Foundation on High Performance Computing

DSSC
MHPC

Stefano Cozzini
Ivan Girotto
Luca Tornatore

Trieste, 14th September 2015

Scuola Internazionale Superiore
di Studi Avanzati

The Abdus Salam
International Centre
for Theoretical Physics

Master in High Performance Computing

P1.2 course:

DAY 1: Introduction to HPC

Stefano Cozzini

CNR/IOM and eXact-lab srl

ma per seguir virtute e canoscenza
SISSA
Scuola Internazionale Superiore di Studi Avanzati

ICTP The Abdus Salam International Centre for Theoretical Physics

MHPC
Master in High Performance Computing

Agenda: first part

- Prologue: why and where HPC ?
- What is HPC ?
  - Definitions&metrics
- What is HPC infrastructure ?
  - Supercomputers & HPC Cluster
  - CPUs and Accelerators
  - Network/storage

ma per seguir virtute e canoscenza
SISSA
ICTP

MHPC
Master in High Performance Computing

Agenda: second part

- Software stack for HPC
  - Middleware: queue systems
  - Libraries/ Compiler/ performance Tools
- HPC Concepts
  - Parallel programming paradigms
  - Evolution of paradigms
  - Ahmdal law / Gustafson law
  - Strong/weak scalability
- HOMEWORK&LABS

SISSA
ICTP

MHPC
Master in High Performance Computing

Why is HPC important ?

"The next 10 to 20 years will see computational science firmly embedded in the fabric of science – the most profound development in the scientific method in over three centuries" (US Department of Energy).

"A host of technologies are on the horizon that we cannot hope to understand, develop, or utilize without simulation" (US National Science Foundation)

ma per seguir virtute e canoscenza
SISSA
ICTP

MHPC
Master in High Performance Computing

HPC: the challenge (from EU web site)

Societal, scientific and economic needs are the drivers for the next generation of HPC - computing with exascale performance (computers capable of performing 10 to the power of 18 floating point operations per second).

- All scientific disciplines are becoming "computational" today. Modern scientific discovery requires very high computing power and capability to deal with huge volumes of data.
- Industry and SMEs are increasingly relying on the power of supercomputers to invent innovative solutions, reduce cost and decrease time to market for products and services.
- HPC is part of a global race. Many countries (USA, Japan, Russia, China, Brazil, India) have announced ambitious plans for building the next generation of HPC with exascale performance and deploying state-of-the-art supercomputers.

From https://ec.europa.eu/programmes/horizon2020/en/h2020-section/high-performance-computing-hpc

MHPC
Master in High Performance Computing

Why HPC is important?

KEY ENABLING TECHNOLOGY

SISSA

ICTP

"Today, to Out-Compute is to Out-Compete".

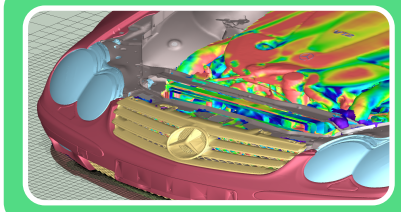
Image from UberCloud

MHPC
Master in High Performance Computing

Where HPC plays a role in Industry ? Some examples..

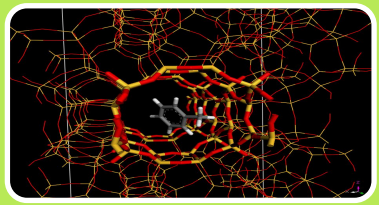
Aeronautics

where the design of airplanes more energy efficient and less noisy cannot be done without simulations involving very large models of the entire aircraft and analysis of physical phenomena at different scales,



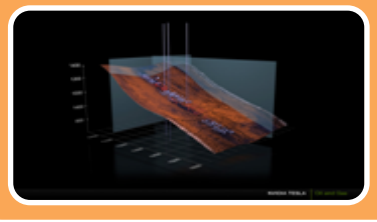
Automotive industry

which wishes to reduce consumption and CO2 emissions while increasing the level of comfort and security,



Pharmaceutical industry

were the discovery of new active molecules and new drugs is accelerated by numerical simulations,



Oil industry

which needs supercomputers to discover new oil fields and to optimize production of existing reservoirs.

MHPC
Master in High Performance Computing

HPC stands for…

Defining HPC (from Intersect survey)

- High Performance Computing (HPC) is the use of servers, clusters, and supercomputers – plus associated software, tools, components, storage, and services – for scientific, engineering, or analytical tasks that are particularly intensive in computation, memory usage, or data management
- HPC is used by scientists and engineers both in research and in production across industry, government and academia.

[to be continued]

ma per seguir virtute e canoscenza
SISSA
ICTP

MHPC
Master in High Performance Computing

Elements of HPC..

- use of servers, clusters, and supercomputers
  → HARDWARE
- associated software, tools, components, storage, and services
  → SOFTWARE
- scientific, engineering, or analytical tasks
  → PROBLEMS TO BE SOLVED..

ALL THE ABOVE DEFINES A COMPUTATIONAL INFRASTRUCTURE aka E-INFRASTRUCTURE

SISSA ICTP

MHPC
Master in High Performance Computing

Elements of e-infrastructure for HPC

- E-infrastructure for HPC includes:
  - Servers/nodes/accelerators
  - High speed Networks
  - High end parallel storage
  - Middleware
  - Scientific/Technical Software
  - Research/Technical data (scientific databases, individual data…)
  - Problems to be solved

IS ALL WHAT WE NEED ?

NO

MHPC
Master in High Performance Computing

Last but not least: people

- Human capital is by far the most important aspect
- Two important roles:
  - HPC providers (plan/install/manage HPC resources)
  - HPC user

MIXING/INTERPLAYING ROLES INCREASES COMPETENCE LEVELS

SISSA
ICTP

MHPC
Master in High Performance Computing

HPC Users…

High Performance Technical Computing (HPTC)

- Applications in science and engineering
- Top Market: Academia/Government Lab/ Manufacturing/ Bio Life/ Oil gas exploration

High Performance Business Computing (HPBC)

- Application included trading/pricing/ risk management/ online gaming/ analytics/fraud detection/ logistics
- Top Market: Banks/ Insurance Companies/ Online games/ Financial services/ Entertainment

SISSA
ICTP

MHPC
Master in High Performance Computing

From computational infrastructure to computational environment…

Goal:

provide a computational environment to satisfy all the different requirements posed by users

- Which kind of requirements ?

All you need to to solve their computational problems

- Which kind of users ?

Any Kind:

- Computa
- R&D de
- Governr
- Videoga
- Etc..etc..

COMPUTATIONAL ENVIRONEMENT IS COMPOSED BY HW/SW and PEOPLE

SISSA ICTP

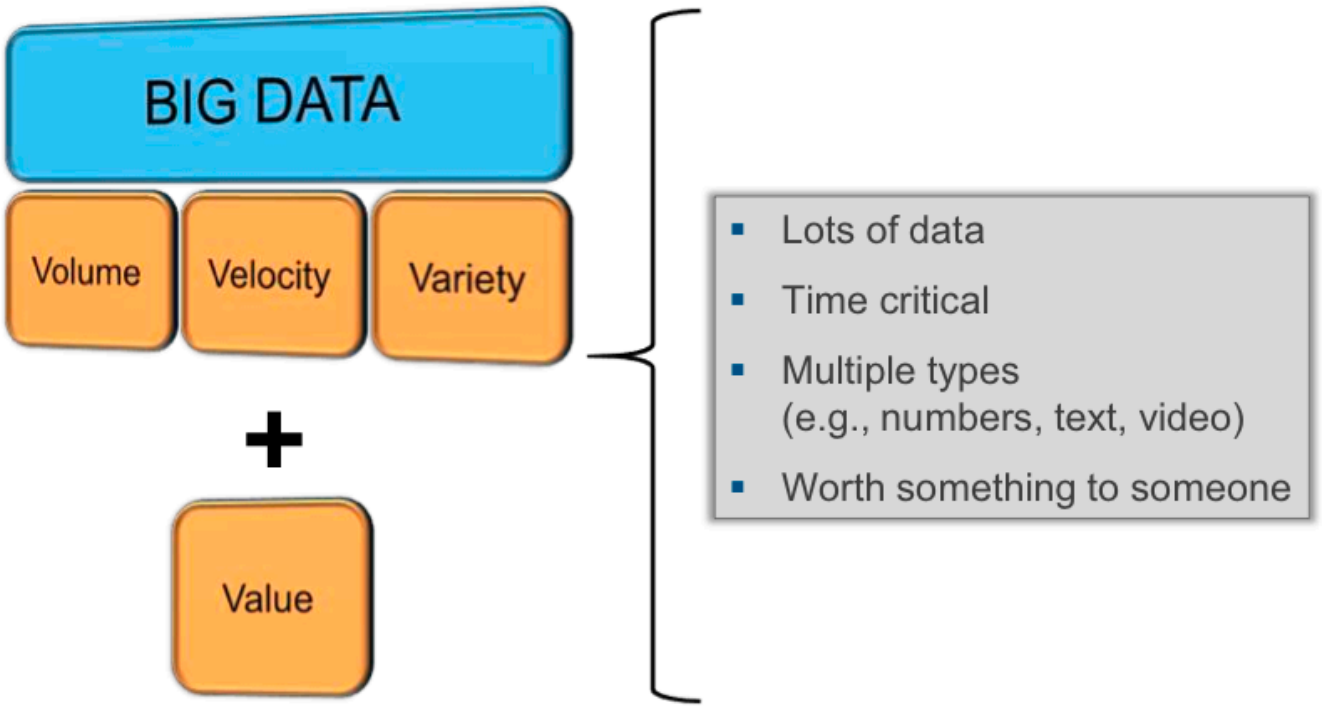
MHPC
Master in High Performance Computing

Challenges ahead HPC (I)

- HPC skilled people
- Big data !

ma per seguir virtute e canoscenza
SISSA
ICTP

Big Data: general definition



MHPC
Master in High Performance Computing

The 3 V's of big data..

- Velocity

  Data are produced at speed higher than the speed you are able to move/analyze and understand them..
- Variety
  - Data range from simulation to remote sensing information, from instruments to market analysis etc..
  - datasets come in a variety of data formats and span a variety of metadata standards
- Volume

  From the dawn of civilization until 2003, humankind generated five exabytes of data. Now we produce five exabytes every two days… and the pace is accelerating"

ma per seguir virtute e canoscenza
SISSA
ICTP

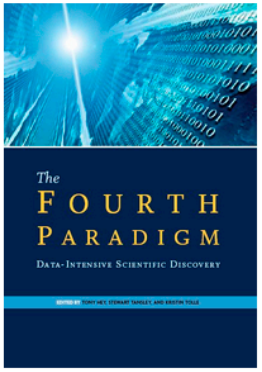
MHPC
Master in High Performance Computing

Data-intensive science

- A "fourth paradigm" after experiment, theory, and computation..

The Fourth Paradigm: Data-Intensive Scientific Discovery

Presenting the first broad look at the rapidly emerging field of data-intensive science



Increasingly, scientific breakthroughs will be powered by advanced computing capabilities that help researchers manipulate and explore massive datasets.

The speed at which any given scientific discipline advances will depend on how well its researchers collaborate with one another, and with technologists, in areas of eScience such as databases, workflow management, visualization, and cloud computing technologies.

In *The Fourth Paradigm: Data-Intensive Scientific Discovery*, the collection of essays expands on the vision of pioneering computer scientist Jim Gray for a new, fourth paradigm of discovery based on data-intensive science and offers insights into how it can be fully realized.

Critical praise for *The Fourth Paradigm*

Download

- Full text, low resolution (6 MB)
- Full text, high resolution (93 MB)
- By chapter and essay

Purchase from Amazon.com

- Paperback
- Kindle version

In the news

- Sailing on an Ocean of 0s and 1s *(Science* Magazine)
- A Deluge of Data Shapes a New Era in Computing *(New York Times)*
- A Guide to the Day of Big Data *(Nature)*

It involves collecting, exploring, visualizing, combining, subsetting, analyzing, and using huge data collections

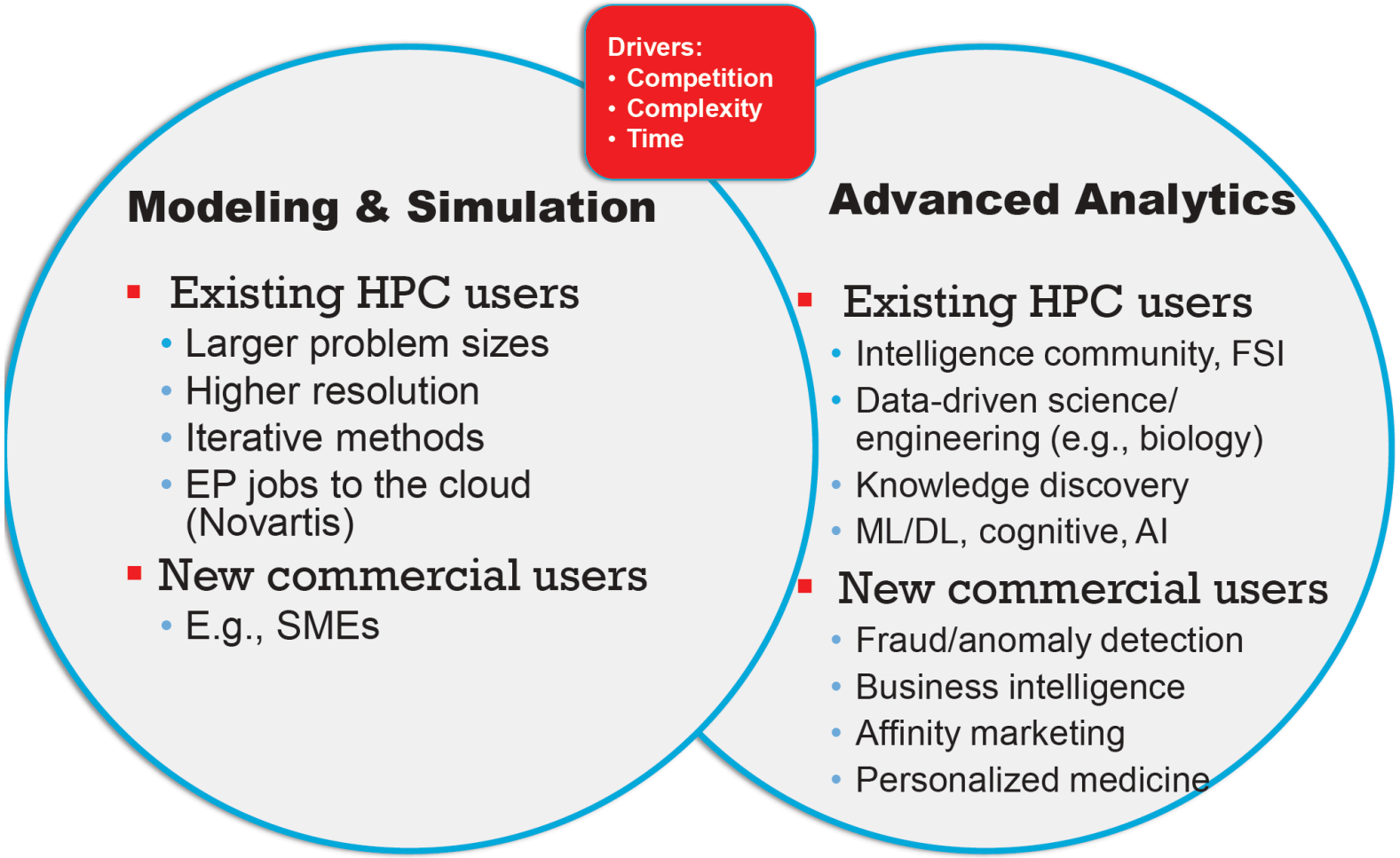
SISSA ICTP

MHPC
Master in High Performance Computing

Big data challenges: HPC is now HPDA

- "*BigData*" is a growing trend affecting many HPC applications touching large datacenters, and research
- Fueled by creation and availability of many data (more complex scientific instruments and sensor networks, from "smart" power grids to the Large Hadron Collider and Square Kilometer Array)
- The proliferation of larger, The increasing transformation of certain disciplines into data-driven sciences. Biology is a notable example, but this transformation extends even to humanities disciplines such as archeology and linguistics.
- The availability of newer advanced analytics methods and tools: MapReduce/Hadoop, graph analytics, semantic analysis, knowledge discovery algorithms, and others
- The escalating need to perform advanced analytics in near-real time—a need that is causing a new wave of commercial firms to adopt HPC for the first time
- Growth in these application areas creates a market opportunity for providers/experts of HPC technologies

HPDA = Data-Intensive Computing Using HPC



MHPC
Master in High Performance Computing

P stands just for PERFORMANCE ?

Performance is not always what matters..

> to reflect a greater focus on the productivity, rather than just the performance, of large-scale computing systems, many believe that HPC should now stand for High Productivity Computing.
>
> [from wikipedia]

- P should also stand for PROFITABILITY

SISSA
ICTP

MHPC
Master in High Performance Computing

Performance vs Productivity

- A possible definition:
  - Productivity = (application performance) / (application programming effort)
- people in HPC arena have different goals in mind thus different expectations and different definitions of productivity.

Question: Which kind of productivity are you interested in ?

SISSA ma per seguir virtute e canoscenza
ICTP

MHPC
Master in High Performance Computing

HPC stands for... (part II)

Defining HPC (from Intersect survey)

- Within industry, HPC can frequently be distinguished from general business computing in that companies generally will use HPC applications to gain advantage in their core endeavors – e.g., finding oil, designing automobile parts, or protecting clients' investments – as opposed to non-core endeavors such as payroll management or resource planning

MHPC
Master in High Performance Computing

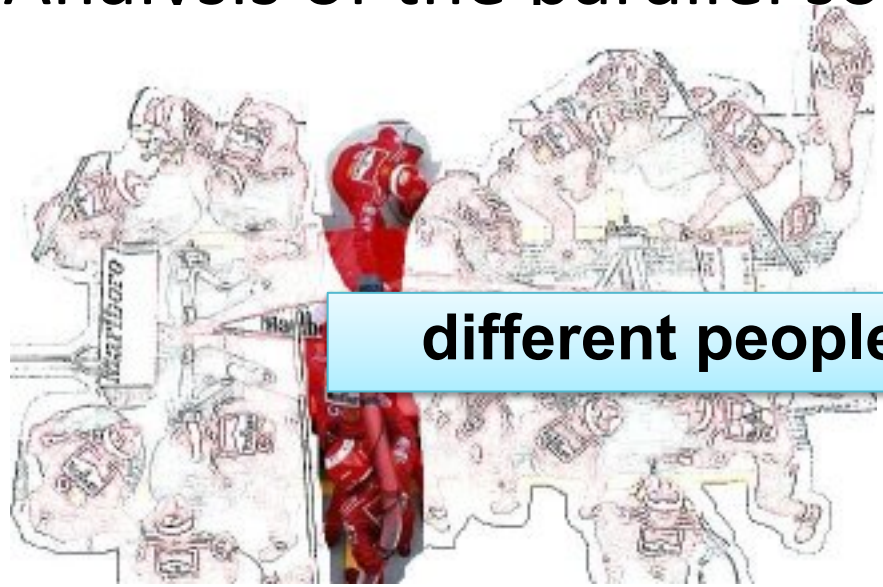
High Performance problem example:

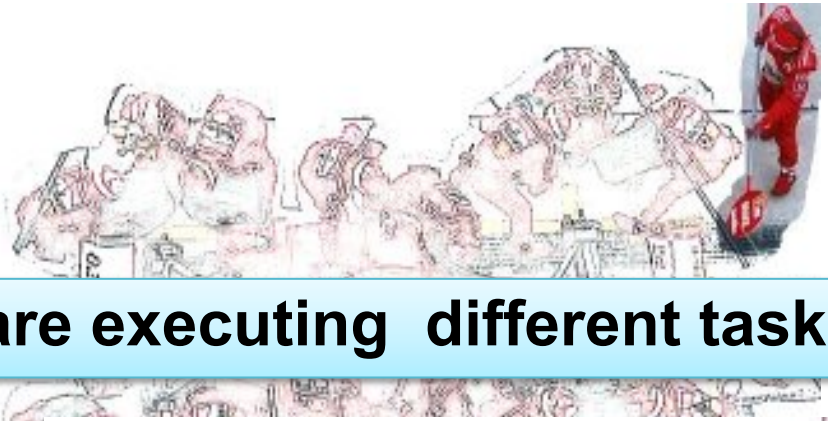
picture from http://www.f1nutter.co.uk/tech/pitstop.php

SISSA ICTP

FUNCTIONAL PARTITIONING

Analysis of the parallel solution:





different people are executing different tasks

DOMAIN DECOMPOSITION

different people are solving the same global task but on smaller subset

HPC

=

Parallel computing

MHPC
Master in High Performance Computing

Measuring speed of HPC systems

- How fast can I crunch numbers on my CPUs ?
- How fast can I move data around ?
  - from CPUs to memory
  - from CPUs to disk
  - from CPUs on different machines
- How much data can I store ?

MHPC
Master in High Performance Computing

Number crunching on CPU: what do we count ?

- Rate of [million/billions of] floating point operations per second ([M|G]flops) FLOPs/S
- Theoretical peak performance:
  - determined by counting the number of floating-point additions and multiplications that can be completed during a period of time, usually the cycle time of the machine

FLOPS=Clock-rate*Number_of_FP_ operation*Number_of_cores

Sustained (peak) performance

Real (sustained) performance: a measure

-measured by taking the time the code requires to run

$$\frac{(\text{Number\_of\_floating\_point\_operations of the code})}{\text{Time measured}}$$

-Number_of_floating_point_operations not easy to be defined for real application

-Top500 list uses HPL Linpack:

-Sustained peak performance is what's matter in TOP500

TOP 500 List

- The TOP500 list www.top500.org
  - published twice a year from 1993
    - ISC conference in Europe (June)
    - Supercomputing conference in USA (November)
  - List the most powerful computers in the world
  - yardstick: Linpack benchmark (LU – decomposition)

HPL: some details

From http://icl.cs.utk.edu/hpl/index.html:

-The code solves a uniformly random system of linear equations and reports time and floating-point execution rate using a standard formula for operation count.

-Number_of_floating_point_operations = $2/3n^3 + 2n^2$ (n=size of the system)

```
================================================================================
T/V                N    NB     P     Q               Time                 Gflops
--------------------------------------------------------------------------------
WR03R2L2       86000  1024     2     1             191.06              2.219e+03
--------------------------------------------------------------------------------
||Ax-b||_oo/(eps*(||A||_oo*||x||_oo+||b||_oo)*N)=        0.0043644 ...... PASSED
```

MHPC
Master in High Performance Computing

HPL&TOP500

- For each machine the following numbers are reported using HPL:
  - Rmax: the performance in GFLOPS for the largest problem run on a machine.
  - Rpeak: the theoretical peak performance GFLOPS for the machine.
  - The measure of the power required to run the benchmark

MHPC
Master in High Performance Computing

And the winner is...

Sunway TaihuLight: National Supercomputing Center in Wuxi
No.1 from Jun 2016 until Jun 2017

Tianhe-2 (MilkyWay-2) : National University of Defense Technology
No.1 from Jun 2013 until Nov 2015

Titan: Oak Ridge National Laboratory
No.1 in Nov 2012

Sequoia: Lawrence Livermore National Laboratory
No.1 in Jun 2012

K Computer: RIKEN Advanced Institute for Computational Science
No.1 from Jun 2011 until Nov 2011

Tianhe-1A: National Supercomputing Center in Tianjin
No.1 in Nov 2010

MHPC
Master in High Performance Computing

2017 winner…

Sunway TaihuLight - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway

Site:	National Supercomputing Center in Wuxi
Manufacturer:	NRCPC
Cores:	10,649,600
Memory:	1,310,720 GB
Processor:	Sunway SW26010 260C 1.45GHz
Interconnect:	Sunway
Performace	
Linpack Performance (Rmax)	93,014.6 TFlop/s
Theoretical Peak (Rpeak)	125,436 TFlop/s
Nmax	12,288,000
HPCG [TFlop/s]	480.8
Power Consumption	
Power:	15,371.00 kW (Submitted)
Power Measurement Level:	2
Software	
Operating System:	Sunway RaiseOS 2.0.5

First EU machine..

Piz Daint - Cray XC50, Xeon E5-2690v3 12C 2.6GHz, Aries interconnect , NVIDIA Tesla P100

Site:	Swiss National Supercomputing Centre (CSCS)
System URL:	http://www.cscs.ch/computers/piz_daint_piz_dora/index.html
Manufacturer:	Cray Inc.
Cores:	361,760
Memory:	340,480 GB
Processor:	Xeon E5-2690v3 12C 2.6GHz
Interconnect:	Aries interconnect
Performace	
Linpack Performance (Rmax)	19,590 TFlop/s
Theoretical Peak (Rpeak)	25,326.3 TFlop/s
Nmax	3,569,664
HPCG [TFlop/s]	470.0
Power Consumption	
Power:	2,271.99 kW (Optimized: **1631.13** kW)
Power Measurement Level:	3
Measured Cores:	361,760
Software	
Operating System:	Cray Linux Environment

MHPC
Master in High Performance Computing

Challenges ahead HPC (I)

- HPC skilled people !
- Big data ! !
- Sustained performance not just HPL !!!
- Energy !!!!

ma per seguir virtute e canoscenza
SISSA
ICTP

MHPC
Master in High Performance Computing

Sustained peak performance on real scientific codes

Blue-waters at NCSA: 22,640 AMD 6276 processors
Theoretical peak performance: 13 Petaflops
Sustained performance on real scientific codes:..

Scientific code	Number of cores	Performance achieved(PF)	runtime (hour)
VPIC	22528	1.25	2.5
PPM	21417	1.23	~ 1
QMCPACK	22500	1.037	~1
SPECF3MD	21675	>1	Not reported
WRF	8192	0,160	<0.50

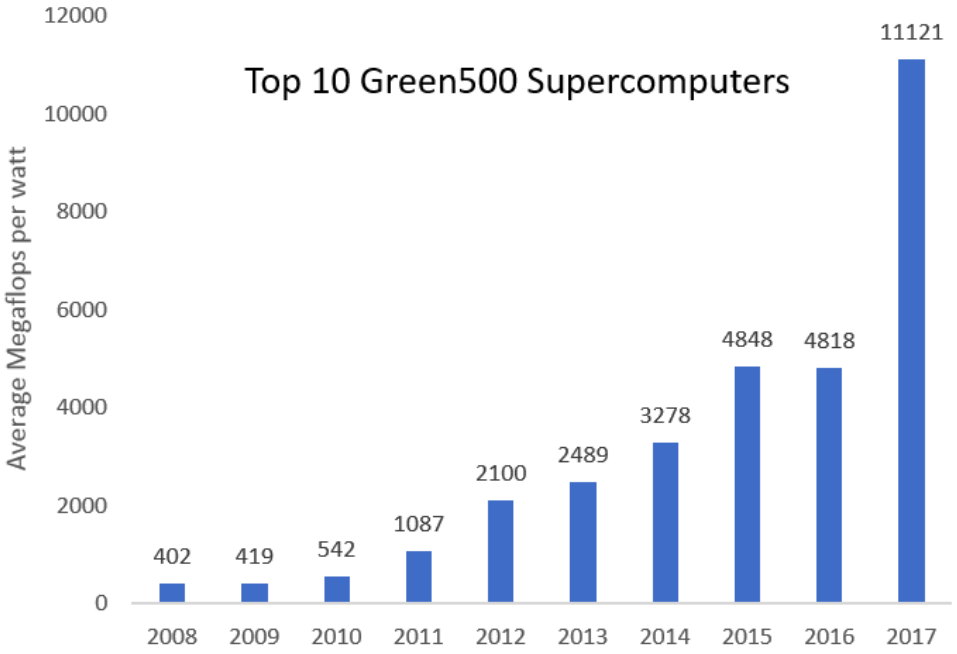
http://www.cray.com/sites/default/files/resources/XE6-NCSA-PFApplications-0514.pdf

SISSA ICTP

MHPC
Master in High Performance Computing

Top500&Green500

- Over the last year, the greenest supercomputers more than doubled their energy efficiency
- If such a pace can be maintained, exascale supercomputers operating at less than 20 MW will be possible in as little as two years.
- But that's a big if.



SISSA
ICTP

Top Green500 June 2017

Rank	TOP500 Rank	System	Cores	Rmax (TFlop/s)	Power (kW)	Power Efficiency (GFlops/watts)
1	61	**TSUBAME3.0** - SGI ICE XA, IP139-SXM2, Xeon E5-2680v4 14C 2.4GHz, Intel Omni-Path, NVIDIA Tesla P100 SXM2 , HPE GSIC Center, Tokyo Institute of Technology Japan	36,288	1,998.0	142	14.110
2	465	**kukai** - ZettaScaler-1.6 GPGPU system, Xeon E5-2650Lv4 14C 1.7GHz, Infiniband FDR, NVIDIA Tesla P100 , ExaScalar Yahoo Japan Corporation Japan	10,080	460.7	33	14.046
3	148	**AIST AI Cloud** - NEC 4U-8GPU Server, Xeon E5-2630Lv4 10C 1.8GHz, Infiniband EDR, NVIDIA Tesla P100 SXM2 , NEC National Institute of Advanced Industrial Science and Technology Japan	23,400	961.0	76	12.681
4	305	**RAIDEN GPU subsystem** - NVIDIA DGX-1, Xeon E5-2698v4 20C 2.2GHz, Infiniband EDR, NVIDIA Tesla P100 , Fujitsu Center for Advanced Intelligence Project, RIKEN Japan	11,712	635.1	60	10.603
5	100	**Wilkes-2** - Dell C4130, Xeon E5-2650v4 12C 2.2GHz, Infiniband EDR, NVIDIA Tesla P100 , Dell University of Cambridge United Kingdom	21,240	1,193.0	114	10.428

MHPC
Master in High Performance Computing

Exercise 1 : compute Theoretical Peak performance for your laptop/desktop: (mandatory for MHPC&DSCC optional)

Identify the CPU
Identify the frequency
Identify the number of floating point for cycle
Identify how many cores
Put all together in one single number and tell me

Exercise 2 : compute sustained Peak performance for your cell-phone (mandatory for all MHPC&DSSC)

Identify an app to run HPL
Run it
Tune it and get the best number you can…

SISSA
ICTP

MHPC
Master in High Performance Computing

Moving data around: bits and Mb/sec

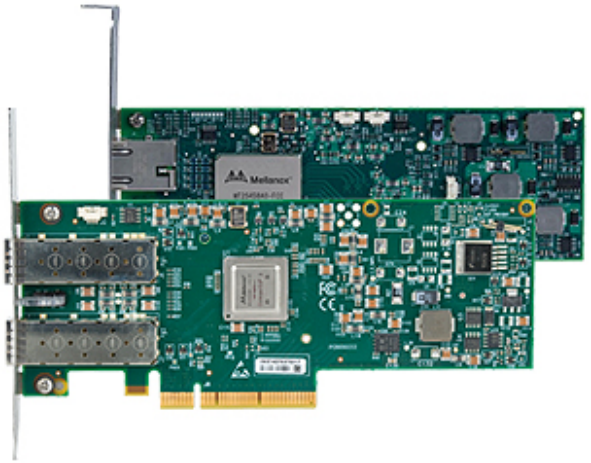
bit/second transmitted

within the computer:

- CPU-Memory: thousand of Mb/sec GByte/s
  - 10 - 100 Gbit
- CPU- Disks : MByte/sec
  - 50 ~ 100 MB up 1000MB/sec

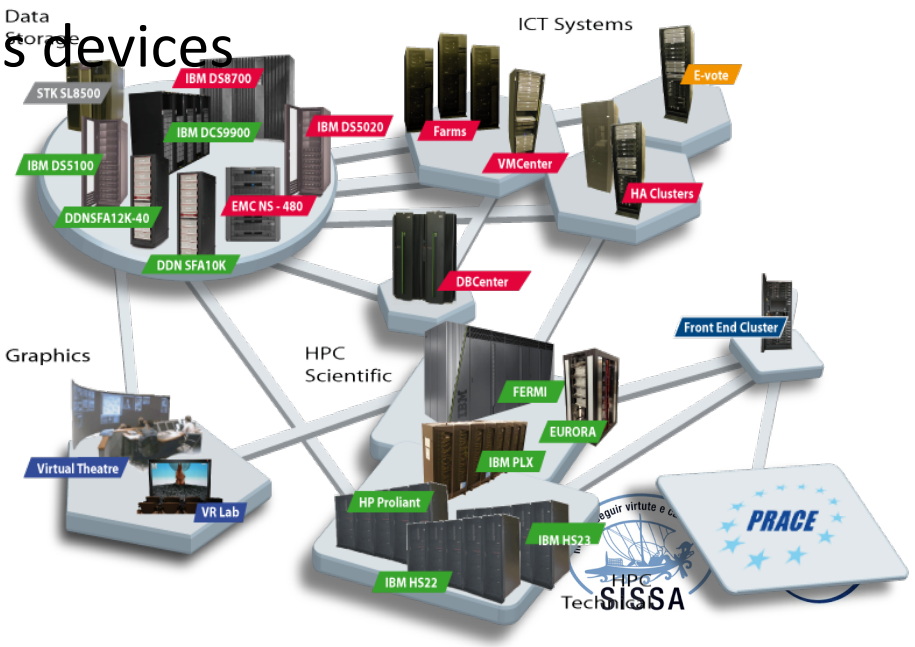
among computers: networks

- -default (commodity)
  - 1000Mbit=1Gbit
- -custom(high speed)
  - 10Gb and now 100 Gb (and even more)



Storage size: bytes

- size of storage devices:
  - kbyte/Mbyte -->caches/RAM
  - Gigabyte ---> RAM/hard disks
  - Terabyte ---> Disks/SAN
  - Petabyte ----> SAN / Tapes devices

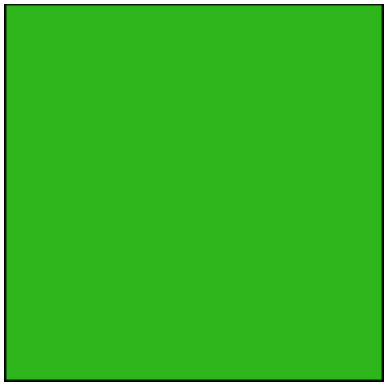


IOPS vs FLOPS

- HPC is too compute-centric
- Modern Scientific&technical computing requires access to data and computing

computing 1 calculation
≈ 1 picojoule

moving 1 calculation
≈ 100 picojoule



Source: IDC Direction 2013

MHPC
Master in High Performance Computing

HPC

=

Parallel computing

SISSA

ICTP

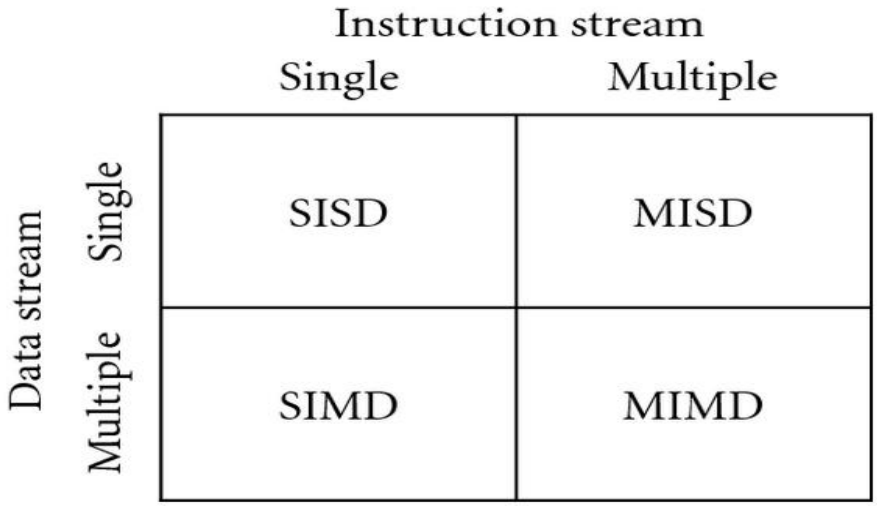
MHPC
Master in High Performance Computing

PARALLEL HARDWARE

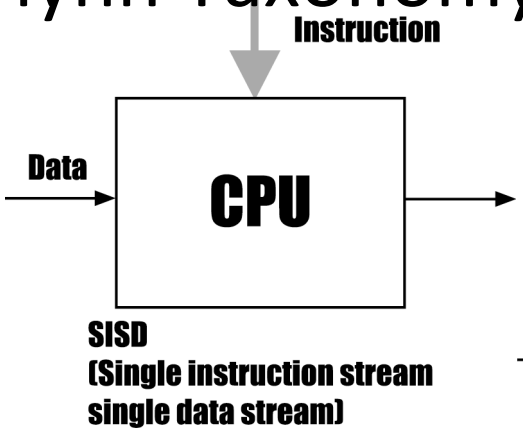
SISSA
ICTP

Parallel computers

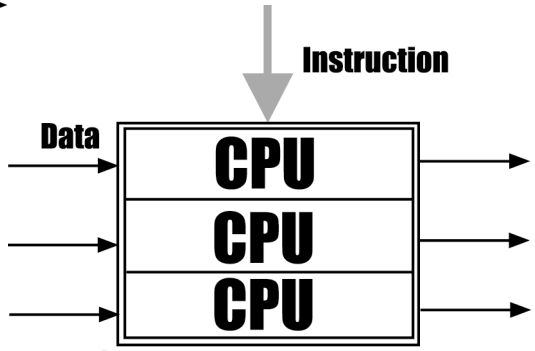
- Tons of different machines !
- Flynn Taxonomy (1966): helps (?) us in classifying them:
  - Data Stream
  - Instruction Stream



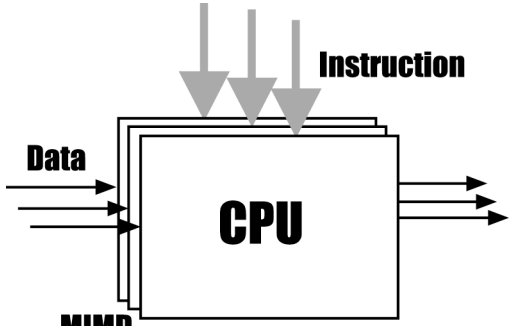
Flynn Taxonomy (graphical view)



SISD
(Single instruction stream
single data stream)



SIMD
(Single instruction stream
multipe data stream)



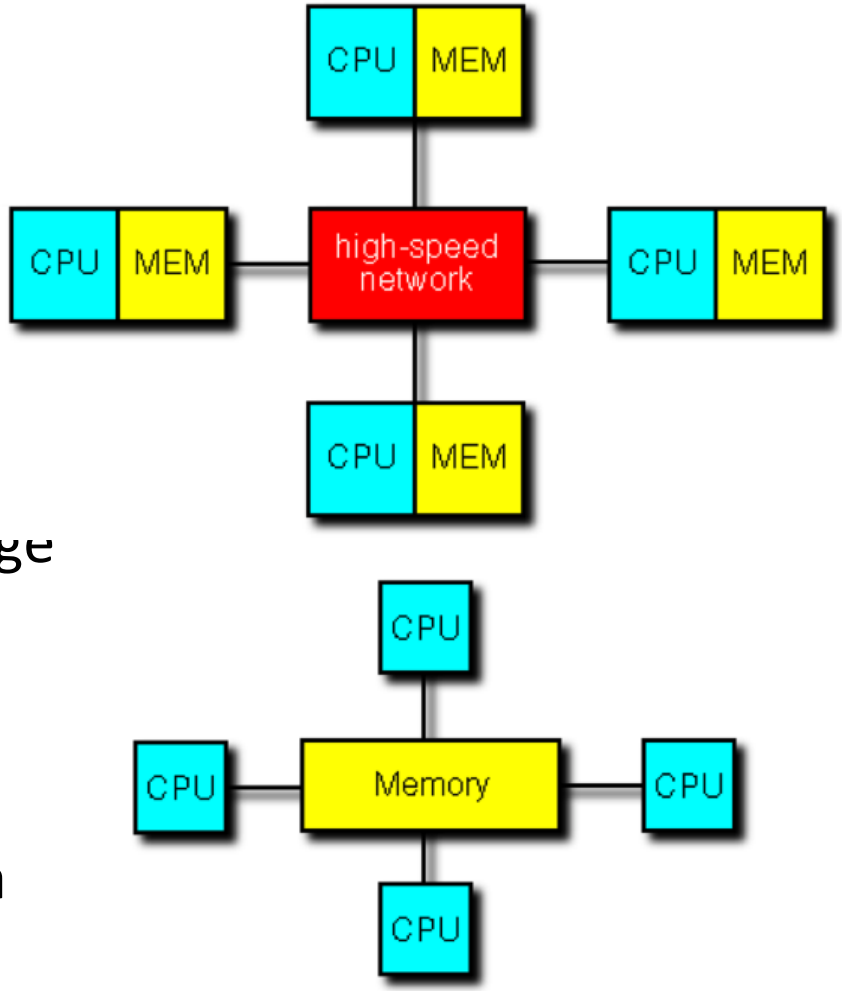
MIMD
(Multipe instruction stream
mingle data stream)

Another important question:

- MEMORY: The simplest and most useful way to classify modern parallel computers is by their memory model:
  - SHARED MEMORY
  - DISTRIBUTED MEMORY
  - MIXED SITUATION

Shared vs Distributed me

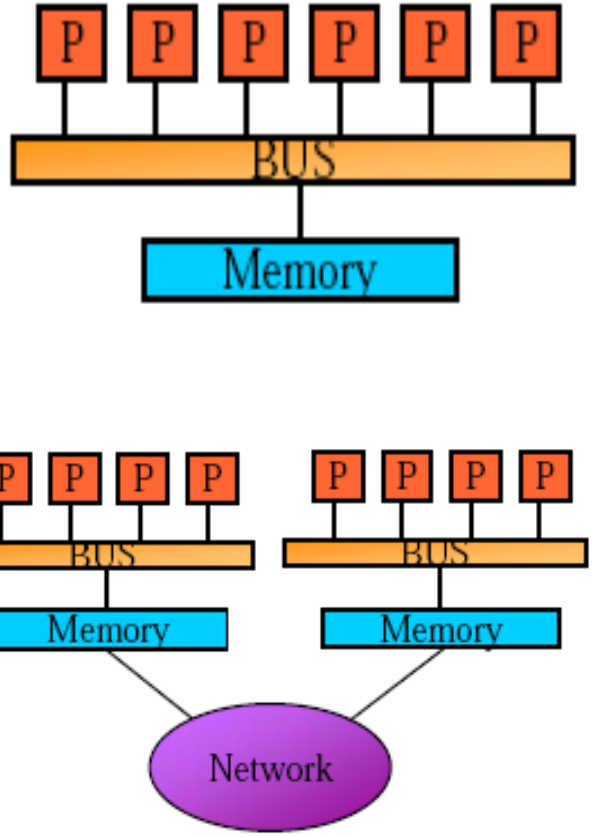
- Distributed memory
  - each processor has its own local memory. Must do message passing to exchange data between processors
- Shared Memory
  - single address space. All processors have access to a pool of shared memory.



Shared memory: UMA vs NUMA

Uniform memory access (UMA): Each processor has uniform access to memory. Also known as symmetric multiprocessors (SMP)

Non-uniform memory access (NUMA): Time for memory access depends on location of data. Local access is faster than non-local access.



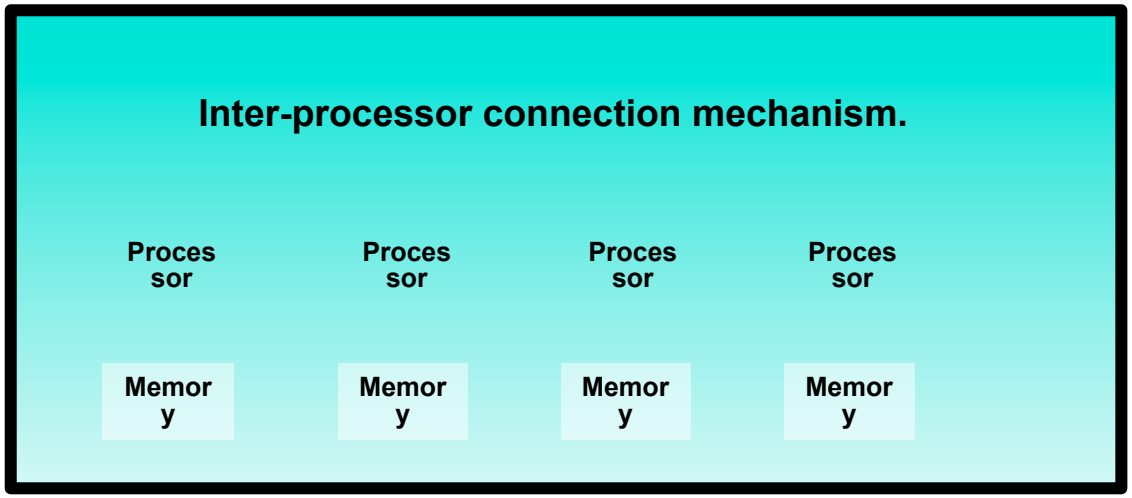
More on distributed memory machines

- The memory is physically distributed among the processors (local memory). Each processor can access directly only to its own local memory

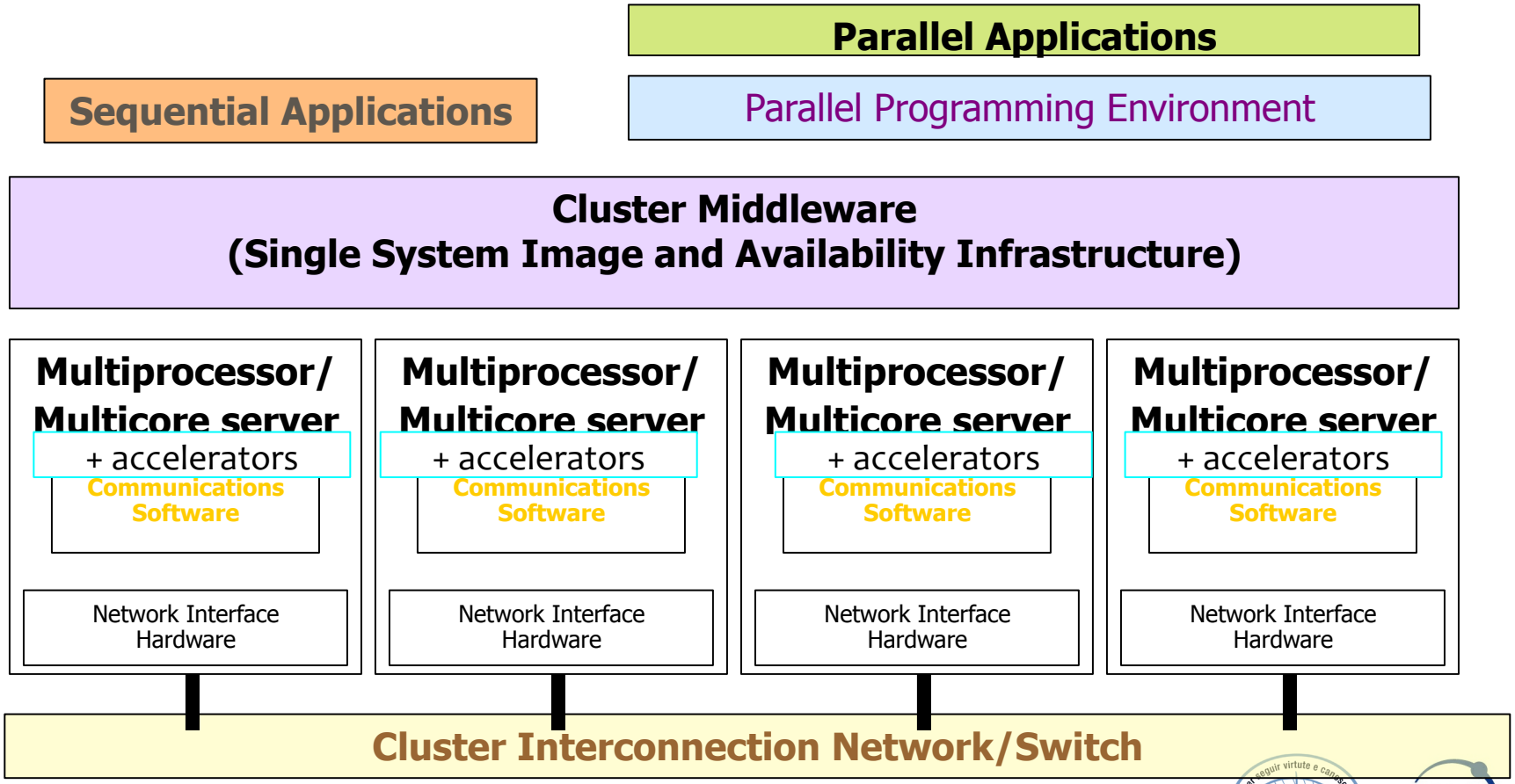
  NO-Remote Memory Access (NORMA) model
- Communication among different processors occurs via a specific communication protocol (message passing).
- In general distributed memory systems can scale-up from a small number of processors $O(10^2)$ to huge numbers of processors $O(10^7)$
- The performance of the system are influenced by:
  - Features of the node (RAM/cores/CPU frequency/ accelerator)
  - Features (Topology and other) of the interconnection network

Distributed memory architecture: Clusters !

- CLUSTER: independent machines combined into a unified system through software and networking



HPC Cluster Computer Architecture



MHPC
Master in High Performance Computing

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..

SISSA
ICTP

MHPC
Master in High Performance Computing

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

SISSA
ICTP

MHPC
Master in High Performance Computing

Total Cost of Ownership

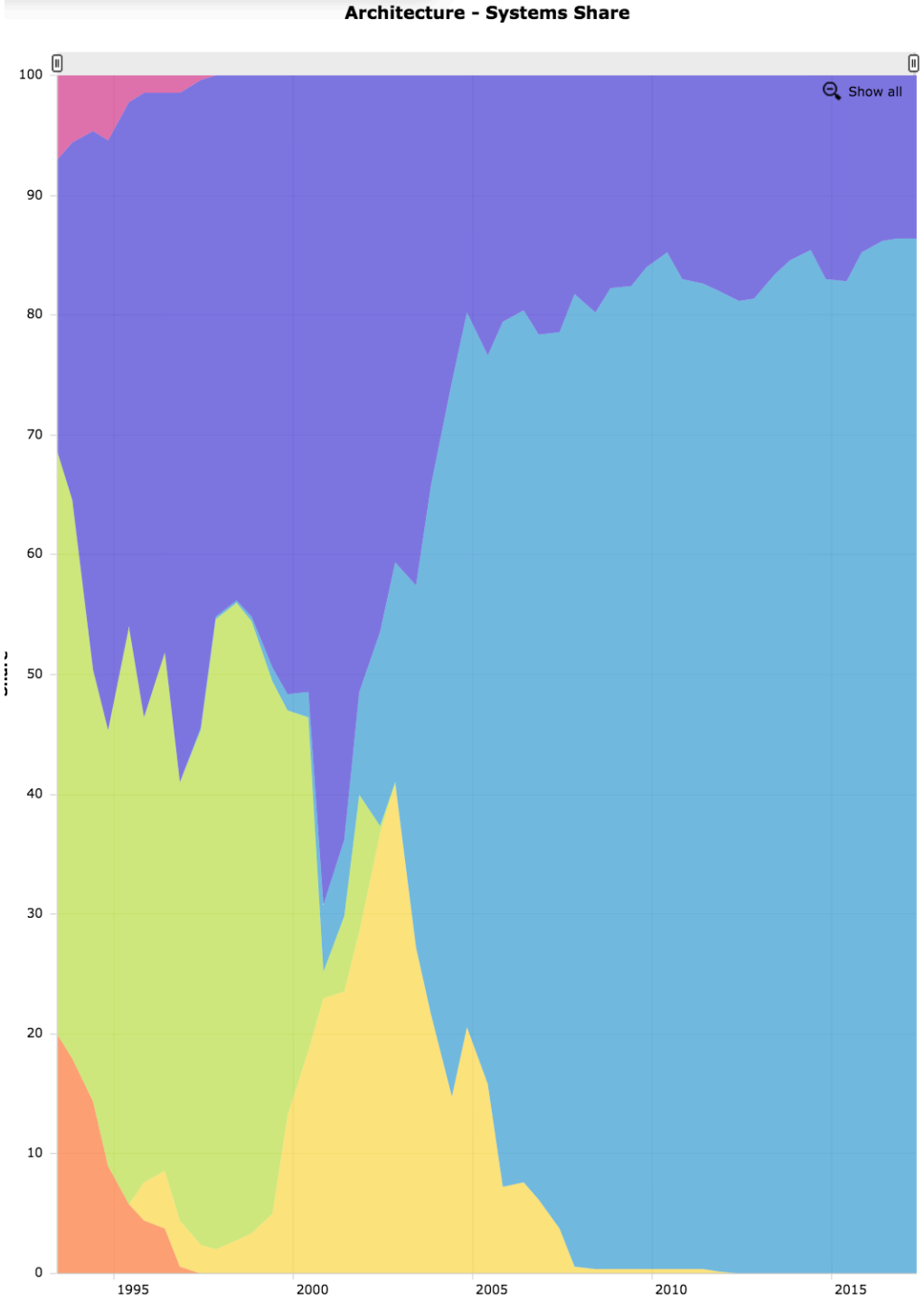
- It is the sum of all of the costs that a customer incurs during the lifetime of a technology solution.
- In the High Performance Computing (HPC) field, the Total Cost of Ownership is normally referred to the data center costs.
- Cost to the owner to build, operate and maintain the data center.
- Cost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

SISSA
ICTP

MHPC
Master in High Performance Computing

Top500 supercomputer

- 432 CLUSTERS
- 68 MPP (Massive Parallel Processors)

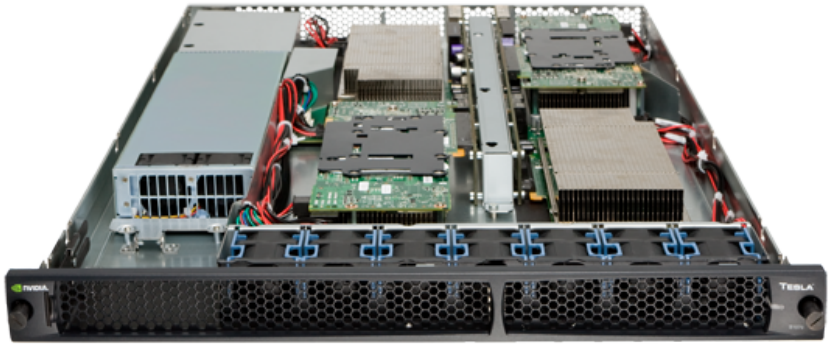


Elements of the clusters

- 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

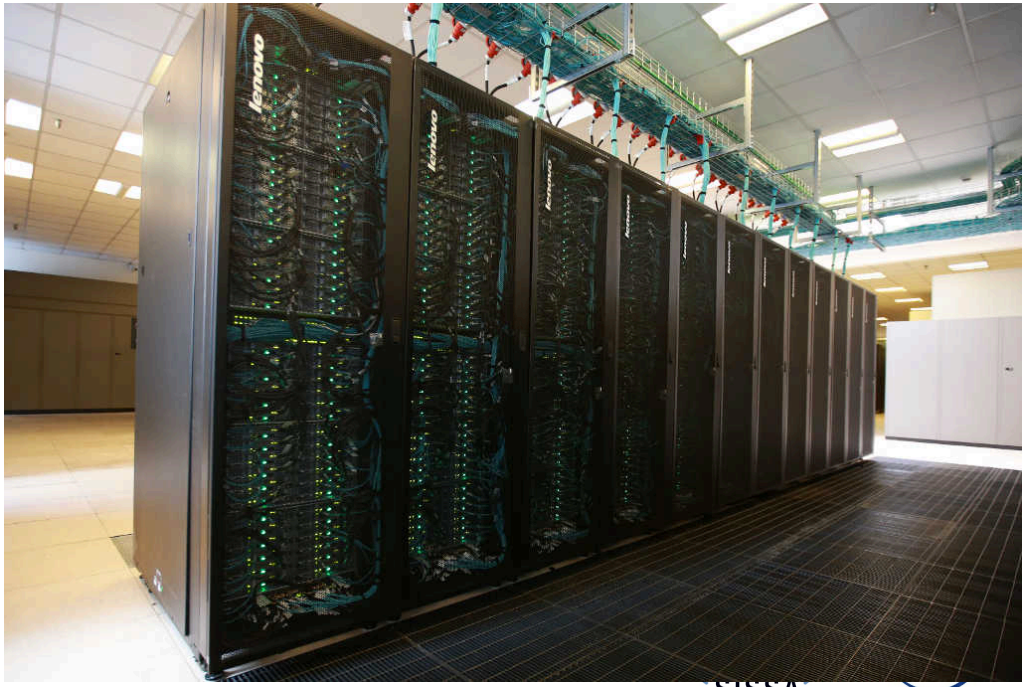


1 or 2 accelerators

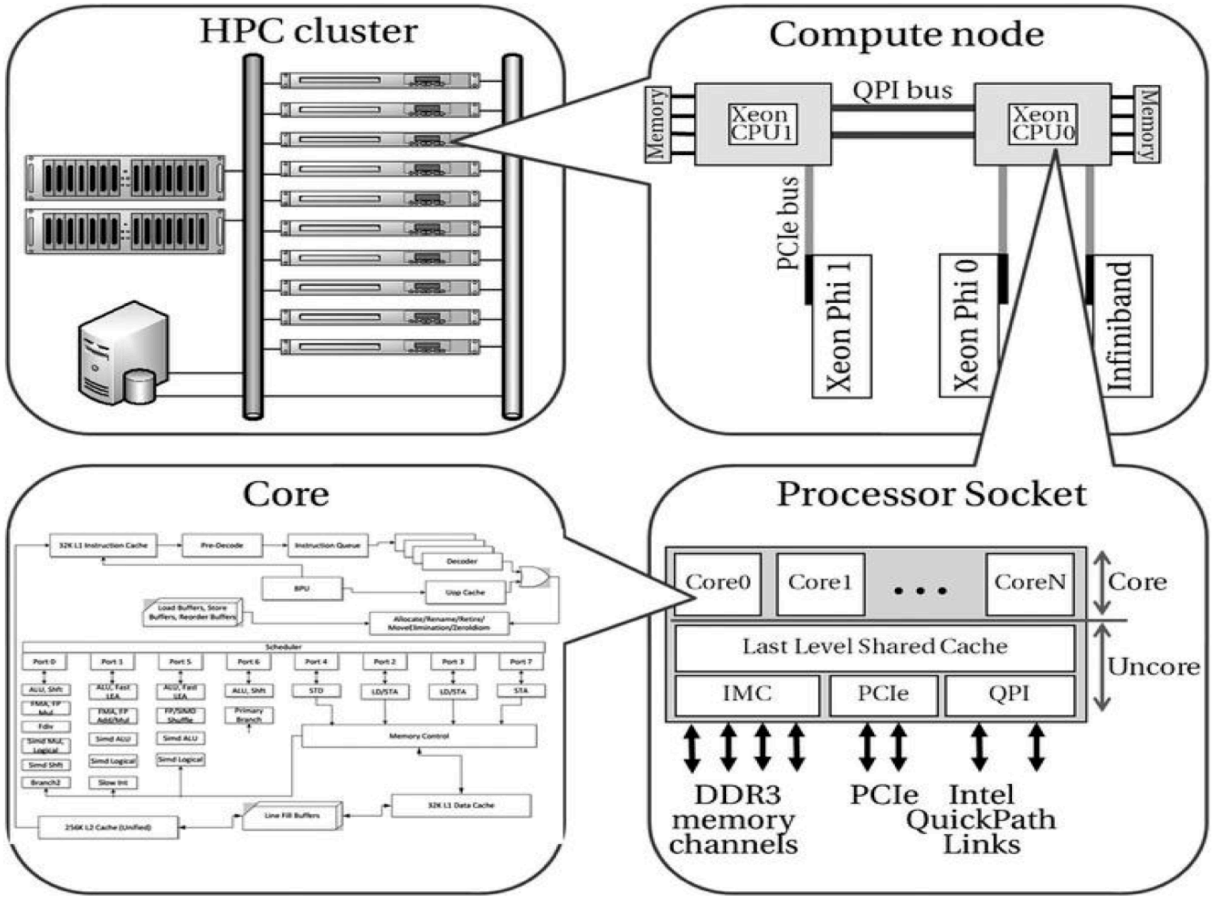
A shared memory machine (SMP or NUMA)

MHPC
Master in High Performance Computing

Some times also blades.. In racks

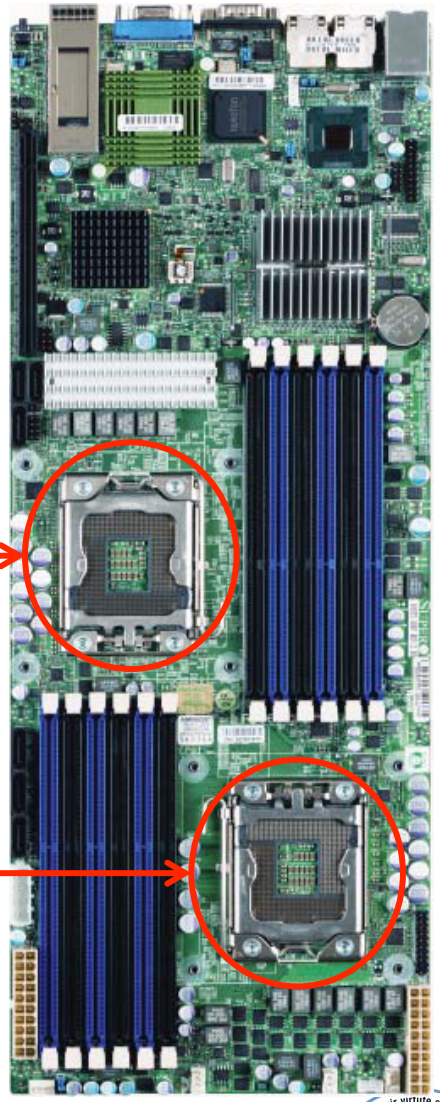


Building blocks of a cluster



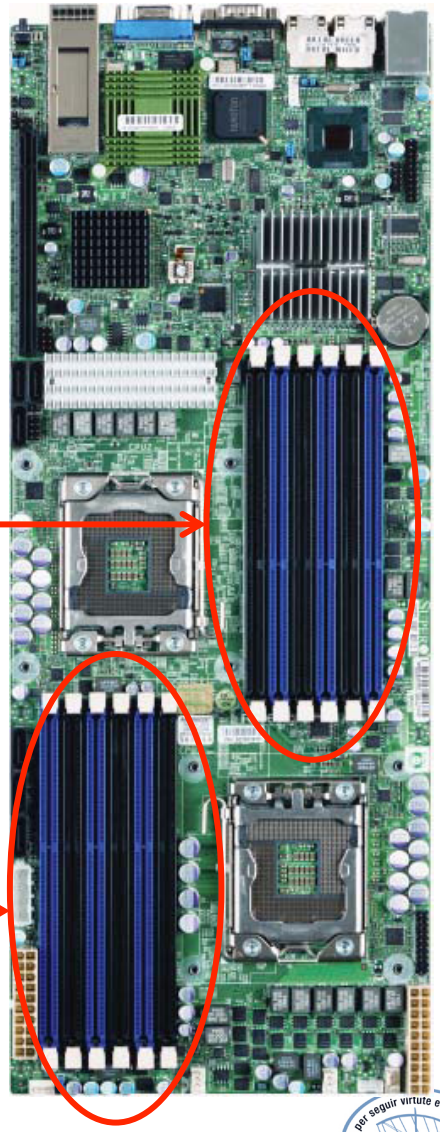
The motherboard components

Dual socket CPU



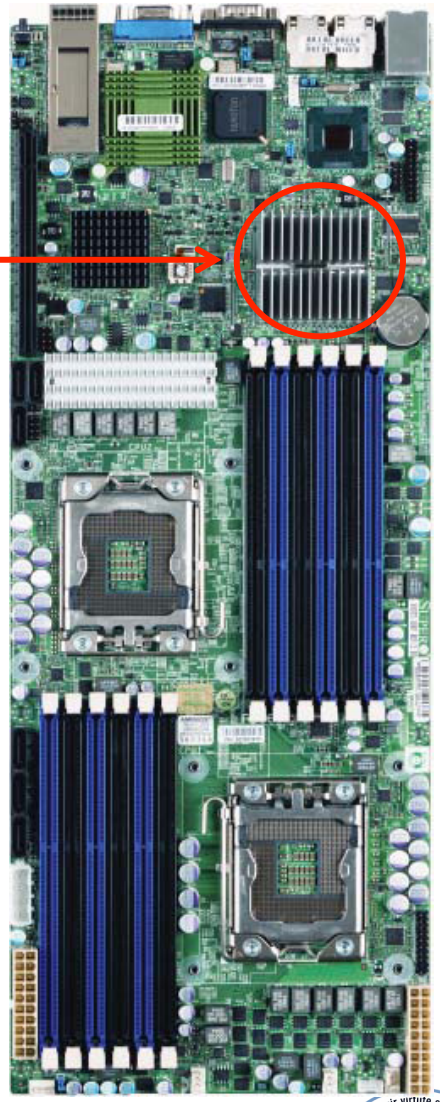
The motherboard components

DDR3 Ram 12 slots



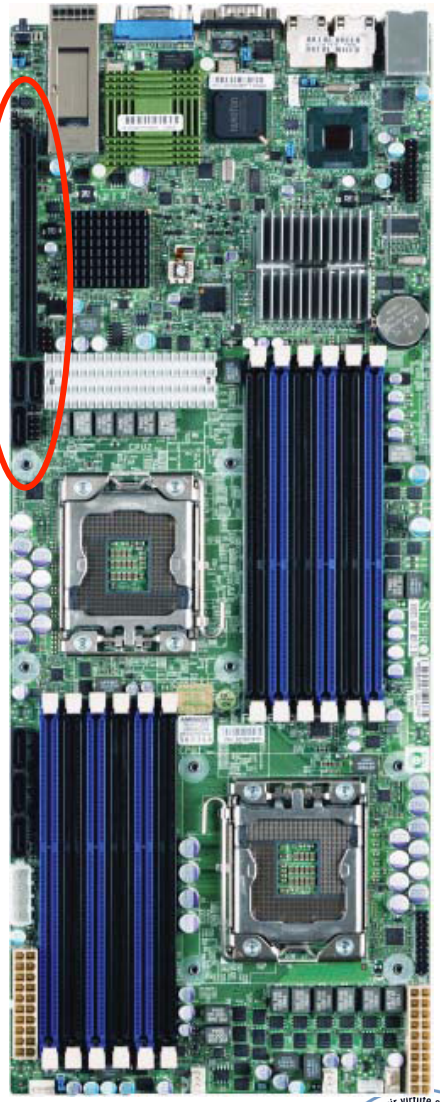
The motherboard components

Northbridge



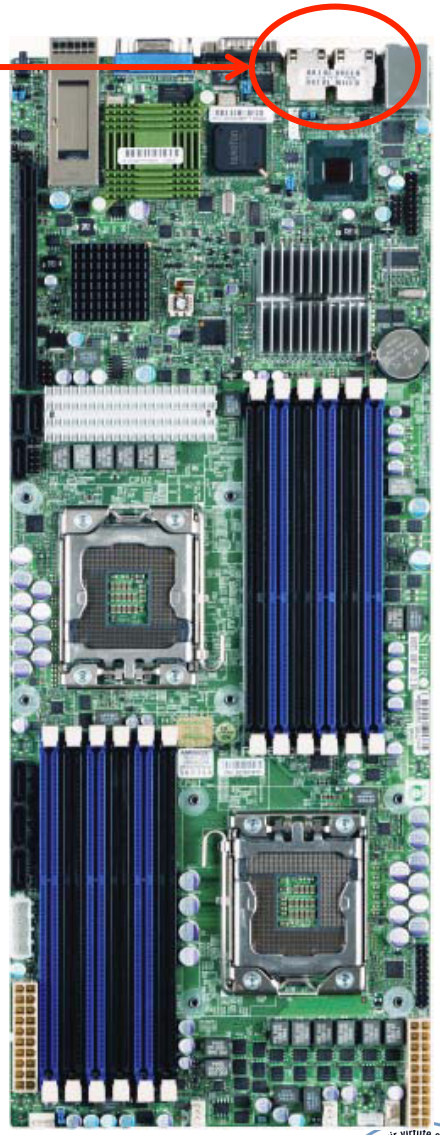
The motherboard components

PCI 8x 1 slot



The motherboard components

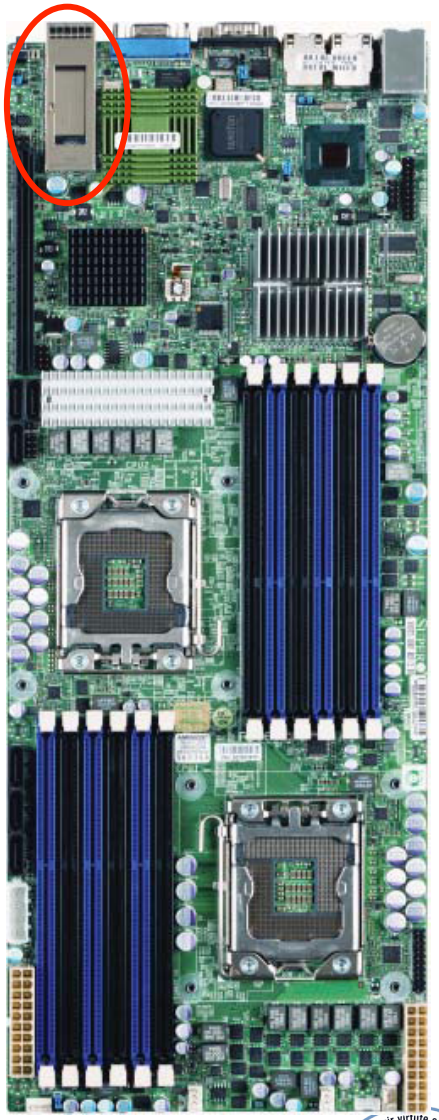
Gb Ethernet dual port



MHPC
Master in High Performance Computing

The motherboard components

Infiniband 4x QDR



ma per seguir virtute e canoscenza
SISSA

ICTP

MHPC
Master in High Performance Computing

Motherboard components

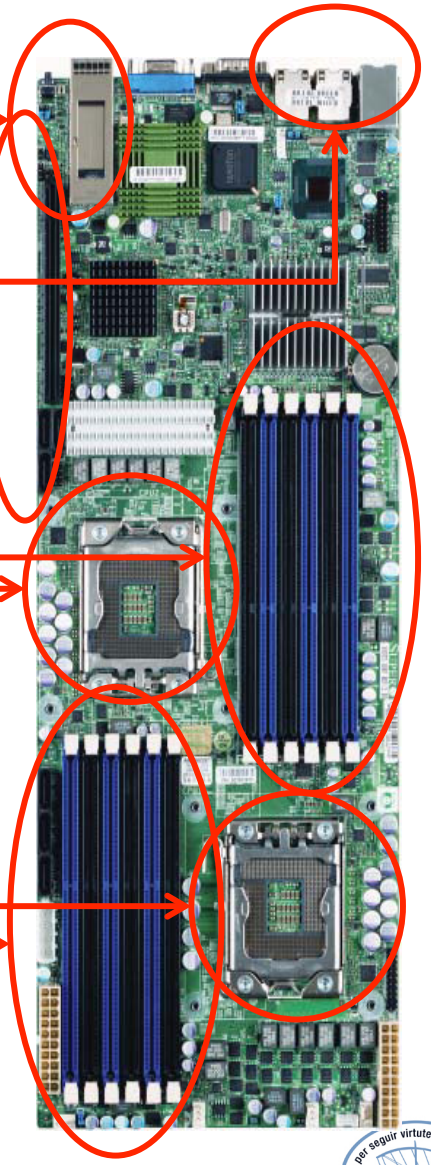
Gb Ethernet dual port

PCI 8x 1 slot

Infiniband 4x QDR

Dual socket CPU

DDR3 Ram 12 slots



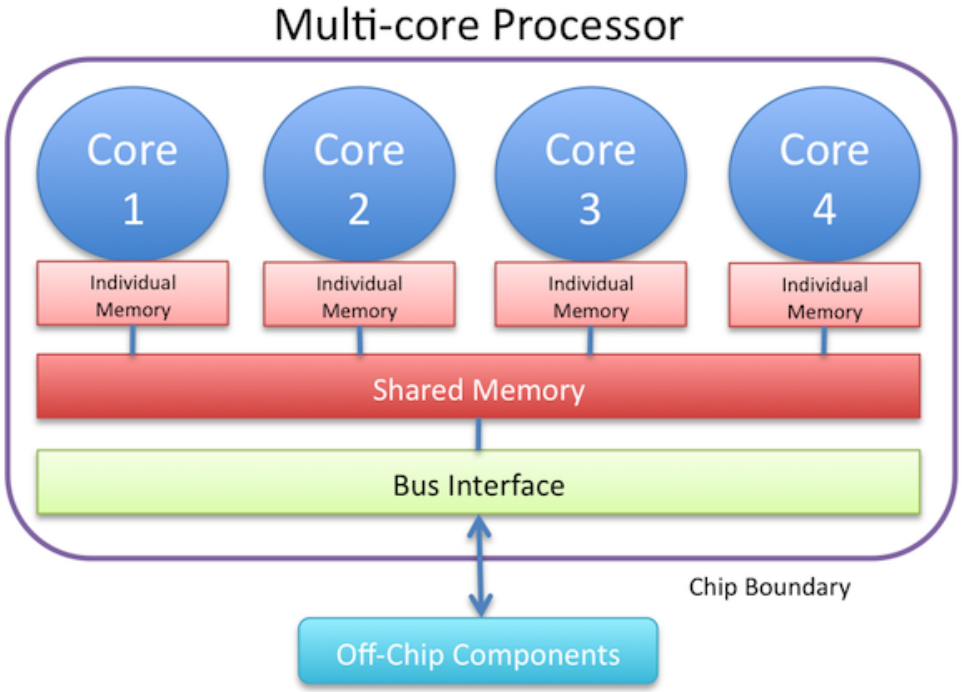
Slides from P.Altoe E4company

SISSA

ICTP

Modern CPU are multicore

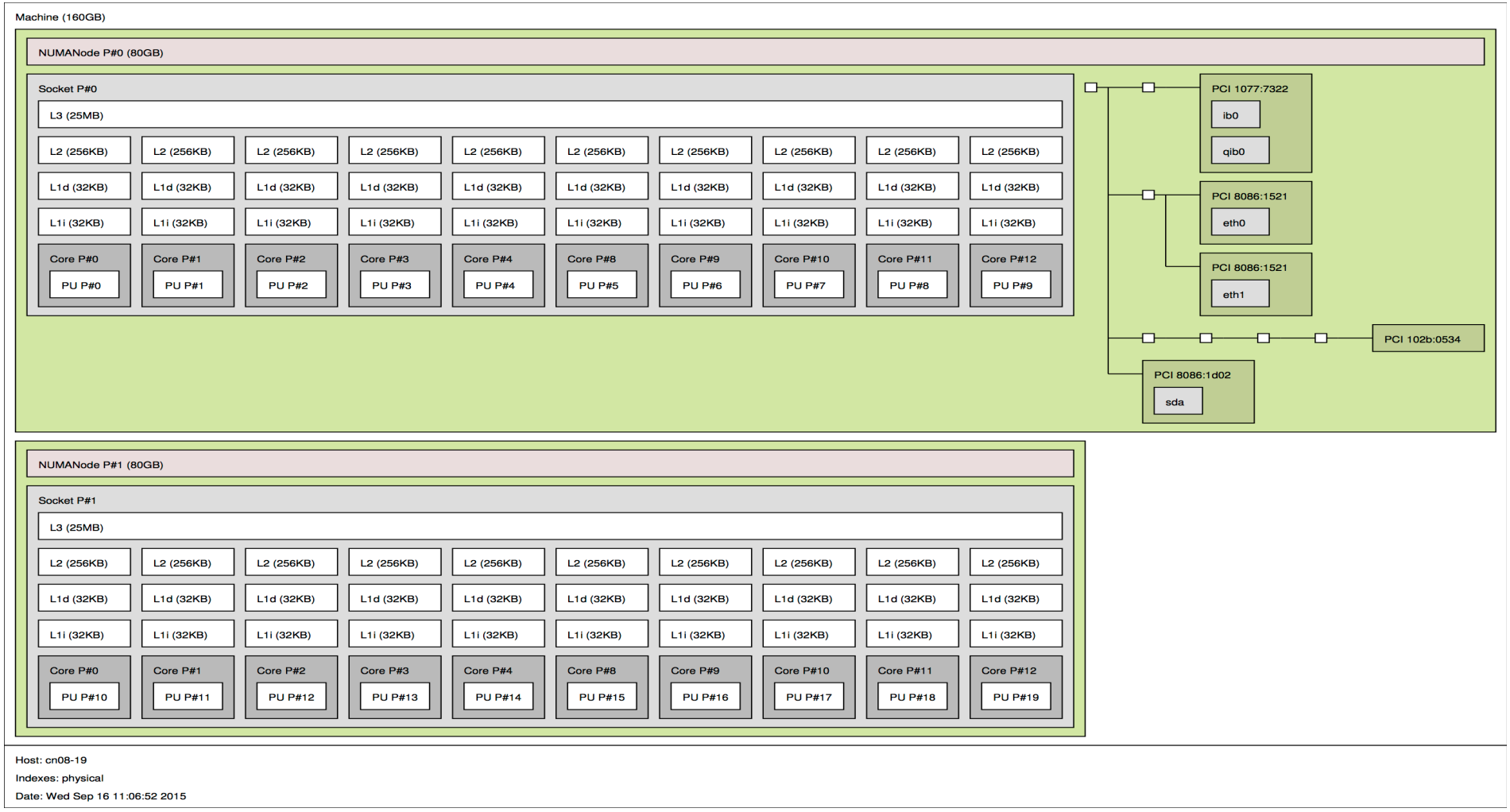
- Because of power, heat dissipation, etc increasing tendency to actually lower clock frequency but pack more computing cores onto a chip.
- These cores will share some resources, e.g. memory, network, disk, etc but are still capable of independent calculations



Picture from: http://www.cse.wustl.edu/~jain/cse567-11/ftp/multcore/

MHPC
Master in High Performance Computing

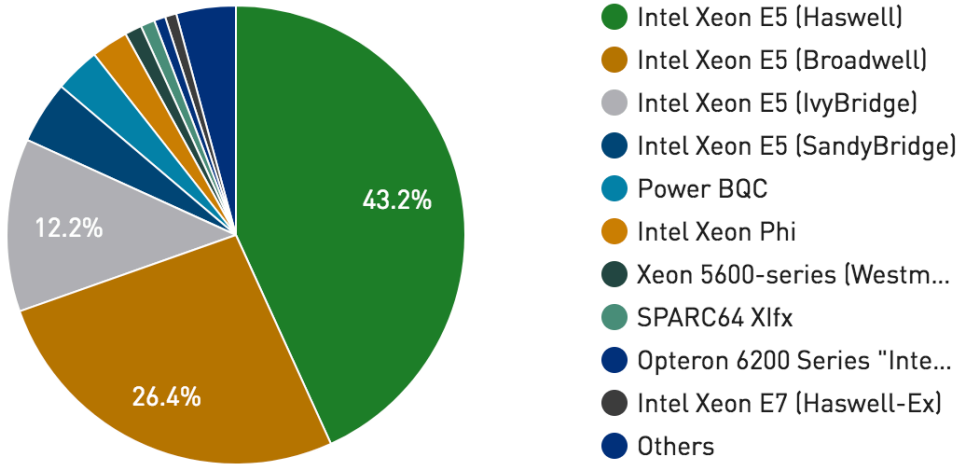
A modern node picture (Ullysses node)



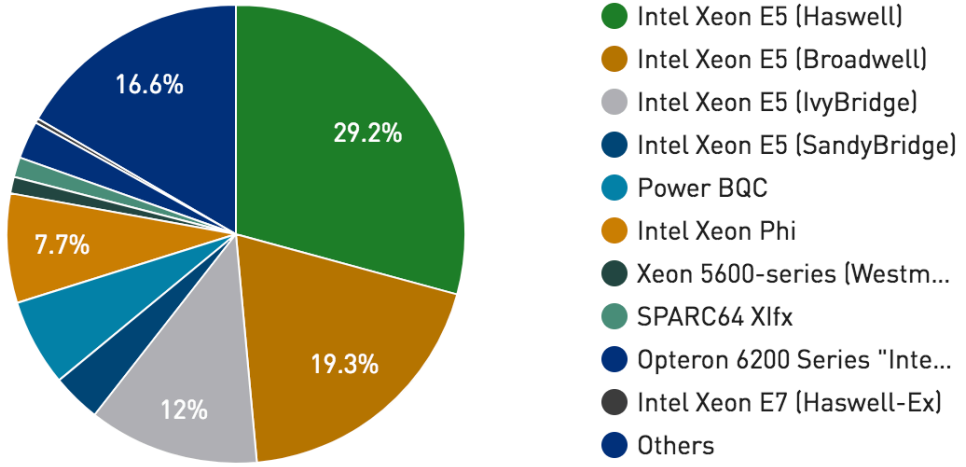
MHPC
Master in High Performance Computing

Which CPUs on TOP500 system ?

Processor Generation System Share



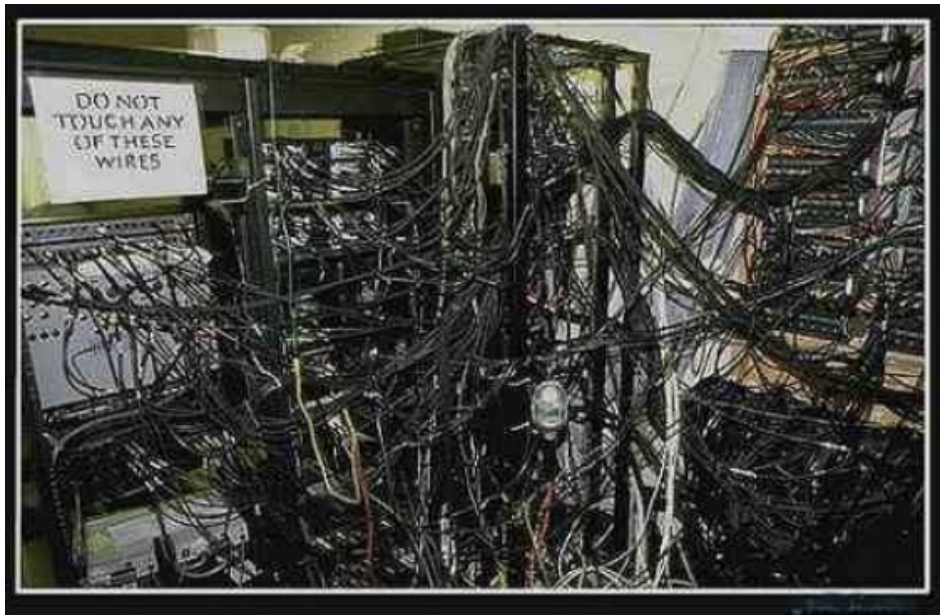
Processor Generation Performance Share



MHPC
Master in High Performance Computing

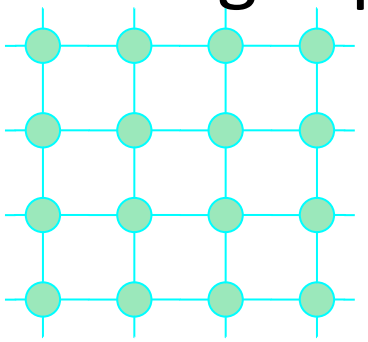
About network for cluster

- 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.

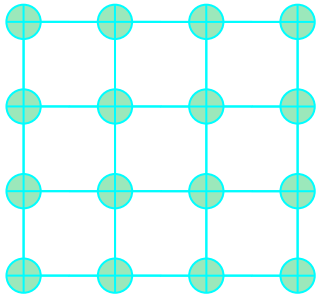


MHPC
Master in High Performance Computing

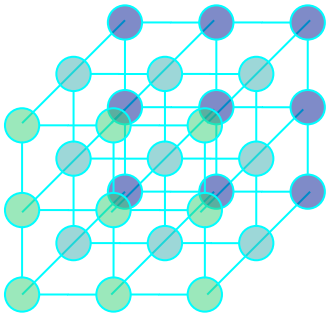
Clustering topology



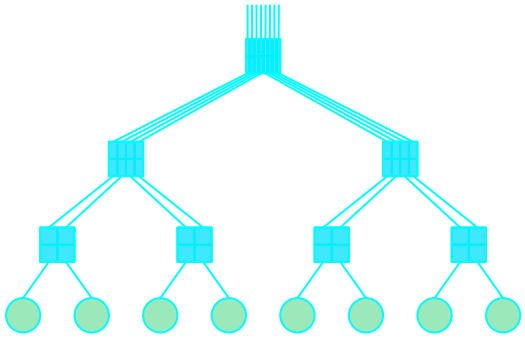
2D Mesh



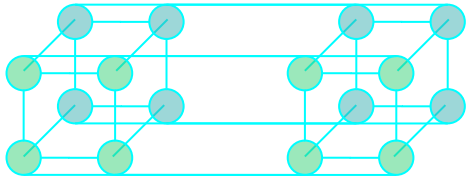
2D Torus



3D Mesh



FAT TREE



Hypercube (4-cube)

SISSA

ICTP

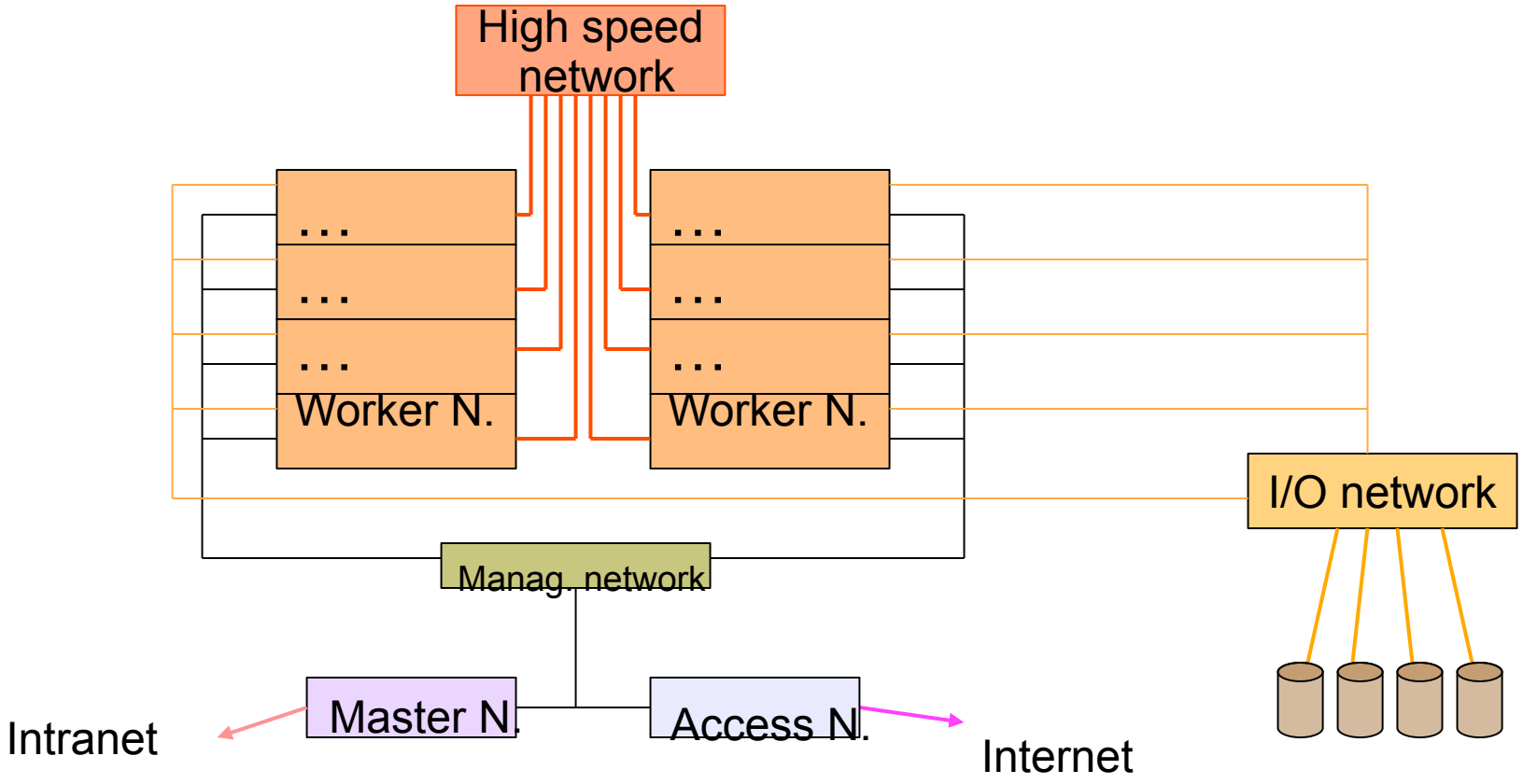
Latency&bandwidth

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

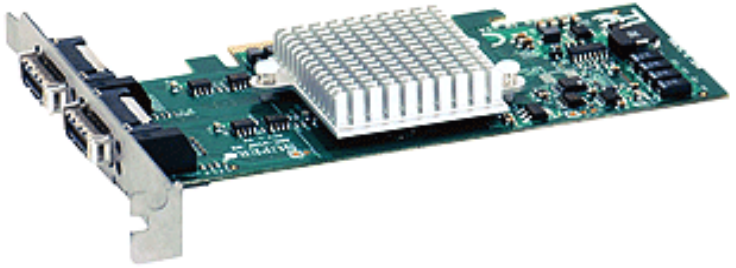
What is the UNIT OF MEASURE OF LATENCY ?

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

HPC cluster logical structure..



HPC cluster: 3 kind of network

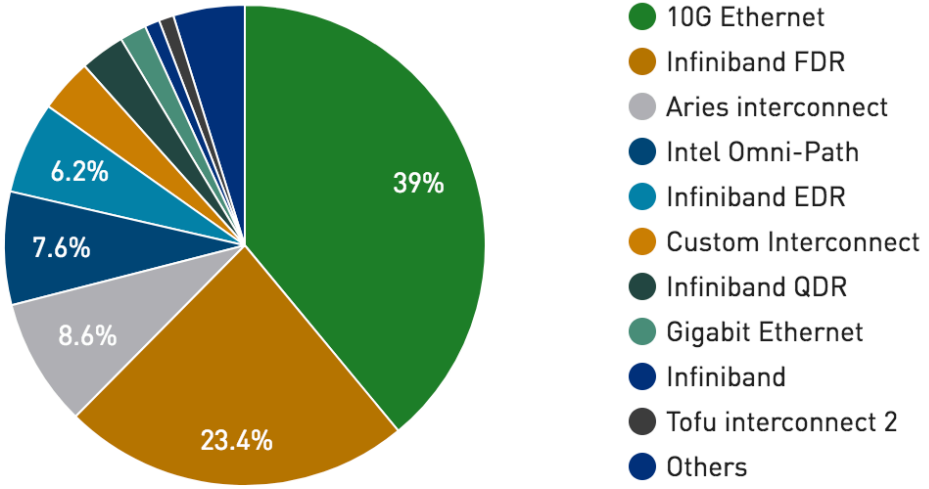


- HIGH SPEED NETWORK
  - parallel computation
  - low latency /high bandwidth
  - Usual choices: Infiniband…
- I/O NETWORK
  - I/O requests (NFS and/or parallel FS)
  - latency not fundamental/ good bandwidth
  - GIGABIT is ok
- Management network
  - management traffic
  - any standard network

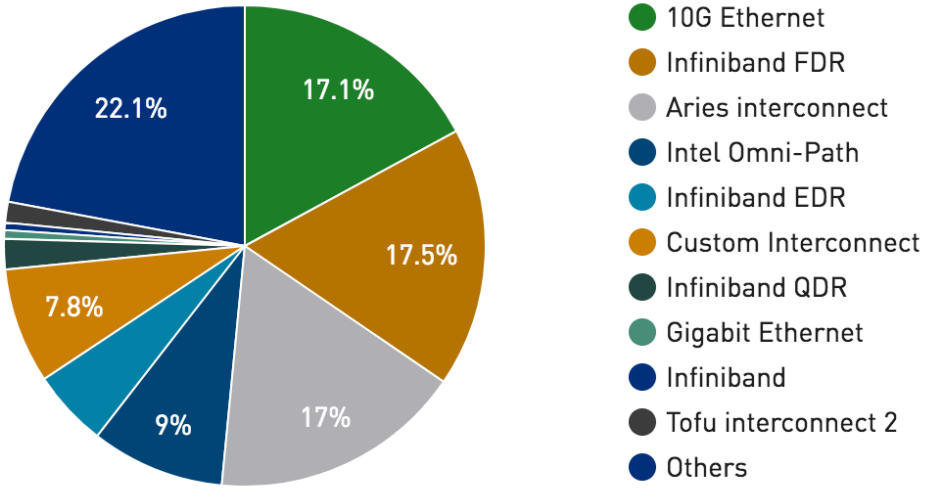
MHPC
Master in High Performance Computing

Network in Top500

Interconnect System Share



Interconnect Performance Share



Accelerators: GPU

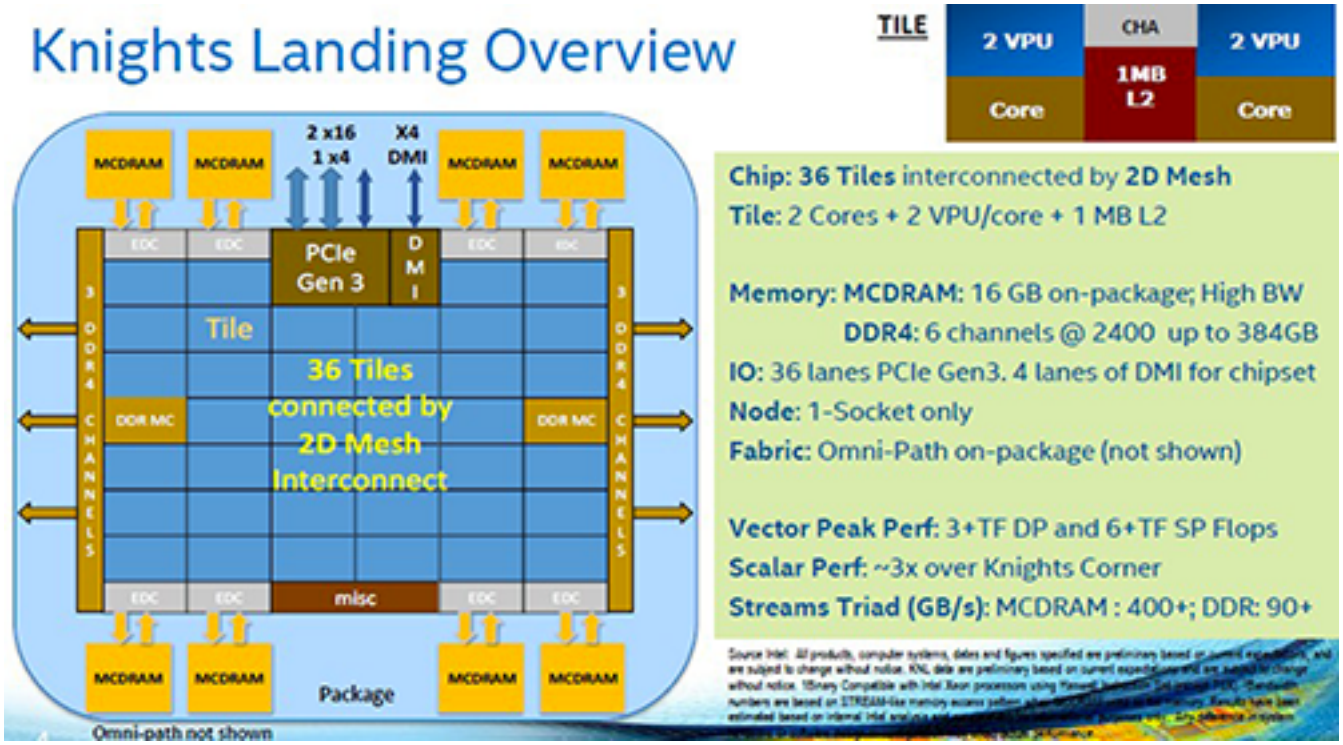
- Co-processors or accelerators have been around for a while
- Big burst in its adoption in HPC when Nvidia released CUDA (2006).
- GPGPUs or simply GPUs work in a different way to conventional CPUs. Emphasis on stream processing.
- Acceleration can be significant but depends on application.
- Nvidia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

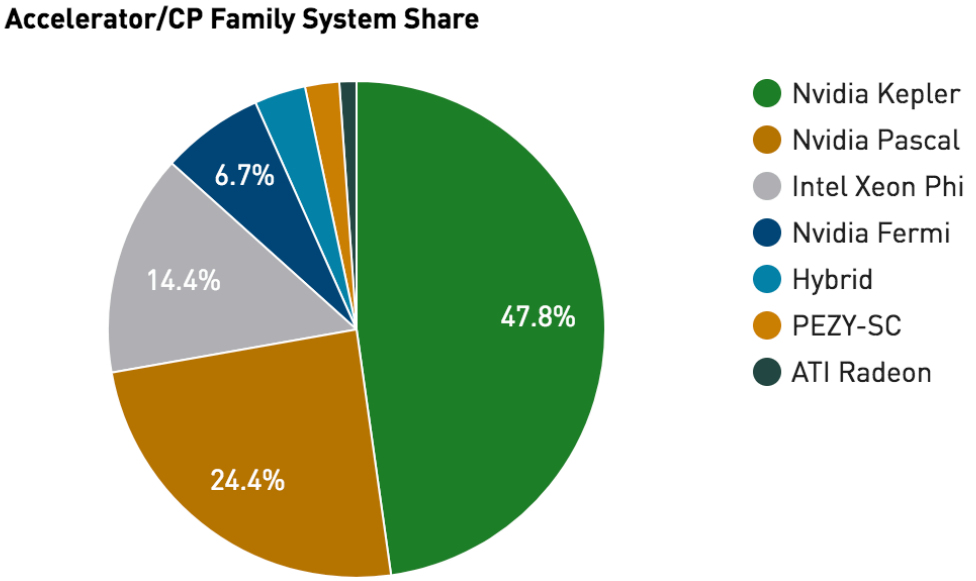
Accelerators: Intel PHI (MIC)

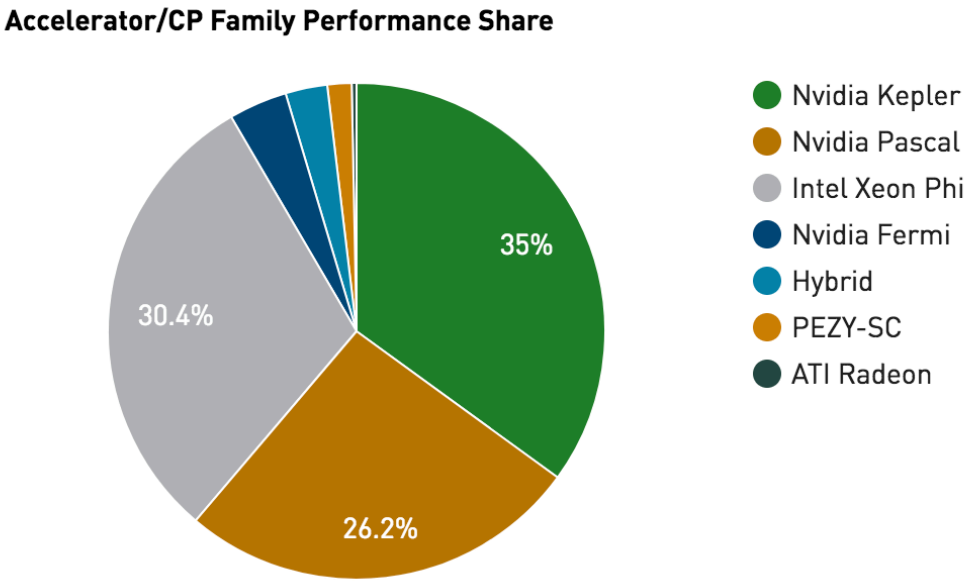
- Also an accelerator but more similar to a conventional multicore CPU.
- • Cores connected in a ring topology.
- • No need to write CUDA or OpenCL as Intel compilers will compile Fortran or C code for the MIC.



MHPC
Master in High Performance Computing

Accelerators in Top500





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).

MHPC
Master in High Performance Computing

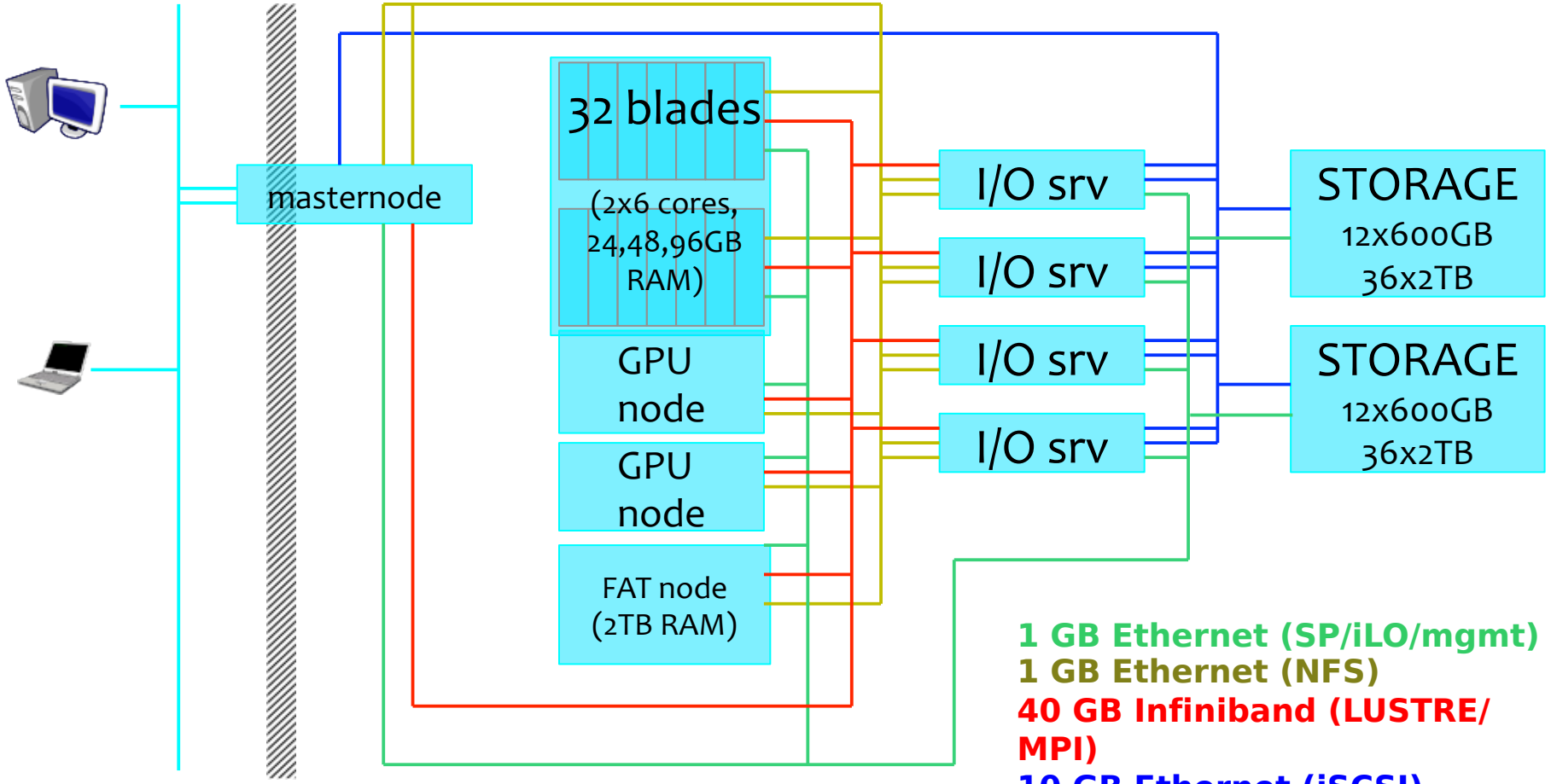
Last but not least: Storage

- 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

SISSA ICTP

MHPC
Master in High Performance Computing

Cluster example



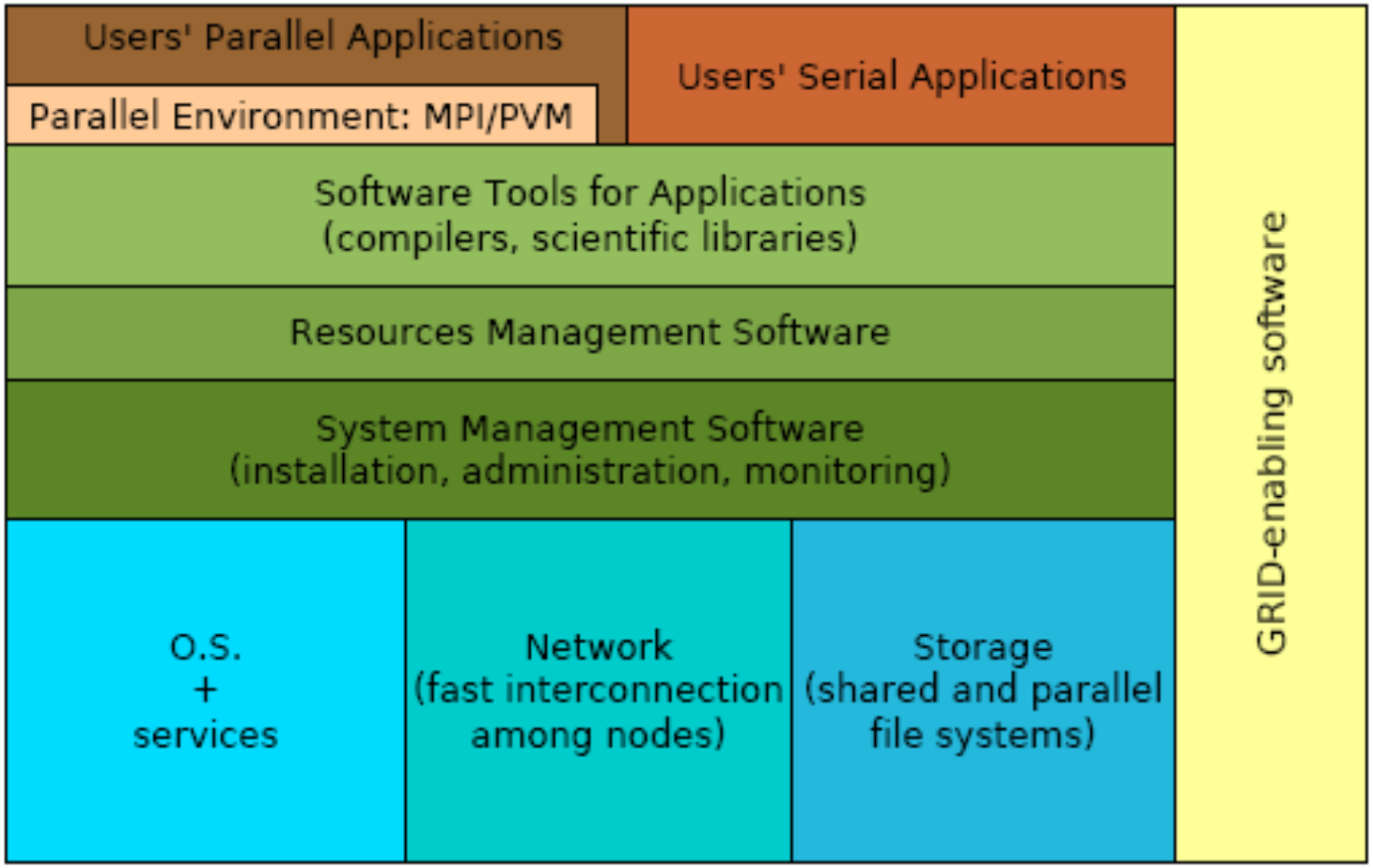
1 GB Ethernet (SP/iLO/mgmt)
1 GB Ethernet (NFS)
40 GB Infiniband (LUSTRE/ MPI)
10 GB Ethernet (iSCSI)
1 GB (LAN)

SISSA
ICTP

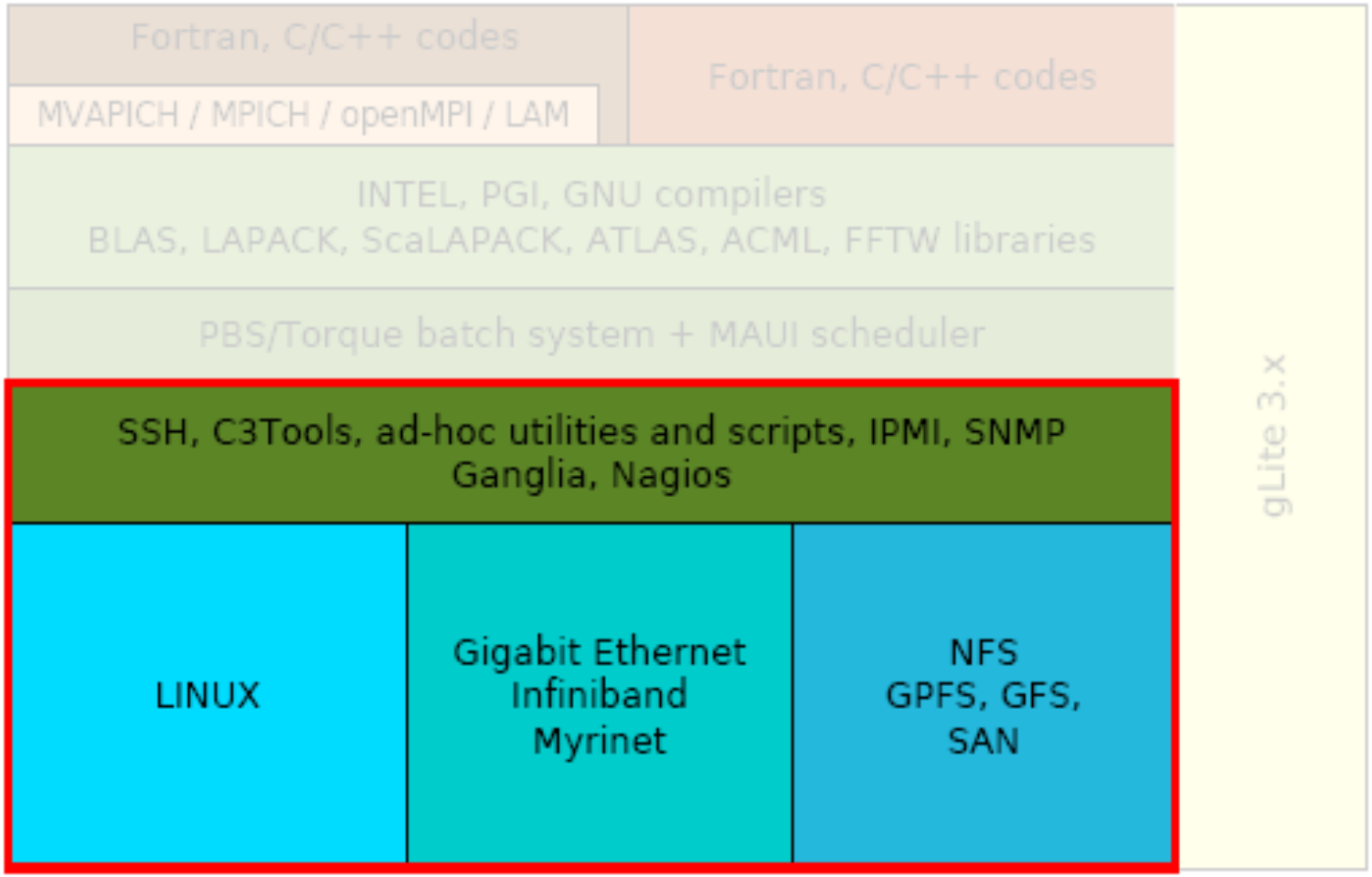
MHPC
Master in High Performance Computing

SOFTWARE

HPC platform: the software stacks

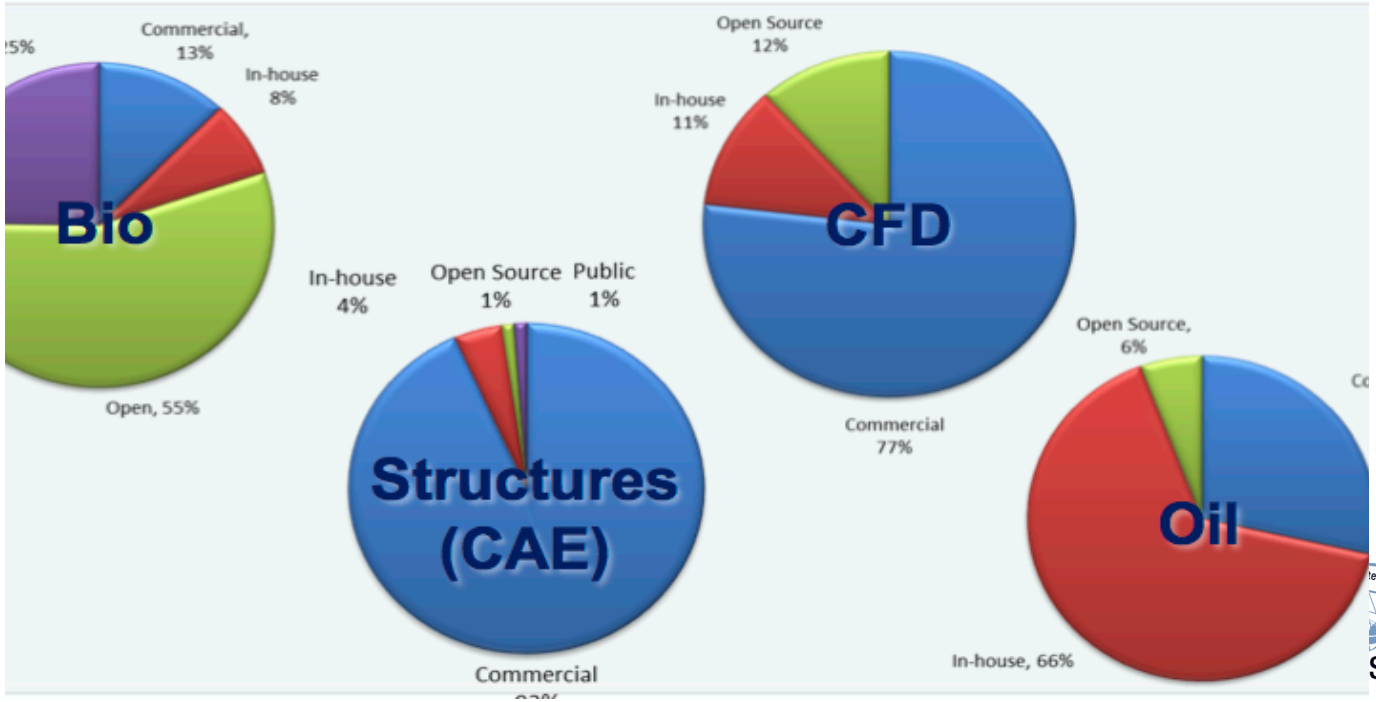


the software stacks: sys. adm...



Software Components

- MIDDLEWARE:
  - To run/manage the HPC resources
- TECHNICAL SOFTWARE:



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

MHPC
Master in High Performance Computing

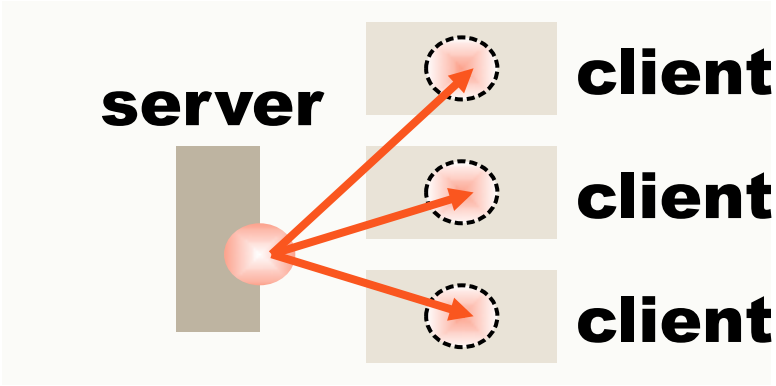
Cluster middleware

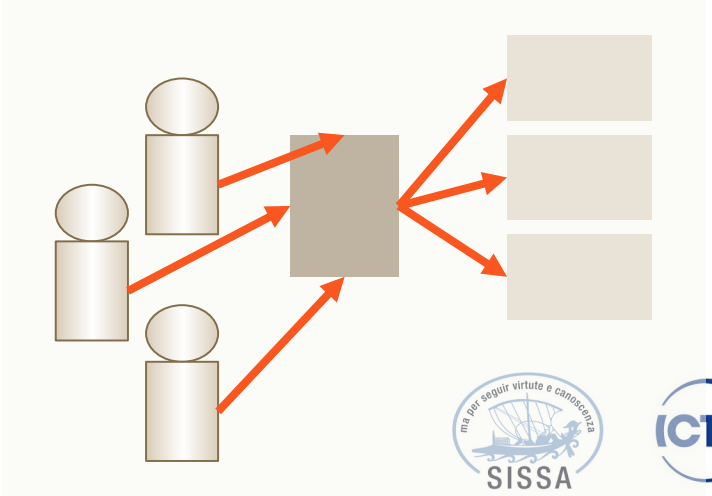
Administration software:

- user accounts
- NTP/NFS/ etc…

Resource management and scheduling software (LRMS)

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





Parallel Programming Paradigms

Two basic schemes for parallel programming dictated by hardware:

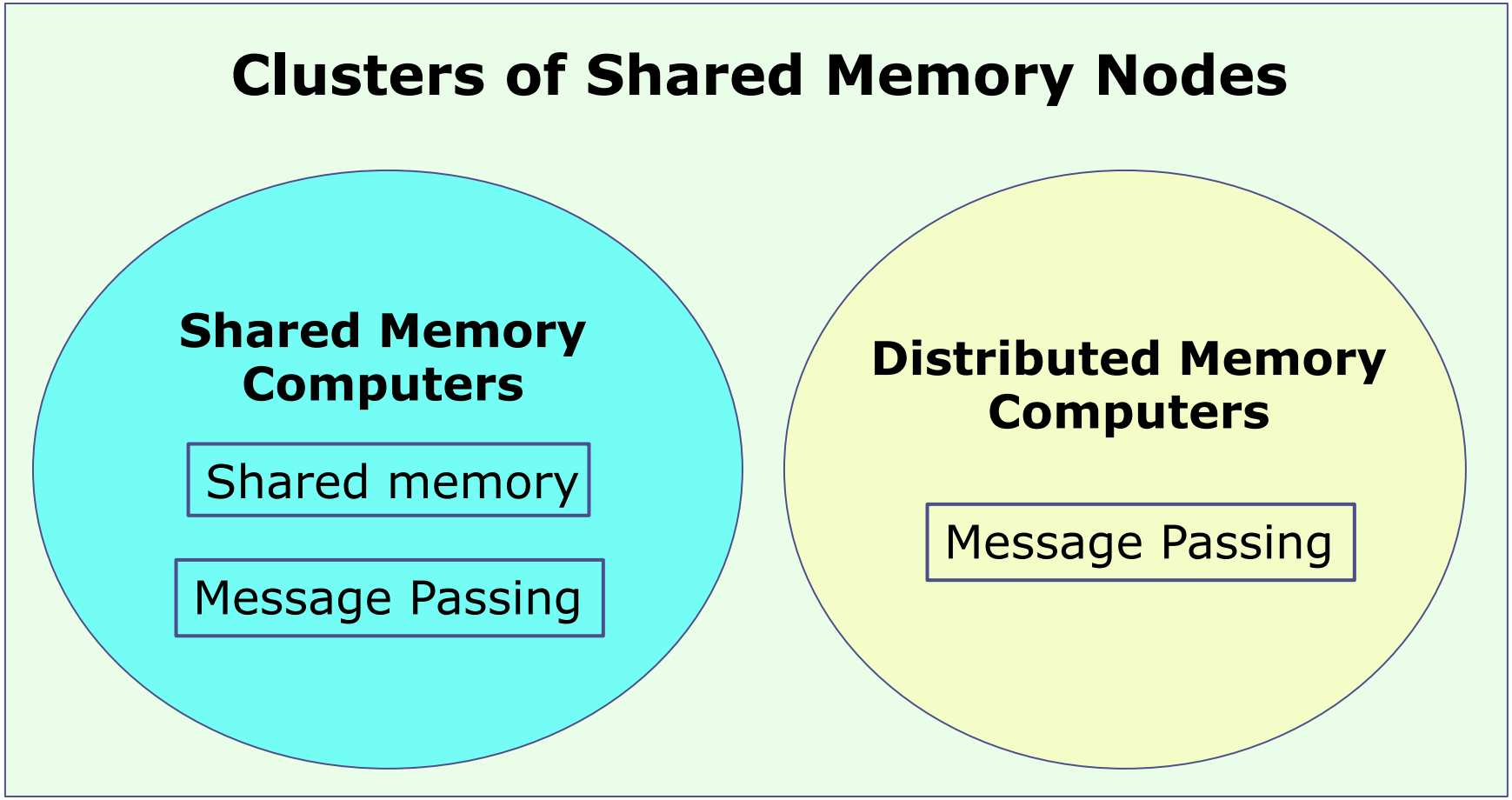
Shared memory

Single memory view, all processes (usually threads) could **directly**access the whole memory

Message Passing (distributed memory)

all processes could **directly** access only their local memory

Architectures vs. Programming Paradigms



How to program with shared memory ?

- Automatic (implicit) parallelization:
  - -compilers do (?) the job for you
- Manual parallelization:
  - -Insert parallel directives by yourself to help compilers
  - -OpenMP THE standard
- Multi threading programming:
  - -more complex but more efficient
  - -use a threads library to create task by yourself
- Use already threaded libraries..

MHPC
Master in High Performance Computing

How to program using Message Passing ?

Using the de-facto standard : MPI message passing interface

- -A standard which defines how to send/receive message from a different processes

Many different implementation

- -OpenMPI
- -Intel-MPI
- -They all provide a library which provide all communication routines

To compile your code you have to link against a library

- -Generally a wrapper is provided (mpif90/mpicc)

SISSA
ICTP

Other paradigm are now available

Mixed/hybrid approach..

- -MPI + OpenMP

Specific SDK for specific devices

- -CUDA for Nvdia GPU

Write once run everywhere:

- -OpenCL
- -OpenACC:

OpenCL

OpenACC is about giving programmers a set of tools to port their codes to new heterogeneous system without having to rewrite the codes in proprietary languages.

MHPC
Master in High Performance Computing

Principle of parallel computing

- Speedup, efficiency, and Amdahl's Law
- Finding and exploiting parallelism
- Finding and exploiting data locality
- Load balancing
- Coordination and synchronization
- Performance modeling

All of these things make parallel programming more difficult than sequential programming.

SISSA ICTP

Speed up

The *speedup* of a parallel application is

Speedup(p) = Time(1)/Time(p)

where

Time(1) = execution time for a single processor

Time(p) = execution time using p parallel processor

If Speedup(p) = p we have *perfect speedup* (also called *linear scaling*)

speedup compares an application with itself on one and on p processors

more useful to compare

The execution time of the best serial application on 1 processor

versus

The execution time of best parallel algorithm on p processors

Efficiency

The *parallel efficiency* of an application is defined as

$$\text{Efficiency}(p) = \text{Speedup}(p)/p$$

- Efficiency(p) <= 1
- For perfect speedup Efficiency (p) = 1

We will rarely have perfect speedup.

- Lack of perfect parallelism in the application or algorithm
- Imperfect load balancing (some processors have more work)
- Cost of communication
- Cost of contention for resources, e.g., memory bus, I/O
- Synchronization time

Understanding why an application is not scaling linearly will help finding ways improving the applications performance on parallel computers.

Superlinear speedup

Question: can we find "*superlinear*" speedup, that is

Speedup(p) > p ?

Choosing a bad "baseline" for T(1) WRONG !!!

Old serial code has not been updated with optimizations

Shrinking the problem size per processor GOOD

- May allow it to fit in small fast memory (cache)

Amdahl's law

Suppose only part of an application runs in parallel

- -Let s be the fraction of work done serially,
- -So (1-s) is fraction done in parallel
- -What is the maximum speedup for P processors?

Speedup(p) = T(1)/T(p)

T(p) = (1-s)*T(1)/p +s*T(1)
T(p) = T(1)*((1-s) + p*s)/p

assumes perfect speedup for parallel part

$$Speedup(p) = p/(1 + (p-1)*s)$$

Even if the parallel part speeds up perfectly,
we may be limited by the sequential portion of code.

Amdahal's law

Which fraction of serial code is it allowed ?

>	2	4	8	32	64	256	512	1024
5%	1.91	3.48	5.93	12.55	15.42	18.62	19.28	19.63
2%	1.94	3.67	6.61	16.58	22.15	29.60	31.35	32.31
1%	1.99	3.88	7.48	24.43	39.29	72.11	83.80	91.18

What about Scalability ???

Problem scaling

- Amdahl's Law is relevant only if serial fraction is independent of problem size, which is rarely true
- Fortunately "The proportion of the computations that are sequential (non parallel) normally decreases as the problem size increases " (a.k.a. Gustafon's Law)

So What Is Scalability?

- to get N times more work done on N processors
- compute a fixed-size problem N times faster
  - Strong scaling
    - Speedup S = T1 / TN ; linear speedup occurs when S = N
    - Can't achieve it due to Amdahl's Law (no speedup for serial parts)
- compute a problem N times bigger in the same amount of time:
  - Weak scaling
    - Speedup depends on the amount of serial work remaining constant or increasing slowly as the size of the problem grows
    - Assumes amount of communication among processors also remains constant or grows slowly

MHPC
Master in High Performance Computing

Why weak scaling tends to work better..

Strong scaling: fixed data/problem set; measure speedup with more processors

-Ahmdal law

Weak scaling: data/problem set increases with more processors; measure if speed(efficiency) is the same

-Gustafson law

Exercise in the afternoon:
evaluate strong/weak scalability of a toy parallel code.

SISSA ICTP

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 lecture/courses where all what we discussed today will be analyzed in details

Lesson learned

- Nobody is aware of TCO (especially in academic world)
- Very few have a clear idea of the computational tasks and associated requirements..
- HPC just still means Performance not Productivity or Profitability