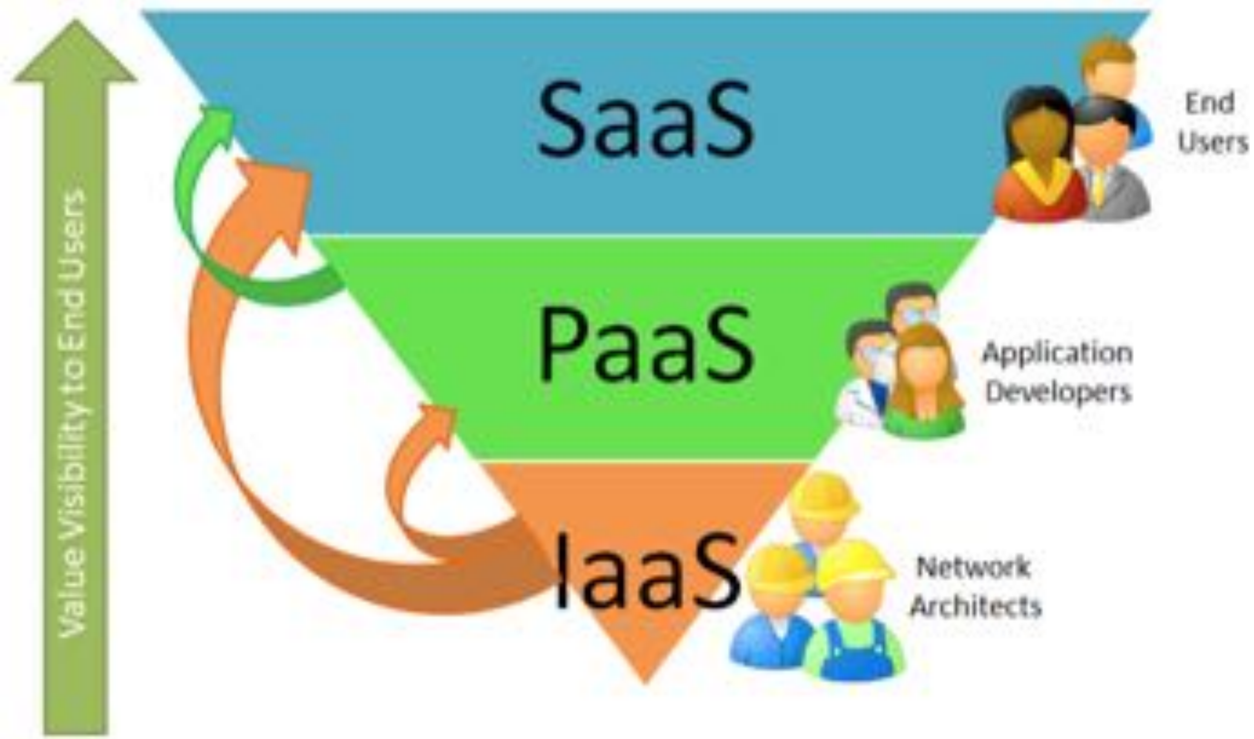
- Investment, operation and maintenance costs:
  - Hardware: servers, storage, networking, cabling, etc.
  - Electrical equipment: power distribution units, UPS, generators, etc.
  - Cooling systems: air conditioners, water cooling, etc.
- Infrastructure for the data center, power adaptation issues, etc.
- Energy consumption of the hardware and cooling systems
- Software licenses
- Human resources
- Maintenance

HOW can I reduce TCO ?

Cloud computing is the answer ?



cloud approach



The dream

- Cloud computing offers almost unlimited storage and instantly available and scalable computing resources,
- All the above at a reasonable metered cost.. (pay per use)
- However…
- the use of a typical cloud needs a bit of care..
- Remote HPC services can range from shared HPC clusters to fully virtualized cloud environments.

Cloud computing and HPC

- *The case for HPC in the cloud is growing stronger, but still has a way to go, especially for the more traditional HPC segments in the public sector.*

- https://www.hpcwire.com/2018/03/15/how-the-cloud-is-falling-short-for-research-computing/

HPC on cloud..

- cloud computing represented about 2% of the HPC market by total revenue in 2016.
- About 35% of HPC users make occasional use of public cloud resources.
- A number of vendors already exist within the industry providing HPC in the cloud solutions.

HPC cloud providers

- AMAZON WEB SERVICES (AWS)
- GOOGLE CLOUD PLATFORM
- MICROSOFT AZURE
- IBM SPECTRUM COMPUTING
- PENGUIN COMPUTING ON DEMAND (POD)

MHPC
Master in High Performance Computing

Conclusions

- HPC is about performance but not only
- Supercomputers are clusters !
- Clusters have many different components
- Parallel programming is needed to use HPC systems at best
- Several options/tools are available and sometime more than one approach is needed at the same time
- There are a lot of other lectures where all what we discussed in this first lectures will be analyzed in details

SISSA ICTP