

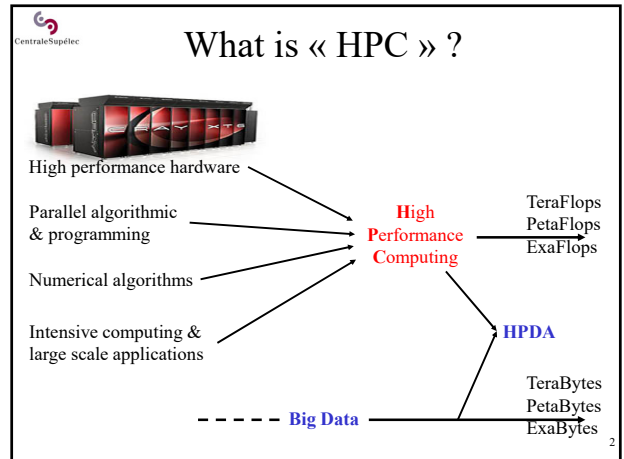
CentraleSupélec 

SG6: High Performance Computing

## Introduction to HPC hardware issues

**Stéphane Vialle**  
 Université Paris-Saclay Sciences et technologies de l'information et de la communication (STIC)  


Stephane.Vialle@centralesupelec.fr  
<http://www.metz.supelec.fr/~vialle>



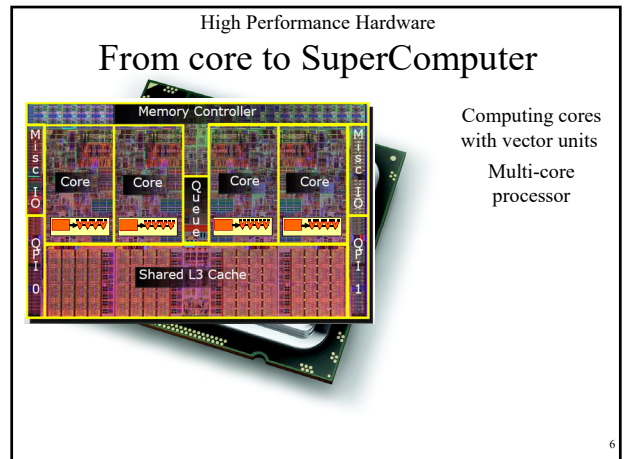
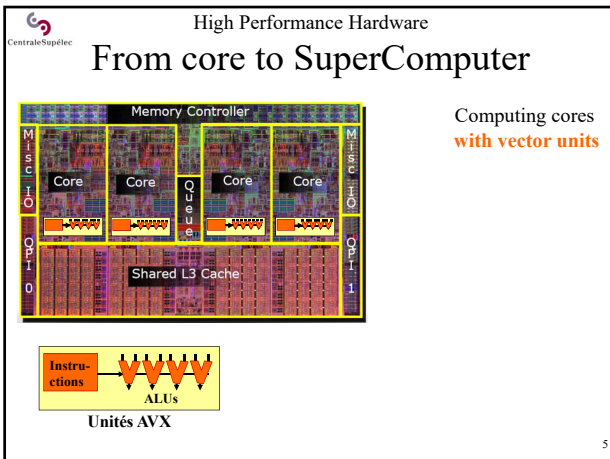
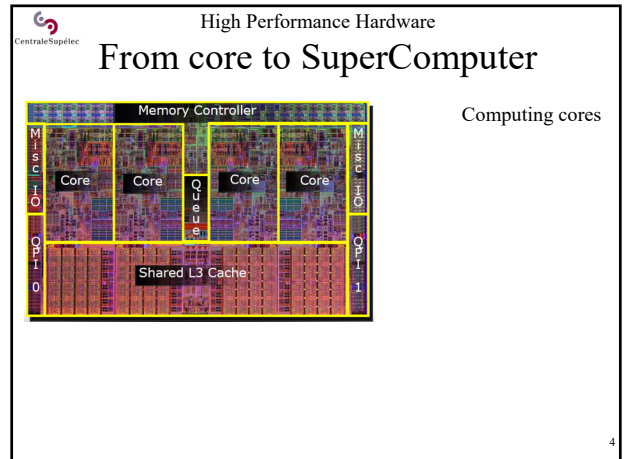
CentraleSupélec

## High Performance Hardware

Inside ....

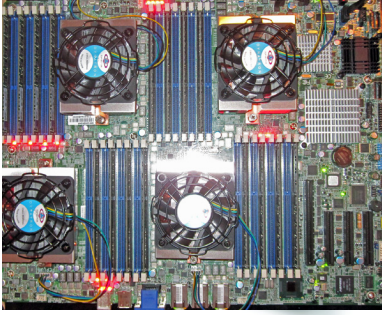


... high performance hardware



High Performance Hardware

## From core to SuperComputer

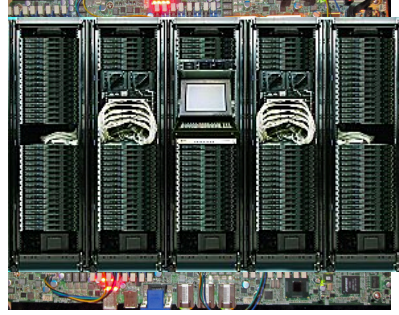


- Computing cores with vector units
- Multi-core processor
- Multi-core PC/node

7

High Performance Hardware

## From core to SuperComputer

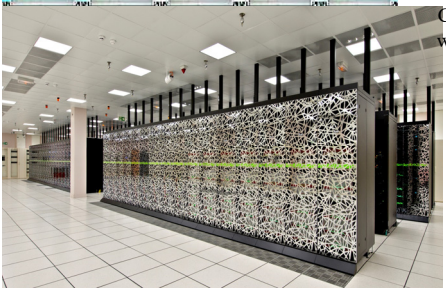


- Computing cores with vector units
- Multi-core processor
- Multi-core PC/node
- Multi-core PC cluster

8

High Performance Hardware

## From core to SuperComputer

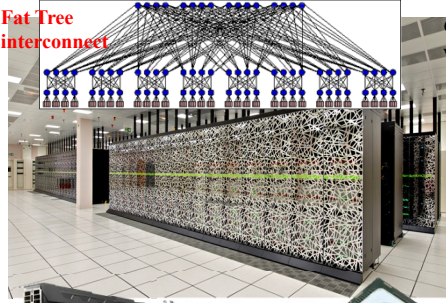


- Computing cores with vector units
- Multi-core processor
- Multi-core PC/node
- Multi-core PC cluster
- Super-Computer

9

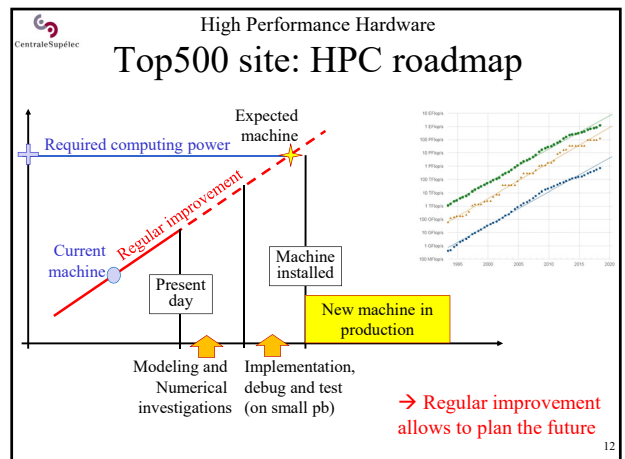
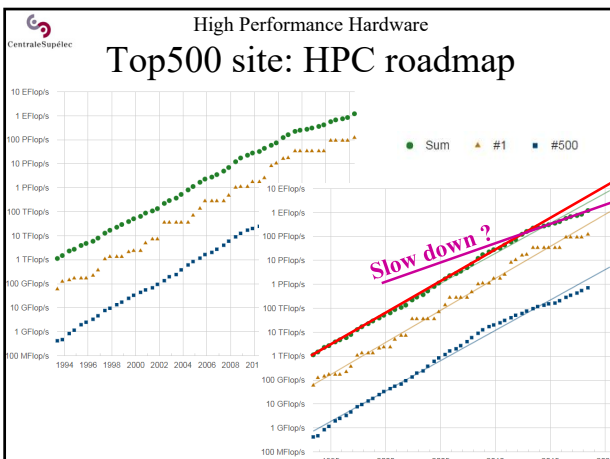
High Performance Hardware

## From core to SuperComputer



- Computing cores with vector units
- Multi-core processor
- Multi-core PC/node
- Multi-core PC cluster
- Super-Computer
- + hardware accelerators

10



### High Performance Hardware HPC in the cloud ?

CentraleSupélec

**Microsoft Azure**  
« AZUR BigCompute » : High performance nodes  
High performance interconnect (Infiniband)  
→ Customers can allocate a part of a HPC cluster

**aws**  
• Allows to allocate a huge number of nodes for a short time  
• No high performance interconnection network  
→ Comfortable for Big Data scaling benchmarks

**Google Cloud**  
• Compliant with very large scale systems and applications


Some HPC or Large Scale PC-clusters exist in some Clouds  
**But no SuperComputer available in a cloud**

13

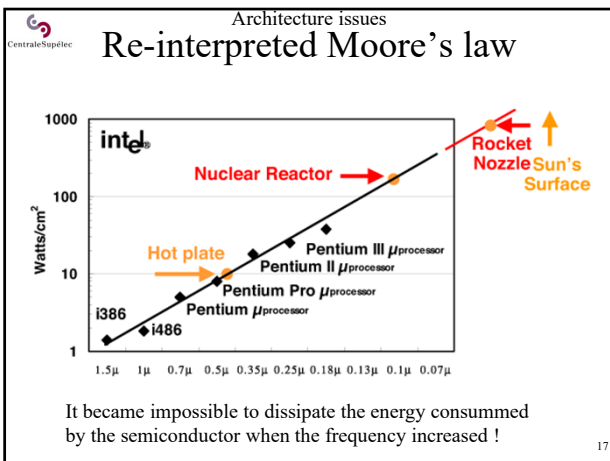
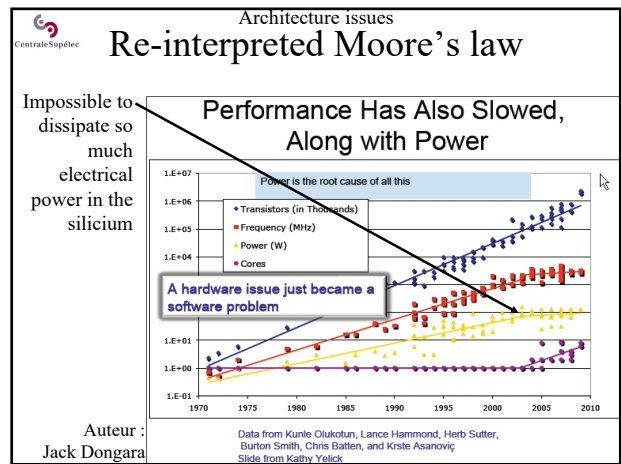
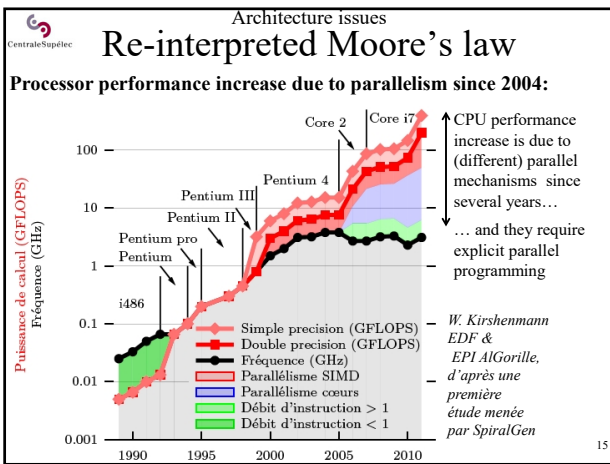
### Architecture issues

CentraleSupélec

Why multi-core processors ?      Shared or distributed memory ?



14



### Architecture issues Re-interpreted Moore's law

#### Power Cost of Frequency

To bound frequency and to increase the nb of cores is energy efficient

- Power  $\propto$  Voltage<sup>2</sup> x Frequency (V<sup>2</sup>F)
- Frequency  $\propto$  Voltage
- Power  $\propto$  Frequency<sup>3</sup>

	Cores	V	Freq	Perf	Power	PE (Bq/μm²)
Superscalar	1	1	1	1	1	1
"New" Superscalar	1X	1.5X	1.5X	1.5X	3.3X	0.45X
Multicore	2X	0.75X	0.75X	1.5X	0.8X	1.88X

$2 \times 0.75^3 = 0.8$

50% more performance with 20% less power  
Preferable to use multiple slower devices, than one superfast device

(Bigger # is better)

Auteur : Jack Dongara

Architecture issues  
**Re-interpreted Moore's law**

**Initial (electronic) Moore's law:**  
each 18 months  $\rightarrow$  x2 number of transistors per  $\mu\text{m}^2$

**Previous computer science interpretation:**  
each 18 months  $\rightarrow$  x2 processor speed

**New computer science interpretation:**  
each 24 months  $\rightarrow$  x2 number of cores

**Leads to a massive parallelism challenge:**  
to split many codes in 100, 1000, .....  $10^6$  threads ...  $10^7$  threads!!

Year	Average Number of Cores
2000	10,000
2001	15,000
2002	20,000
2003	30,000
2004	40,000
2005	60,000
2006	80,000
2007	120,000
2008	180,000
2009	250,000

Architecture issues  
**3 classic parallel architectures**

**Shared-memory machines (Symetric MultiProcessor):**

One principle:  
- several implementations,  
- different costs,  
- different speeds.

Overview of Recent Supercomputers  
Aad J. van der Steen  
Jack J. Dongarra<sup>20</sup>

Architecture issues  
**3 classic parallel architectures**

**Distributed-memory machines (clusters):**

Cluster basic principles, but cost and speed depend on the interconnection network!

Gigabit Ethernet

Highly scalable architecture

21

Architecture issues  
**3 classic parallel architectures**

**Distributed Shared Memory machines (DSM):**

cache coherence Non Uniform Memory Architecture (ccNUMA)  
Extends the cache mechanism

Up to 1024 nodes  
Supports global multithreading

Hardware implementation: fast & expensive...  
Software implementation: slow & cheap!

Overview of Recent Supercomputers  
Aad J. van der Steen  
Jack J. Dongarra

22

Architecture issues  
**3 classic parallel architectures**

- Shared memory « SMP »: Simple and efficient up to ... 16 processors. Limited solution
- Distributed memory « Cluster »: Unlimited scalability. But efficiency and price depend on the interconnect.
- Distributed shared memory « DSM »: Comfortable and efficient solution. Efficient hardware implementation up to 1000 processors.

**2019 : all supercomputers are clusters with a hierarchical architecture**

23

Architecture issues  
**Modern clusters**

One Non-Uniform Memory Architecture node

Cluster of nodes

Cluster of NUMA nodes



Architecture issues

## Hierarchical architecture

Cluster of NUMA nodes

Time to access a data in the RAM of the processor

- < Time to access a data in a RAM attached to another processor (several access times exist inside one NUMA node)
- < Time to send/recv a data to/from another node

→ {data – thread} **co-location** become critic

Architecture issues

## Hierarchical & hybrid architecture

Cluster of multi-processor NUMA nodes with hardware vector accelerators

- Hierarchical and hybrid architectures
- Multi-paradigms programming

*Development problem*

AVX vectorization:  
#pragma simd

+ CPU multithreading:  
#pragma omp parallel for

+ Message passing:  
MPI\_Send(..., Me-1,...);  
MPI\_Recv(..., Me+1,...);

+ GPU vectorization  
myKernel<<<grid,bloc>>>(...)

+ **checkpointing**

26

Architecture issues

## Hierarchical & hybrid architecture

Distributed Software Architecture

Need to achieve an efficient mapping of software and hardware resource:

- One process per node ? per processor ? per core ?
- How many threads per process ? per core ?
- How to distributed the processes and threads to load balance the computations ?
- How to group intensively communicating processes ?
- How to co-localize data and process/threads ?

*Deployment problem*

Distributed Hardware Architecture

« Interconnect »  
(cluster interconnection network)

What is a good interconnection network for a parallel computer ?

28

Interconnect

## Main features of a computing cluster interconnect

- Bandwidth
- Latency
- Contention & saturation resilience  
many algorithms are synchronous ones: all nodes compute, and then enter a communication step at the same time
- Performances of Point-to-Point Communications
- Performances of Collective Communications  
broadcast, scatter, gather, reduce, all\_to\_all,...
- Maximum latency and bandwidth variation between 2 nodes
- Extension capability (to increase machine size)  
ex : hypercubic topology is hard/expensive to extend

Criteria are different of LAN criteria

10/25-Gigabit Ethernet → Infiniband → Proprietary networks

## Fault Tolerance in HPC

Can we run a very large and long parallel computation, and succeed ?

Can a one-million core parallel program run during one week ?

30

Fault tolerance

## Mean Time Between Failures

**MTBF definition:**

Time Between Failures = (down time + up time)

Mean time between failures = MTBF =  $\frac{\sum (\text{start of downtime} - \text{start of uptime})}{\text{number of failures}}$

1976	2010	2020-2022
CRAY-1 : 50h	USA : 8h-15days	Exaflop systems : ???

31

Fault tolerance

## Why do we need for fault tolerance ?

We use more and more cores  
Processor frequency is limited and number of cores increases

We process larger problems in constant time ! (Gustafson's law)

We use more and more cores during the same time  
→ Probability of failure increases ? (yes...)

Current solution:  
« checkpoint-restart »

→ Large parallel applications could never end ?

32

Energy Consumption

1 PetaFlops: 2.3 MW !  
→ 1 ExaFlops : 2.3 GW !! 350 MW ! ..... 20 MW ?

Perhaps we will be able to build the machines,  
but not to pay for the energy consumption !!

33

Energy consumption

## How much electrical power for an Exaflops ?

**1.0 Exaflops should be reached close to 2020-2022:**

- 2.0 GWatts with the flop/watt ratio of 2008 Top500 1<sup>st</sup> machine
- 1.2 GWatts with the flop/watt ratio of 2011 Top500 1<sup>st</sup> machine
- 350 MWatts if the flop/watt ratio increases regularly
- 20 MWatts if we succeed to improve the architecture ? ...  
... « the maximum energy cost we can support ! » (2010)
- 2 MWatts ...  
... « the maximum cost for a large set of customers » (2014)

34

Energy consumption

## From Petaflops to Exaflops

1.03 Petaflops : June 2008

RoadRunner (IBM)  
Opteron + PowerXCell  
122440 « cores »  
500 Gb/s (IO)  
2.35 Mwatt !!!!!

122 Petaflops : juin 2018

Summit – IBM, Oak Ridge - USA  
IBM POWER9 22C 3.07GHz  
NVIDIA Volta GV100  
2 282 544 « cores »  
8.8 Mwatt

1.00 Exaflops : 2018-2020  
2020-2022  
25 Tb/s (IO)  
20 Mwatt max....

×1000 perf  
× 100 cores/node  
× 10 nodes  
× 50 IO  
× 10 energy (only × 10)

Related challenges:

- Elegant & efficient programming
- Training of large programmer teams

35

Energy consumption

## Sunway TaihuLight - China: N°1 2016 - 2017

**93.0 Pflops**

- 41 000 processors Sunway SW26010 260C 1.45GHz  
→ 10 649 600 « cores »
- Sunway interconnect:  
5-level integrated hierarchy (Infiniband like ?)

**15.4 Mwatt**

36

CentraleSupélec

## Energy consumption

### Summit - USA: N°1 June 2018

**122.3 Pflops (×1.31)**

- 9 216 processors IBM POWER9 22C 3.07GHz
- 27 648 GPU Volta GV100  
→ 2 282 544 « cores »
- interconnect: Dual-rail Mellanox EDR Infiniband

**8.8 MWatt (×0.57)**

**Flops/Watt : ×2.3**

CentraleSupélec

## Energy consumption

### Summit - USA: N°1 November 2018

**143.5 Pflops (×1.54)**

- 9 216 processors IBM POWER9 22C 3.07GHz
- 27 648 GPU Volta GV100  
→ 2 282 544 « cores »
- interconnect: Dual-rail Mellanox EDR Infiniband

**9.8 MWatt (×0.64)**

**Flops/Watt : ×2.4**

CentraleSupélec

## Appendix

39

CentraleSupélec

## Cooling

Cooling is close to 30% of the energy consumption

Optimization is mandatory!

40

CentraleSupélec

## Cooling

### Cooling is strategic !

**Des processeurs moins gourmands en énergie :**

- on essaie de limiter la consommation de chaque processeur
- les processeurs passent en mode économique s'ils sont inutilisés
- on améliore le rendement flops/watt

**Mais une densité de processeurs en hausse :**

- une tendance à la limitation de la taille totale des machines (en m<sup>2</sup> au sol)

→ **Besoin de refroidissement efficace et bon marché (!)**

**Souvent estimé à 30% de la dépense énergétique!**

CentraleSupélec

## Cooling

### Optimized air flow

**Optimisation des flux d'air : en entrée et en sortie des armoires**

- Architecture Blue Gene : haute densité de processeurs
- Objectif d'encombrement minimal (au sol) et de consommation énergétique minimale
- **Formes triangulaires ajoutées pour optimiser le flux d'air**

**IBM Blue Gene**

42

CentraleSupélec

## Cooling

### Cold doors (air+water cooling)

**On refroidit par eau une « porte/grille » dans laquelle circule un flux d'air, qui vient de refroidir la machine**

65° F CHILLED WATER RETURNING TO CHILLER AT 75° F

EC FANS

75° F ROOM AIR OUT

STANDARD OR CUSTOM SERVER RACK

75° F ROOM AIR IN

SLAB FLOOR

EXPLODED VIEW OF ACTIVE REAR DOOR RACK COOLING SYSTEM

Flow and return pipes can be fed from the top or from the bottom.

Cabinet rear

Exit air returns over the data centre at or around ambient temperature.

Cabinet front

Side view of cabinet and rear door cooler

Le refroidissement se concentre sur l'armoire.

mo

CentraleSupélec

## Cooling

### Direct liquid cooling

**On amène de l'eau froide directement sur le point chaud, mais l'eau reste isolée de l'électronique.**

- Expérimental en 2009
- Adopté depuis (IBM, BULL, ...)

Carte expérimentale IBM en 2009 (projet Blue Water)

Lame de calcul IBM en 2012 Commercialisée

Based on Standard POWER-7 Processors High Bandwidth, Strong Threads & Larger SMP Images Made to Reduce Bottlenecks and Improve Productivity

CentraleSupélec

## Cooling

### Liquid and immersive cooling

**Refroidissement par immersion des cartes dans un liquide** électriquement neutre, et refroidi.

Refroidissement liquide par immersion testé par SGI & Novec en 2014

**Cray 2 (1985)**

- 4 processeurs
- 1.9 Gflops
- Fluorocarbon

Refroidissement liquide par immersion sur le CRAY-2 en 1985

45

CentraleSupélec

## Cooling

### Extreme air cooling

**Refroidissement avec de l'air à température ambiante :**

- circulant à grande vitesse
- circulant à gros volume

→ Les CPUs fonctionnent proche de leur température max supportable (ex : 35°C sur une carte mère sans pb)

→ Il n'y a pas de refroidissement du flux d'air.

Une machine de Grid'5000 à Grenoble (la seule en Extreme Cooling)

Economique ! Mais arrêt de la machine quand l'air ambiant est trop chaud (l'été) !

46

CentraleSupélec

## Cooling

### Extreme air cooling

**Refroidissement avec de l'air à température ambiante :**

- circulant à grande vitesse
- circulant à gros volume

→ Les CPUs fonctionnent proche de leur température max supportable (ex : 35°C sur une carte mère sans pb)

→ Il n'y a pas de refroidissement du flux d'air.

Installation Ilium à CentraleSupélec à Metz (blockchain - 2018)

Economique ! Mais arrêt de la machine quand l'air ambiant est trop chaud (l'été) !

47

CentraleSupélec

## Introduction to HPC hardware issues

# Questions ?

48