



Mineure CalHau1

A short overview of High Performance Computing

Stéphane Vialle



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

What is « HPC » ?



High performance hardware

Numerical algorithms

- Serial optimization – tiling
- Vectorization
- Multithreading
- Message passing
- Parallel programming

Intensive computing & large scale applications

--- Big Data

High Performance Computing

TeraFlops
PetaFlops
ExaFlops

« HPDA »

TeraBytes
PetaBytes
ExaBytes₂

High Performance Hardware

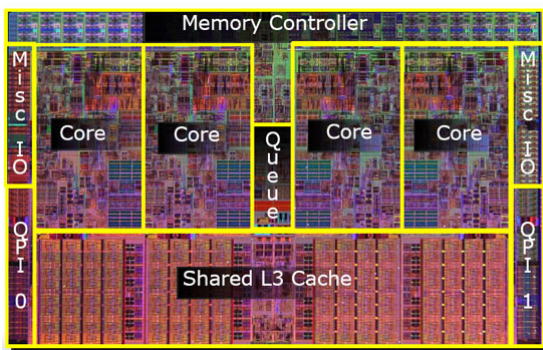
Inside



... high performance hardware

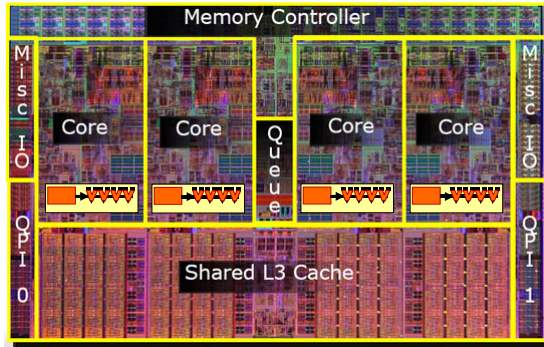
High Performance Hardware

From core to SuperComputer

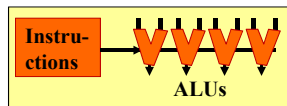


Computing cores

From core to SuperComputer

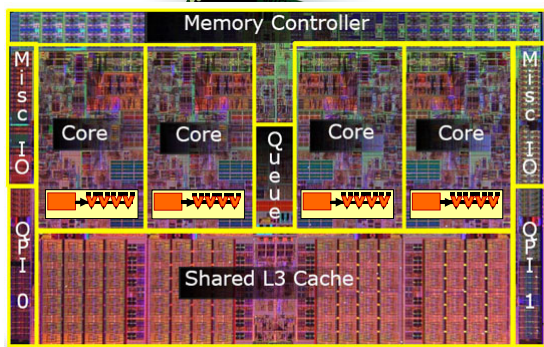


Computing cores
with **vector units**



Unités AVX

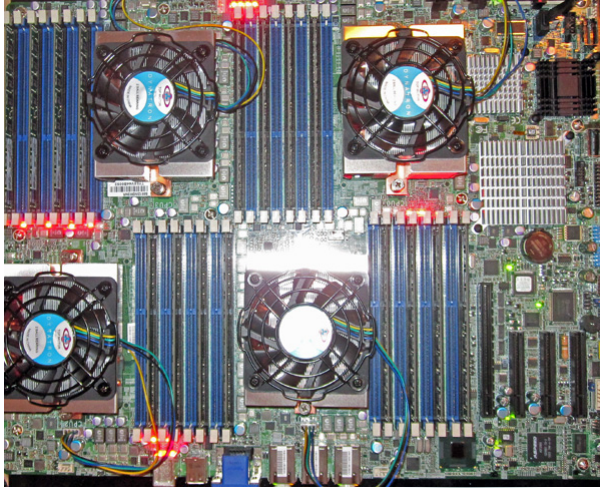
From core to SuperComputer



Computing cores
with vector units
Multi-core
processor

High Performance Hardware

From core to SuperComputer



Computing cores
with vector units

Multi-core
processor

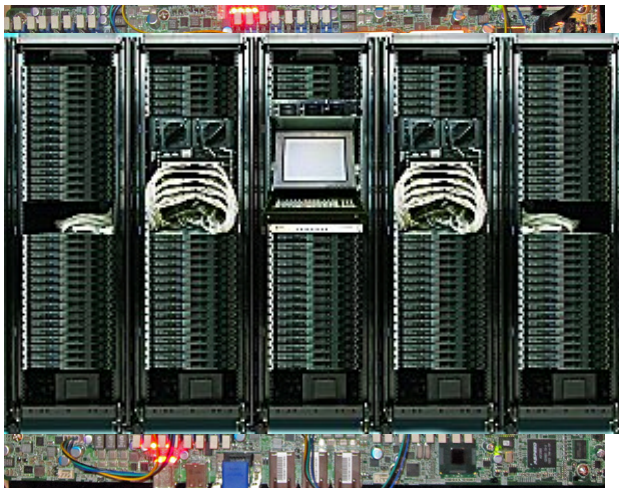
Multi-core
PC/node

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

From core to SuperComputer



Computing cores with vector units

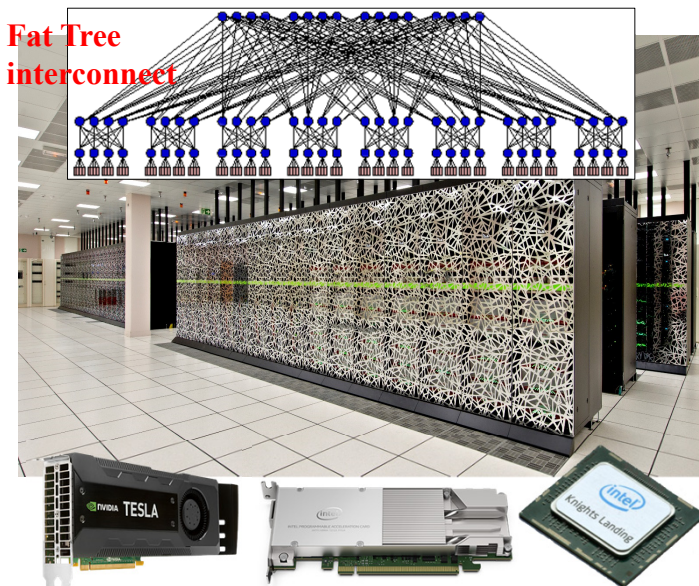
Multi-core processor

Multi-core PC/node

Multi-core PC cluster

Super-Computer

From core to SuperComputer



Fat Tree interconnect

Computing cores with vector units

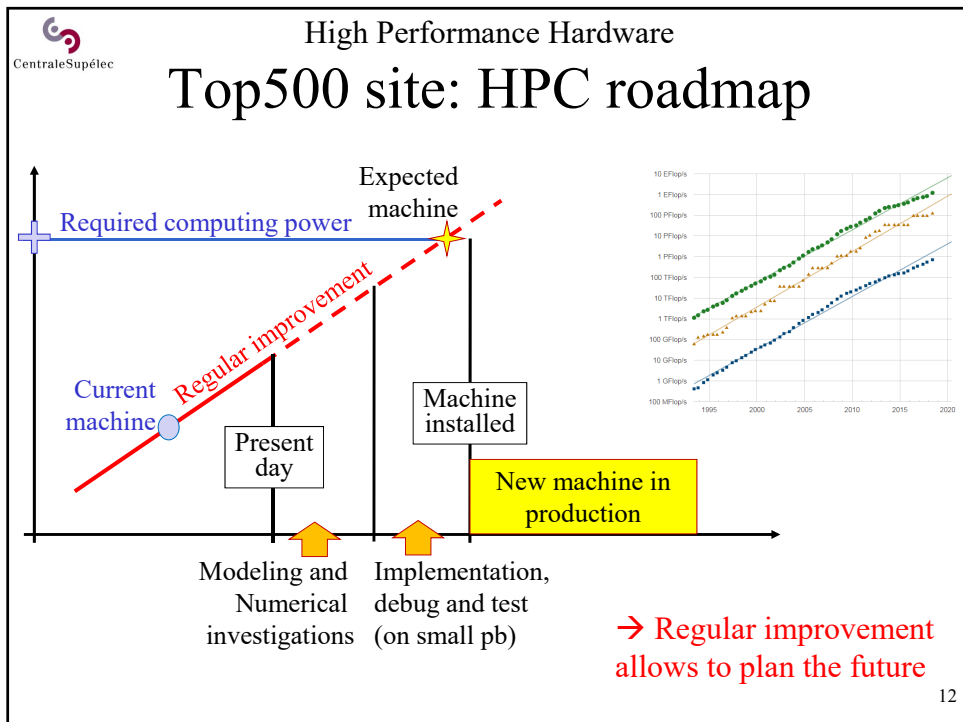
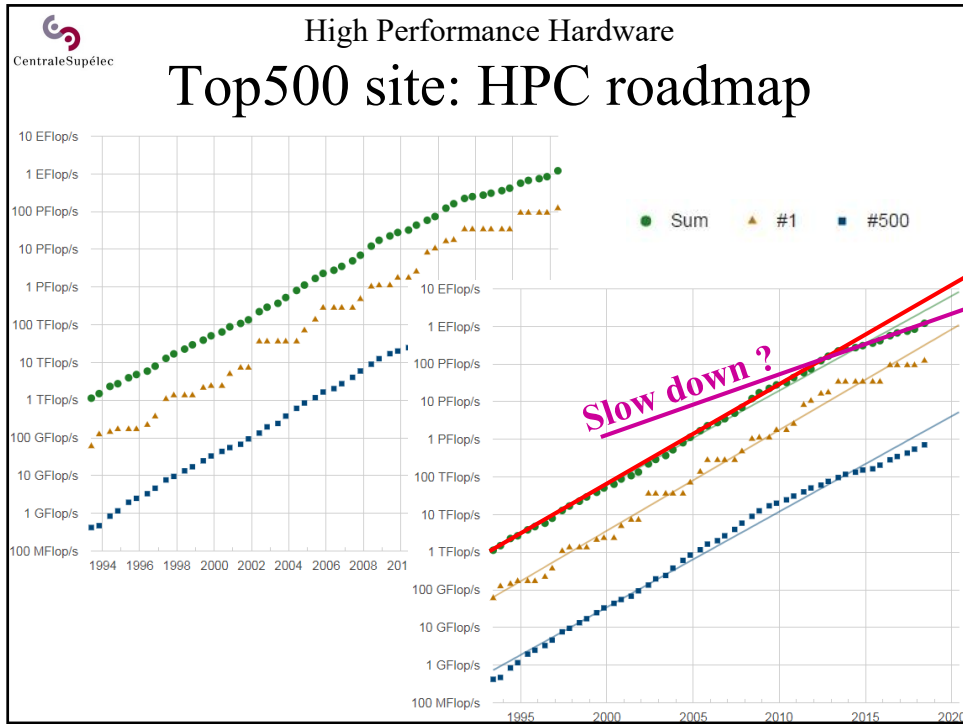
Multi-core processor

Multi-core PC/node

Multi-core PC cluster

Super-Computer

+ hardware accelerators



High Performance Hardware HPC in the cloud ?

Microsoft Azure

« AZUR BigCompute »

- High performance nodes
- High performance interconnect (Infiniband)
- Customers can allocate a part of a HPC cluster



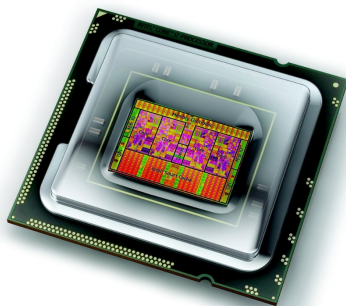
- Allows to allocate a huge number of nodes for a short time
- No high performance interconnection network
- Comfortable for Big Data scaling benchmarks

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

13

Architecture issues

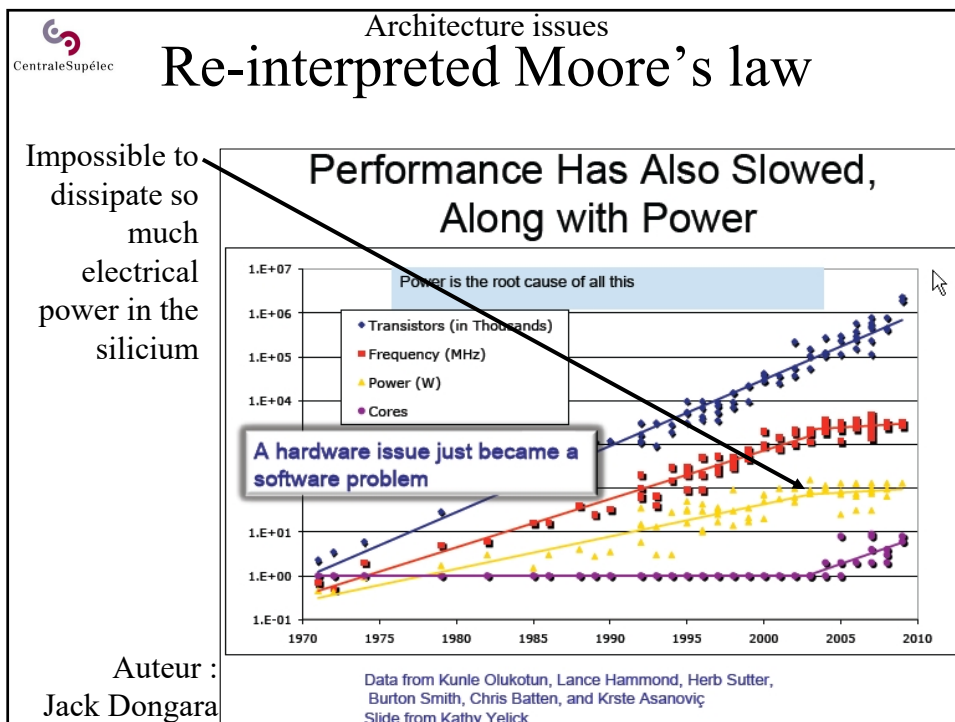
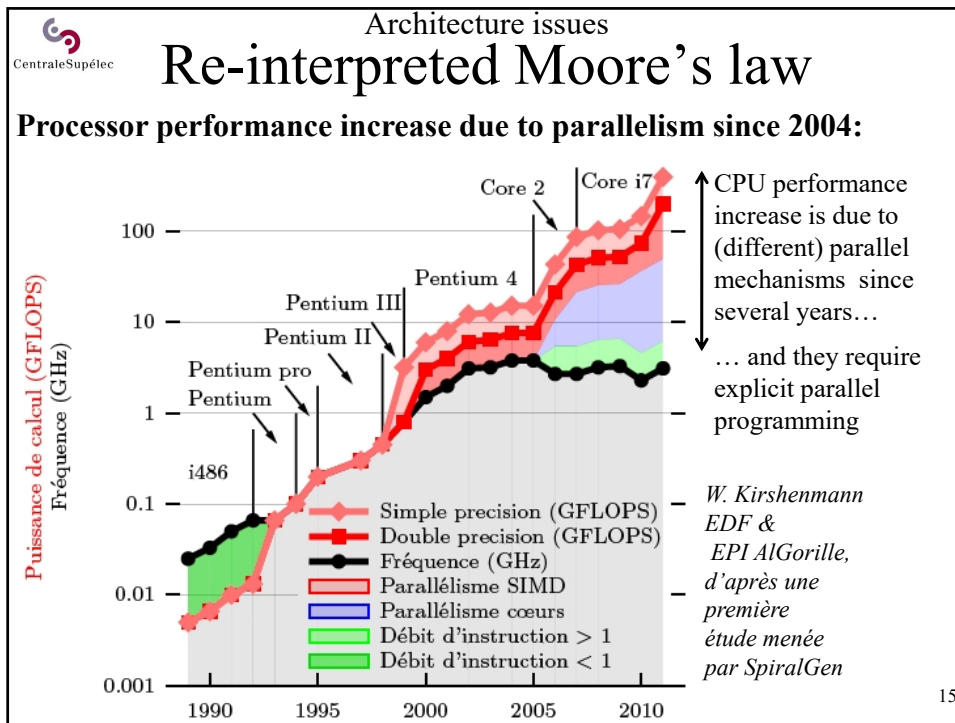
Why multi-core processors ?

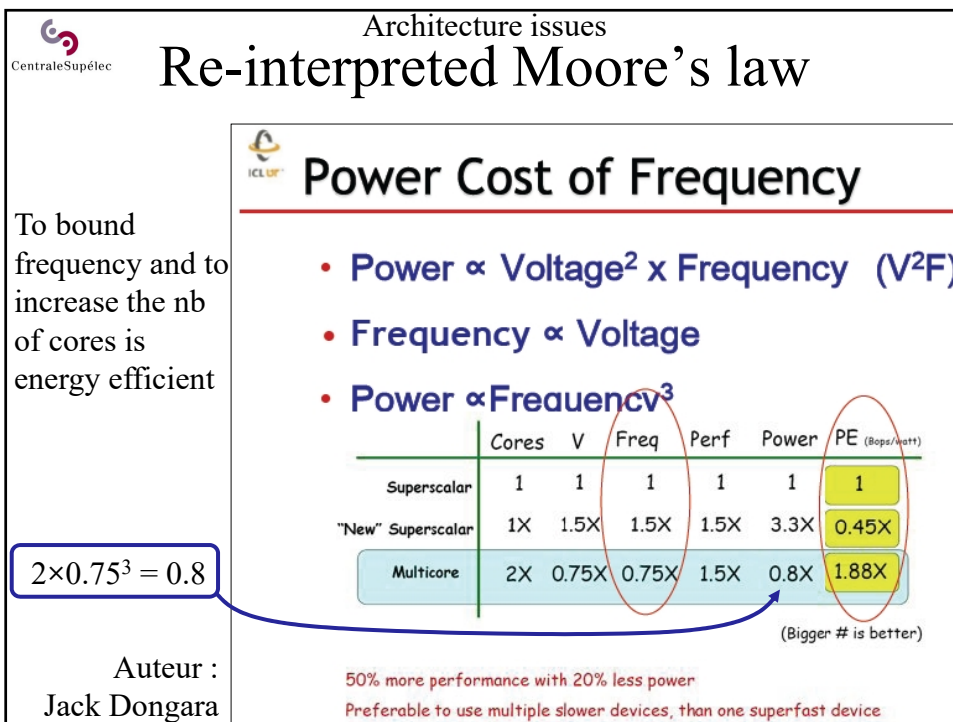
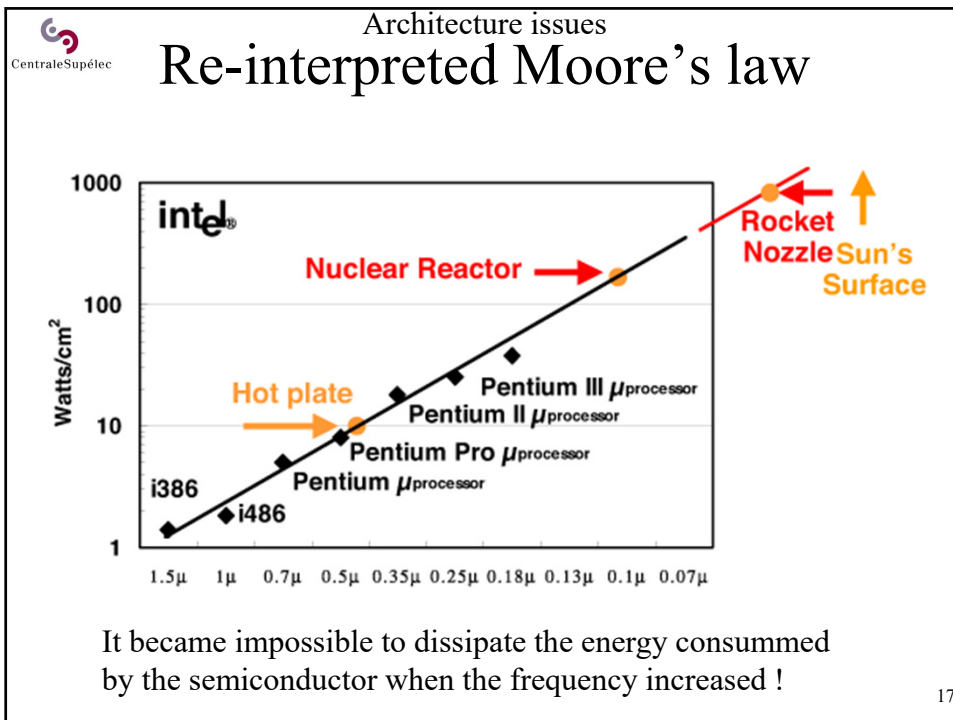


Shared or distributed memory ?



14





Re-interpreted Moore's law

Initial (electronic) Moore's law:

each 18 months \rightarrow x2 number of transistors per μm^2



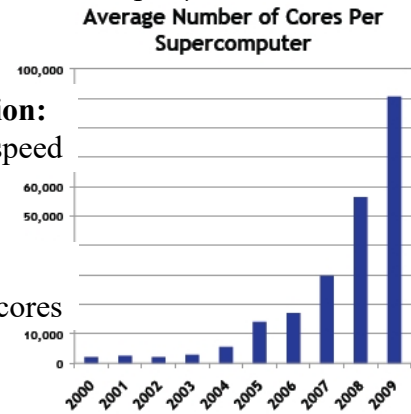
Previous computer science interpretation:

each 18 months \rightarrow x2 processor speed



New computer science interpretation:

each 24 months \rightarrow x2 number of cores

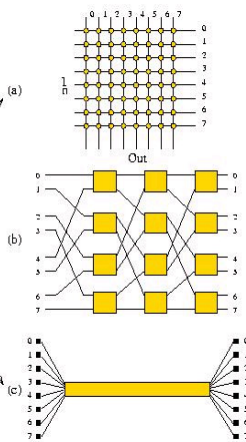
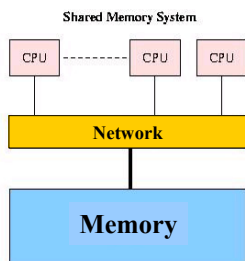


Leads to a massive parallelism challenge:

to split many codes in 100, 1000, 10^6 threads ... 10^7 threads!!

3 classic parallel architectures

Shared-memory machines (*Symmetric MultiProcessor*):



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

Overview of Recent Supercomputers
 Aad J. van der Steen
 Jack J. Dongarra²⁰

Architecture issues

CentraleSupélec

3 classic parallel architectures

Distributed-memory machines (*clusters*):

Mem Mem Mem

proc proc proc

network

Hypercubes

$d=1, \Omega=1$

$d=2, \Omega=2$

$d=3, \Omega=3$

$d=4, \Omega=4$

(a) Hypercubes, dimension 1-4.

Fat trees

(b) A 128-way fat tree.

Gigabit Ethernet

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

Highly scalable architecture

21

Architecture issues

CentraleSupélec

3 classic parallel architectures

Distributed Shared Memory machines (DSM):

Proc. Proc. Proc. Proc. Proc. Proc.

Mem. Mem. Mem.

network

Peripherals

cache coherence Non Uniform Memory Architecture (ccNUMA)

Extends the cache mechanism

Up to 1024 nodes
Support global multithreading

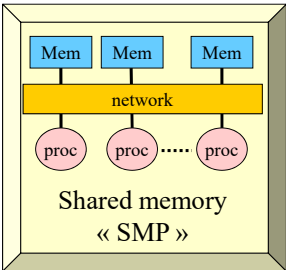
Overview of Recent Supercomputers
Aad J. van der Steen
Jack J. Dongarra

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

22

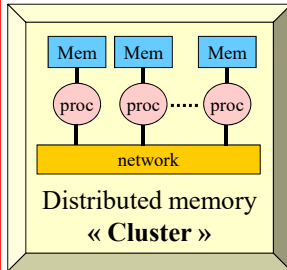
Architecture issues

3 classic parallel architectures



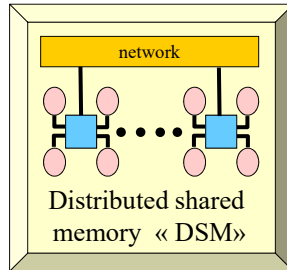
Shared memory
« SMP »

- 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.

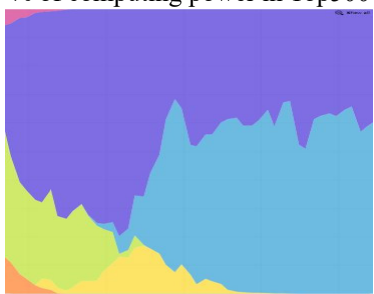
2016 : almost all supercomputers have a **cluster-like** architecture

23

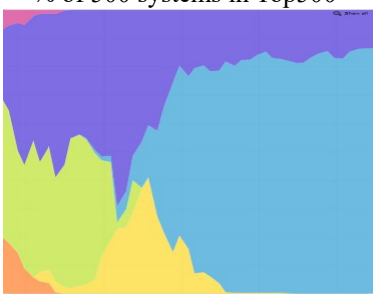
Architecture issues

Evolution of parallel architectures

% of computing power in Top500



% of 500 systems in Top500



■ Pure SIMD

■ Single processor

■ SMP

■ ShM vector machine

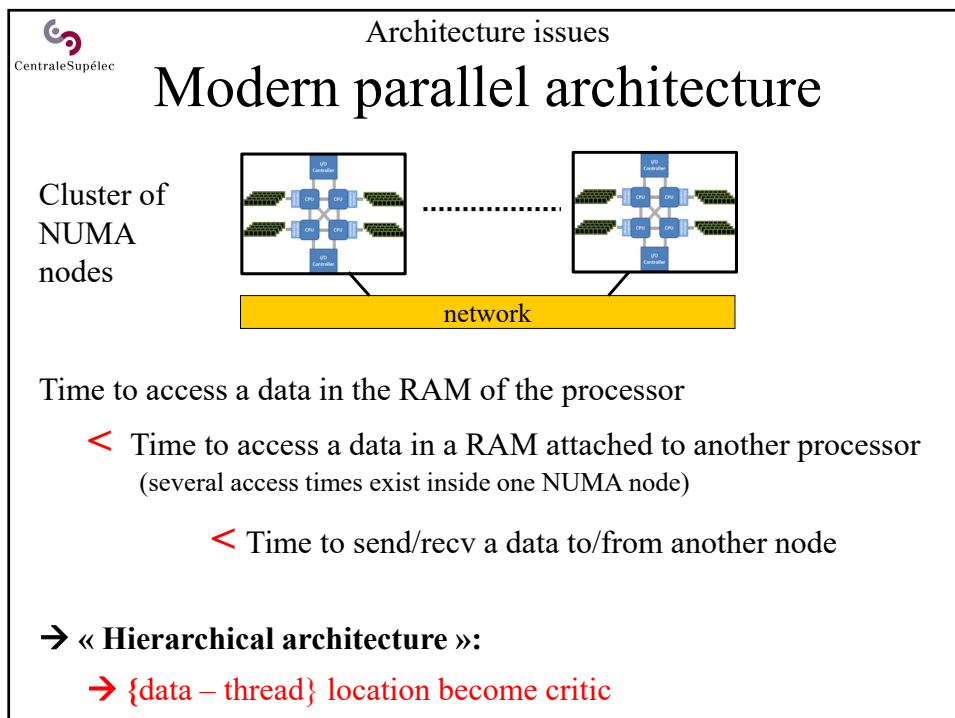
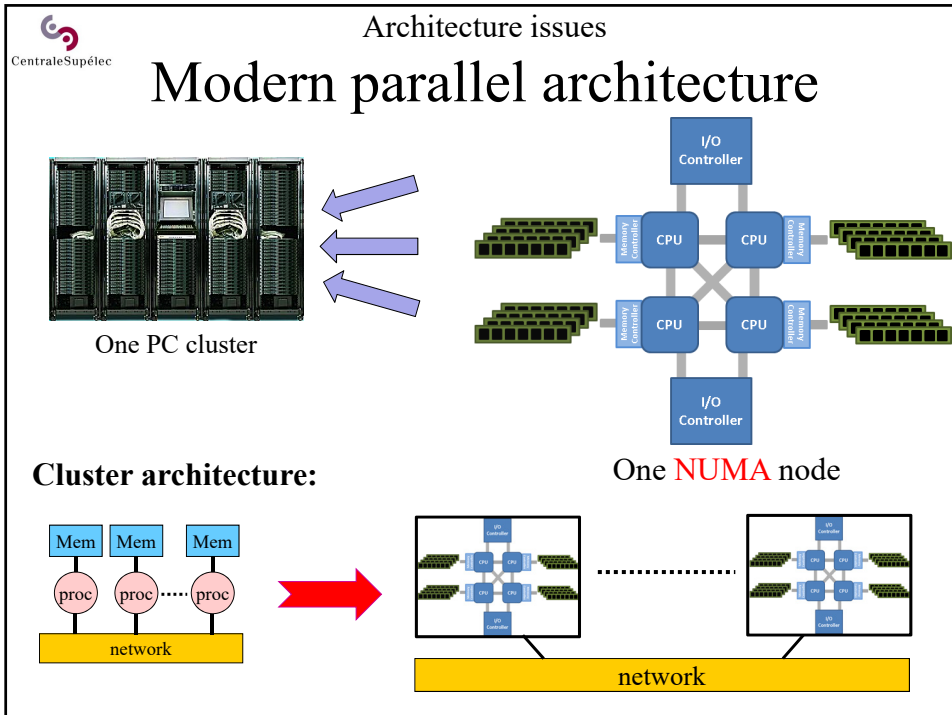
■ MPP (cluster with custom interconnect)

■ Cluster

2016 : almost all supercomputers have a **cluster-like** architecture




BUT ... →

24



Architecture issues

CentraleSupélec **Multi-paradigms programming**

Cluster of multi-processor NUMA nodes with hardware vector accelerators

Hierarchical and hybrid architectures :

→ multi-paradigms programming (or new *high level paradigm...*)

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>>>(...)

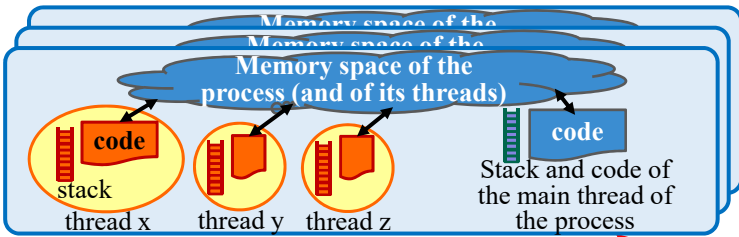
+ checkpointing

27

Architecture issues


CentraleSupélec **Distributed application deployment**

Memory space of the process (and of its threads)



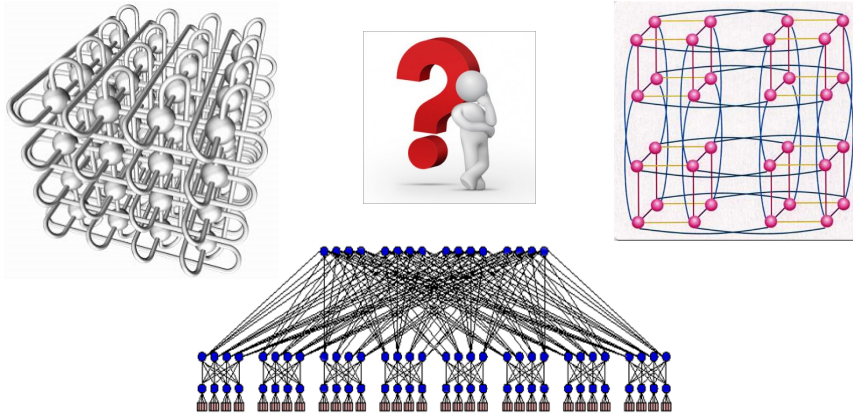
- Achieving **efficient mapping**
How to map software and hardware ressource ?
- Achieving **fault tolerance**
Which strategy and mechanisms ?
- Achieving **scalability**
Are my algorithm, mapping and fault tolerance strategy adapted to larger systems ?

Distributed Hardware Architecture



« Interconnect » (cluster interconnection network)

What is a good interconnection network for a parallel computer ?



29

Interconnect Main features

Main features of a computing cluster interconnect

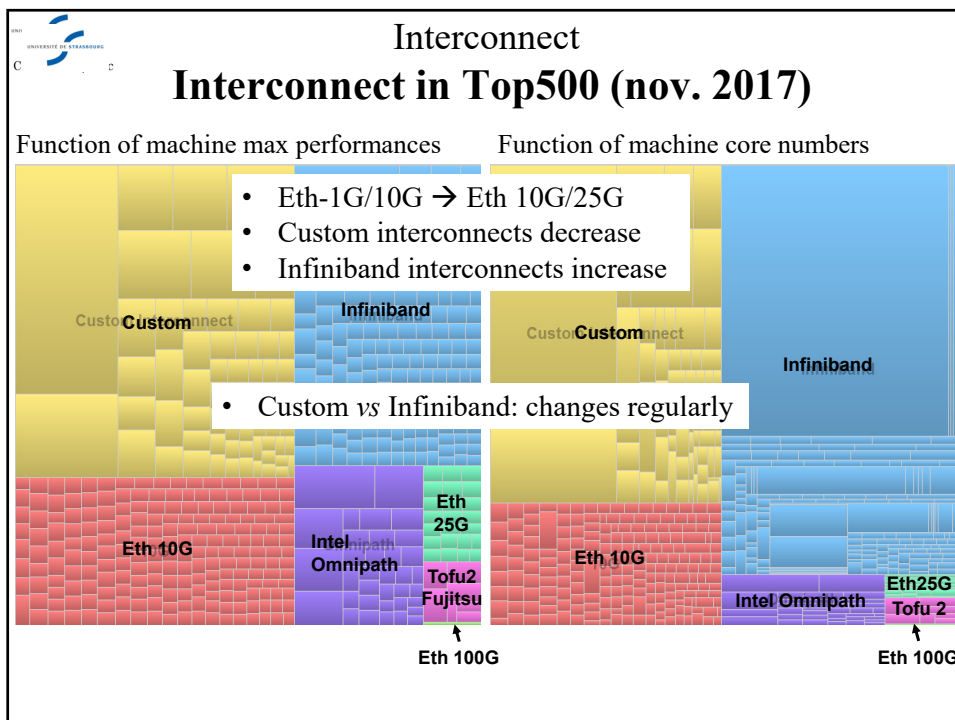
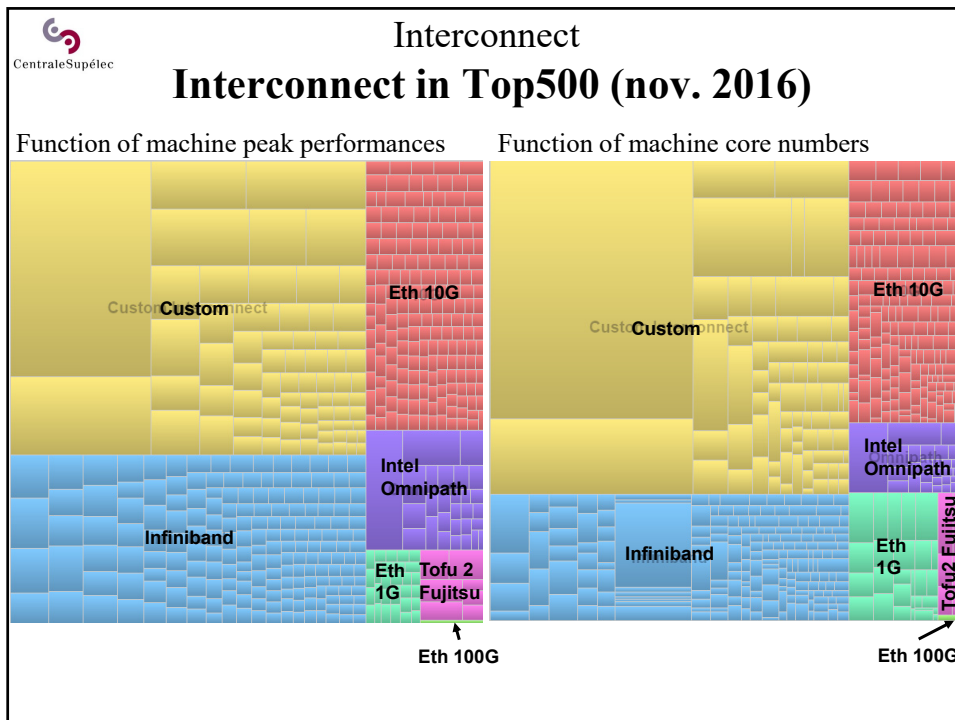
- **Bandwidth**
- **Latency**
- **Contention & saturation resilience**

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

**Sensitive criteria
are different of
LAN criteria**



10-Gigabit Ethernet vs Infiniband

10/25-Gigabit Ethernet :

- Used in many machines in Top500...
...but not in the most powerful
- High latency
- Cheap interconnect!
- Well known technology (not only in HPC)
→ knowledge already exist in any company/institution

Infiniband :

- Used in many machines in Top500...
...more powerful machines than Eth-10G machines
- Latency is lower than Eth-10/25G
- More expensive than Eth-10G (25G ?)
- Used only in HPC → special knowledge is required

Watch out: different versions of Infiniband exist, with different perf !

Proprietary custom networks

Proprietary custom networks

- Majors build SuperComputers with their proprietary Interconnect...
...or customise high quality Infiniband (?)

CRAY/IBM/Fujitsu/Chinese Supercomputers

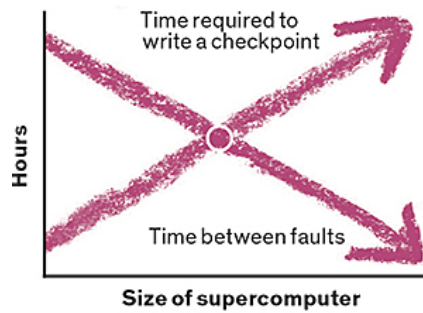
- Different networks and network topologies in one machine
pt-to-pt comm. network, collective comm. network, ctrl network
- They are the key component of a SuperComputer

Ex: Cray T3D has been the first SuperComputer to have an interconnect fast enough for its processor computing power

Fault Tolerance in HPC

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

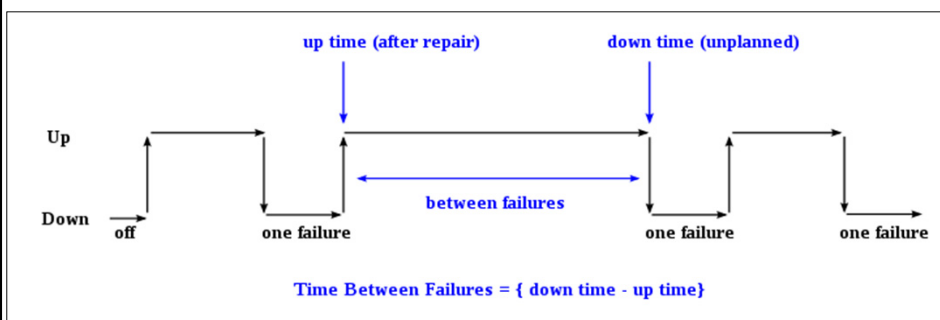
Can a one-million core parallel program run during one week ?



Fault tolerance

Mean Time Between Failures

MTBF definition:



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

Mean Time Between Failures

Experiments:

The Cray-1 required extensive maintenance. Initially, **MTBF was on the order of 50 hours**. MTBF is Mean Time Between Failures, and in this case, it was the average time the Cray-1 worked without any failures. Two hours of everyday was typically set aside for preventive maintenance.... (Cray-1 : 1976)

System Resilience at Extreme Scale White Paper

Prepared for Dr. William Harrod, Defense Advanced Research Project Agency (DARPA)

Today, 20% or more of the computing capacity in a large high-performance computing system is wasted due to failures and recoveries. **Typical MTBF is from 8 hours to 15 days**. As systems increase in size to field petascale computing capability and beyond, the MTBF will go lower and more capacity will be lost.

Addressing Failures in Exascale Computing
report produced by a workshop on “Addressing Failures in Exascale Computing”
2012-2013

37

Why do we need fault tolerance ?

Processor frequency is limited and number of cores increases
→ **we use more and more cores**



We do not attempt to speedup our applications
→ **we process larger problems in constant time ! (Gustafson's law)**



We use more and more cores during the same time
→ **probability of failure increases!**



**We (really) need for fault tolerance
or large parallel applications will never end!**

38

Fault tolerance strategies

High Performance Computing: big computations (batch mode)

- *Checkpoint/restart* is the usual solution
- **Complexify src code**, time consuming, disk consuming !

High Throughput Computing: flow of small and time constrained tasks

- Small and independent tasks
- A task is re-run (entirely) when failure happens

Fault tolerance in HPC remains a « hot topic »

Big Data:

- Data storage redundancy
- Computation on (frequently) incomplete data sets ...

Different approach !

39

Who need for fault tolerance ?

In a HPC cluster: computing resources are checked regularly

- Wrong resources are identified and not allocated
- Many users do not face frequent failures (good!)
(parallel computers are not so bad !)

Which users/applications need for fault tolerance ?

- **When running applications on large numbers of resources during long times**
- Need to restart from a recent checkpoint

Fault tolerance

Remark:

- Critical parallel applications** (with strong dead lines)
- Need for **redundant resources and runs**
- **Impossible** on very large parallel runs

High availability

40

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 !!

Energy consumption

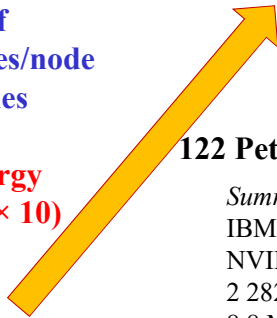
How much electrical power for an Exaflops ?

1.0 Exaflops should be reached close to 2020:

- 2.0 GWatts with the flop/watt ratio of 2008 Top500 1st machine
- 1.2 GWatts with the flop/watt ratio of 2011 Top500 1st 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)

From Petaflops to Exaflops

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



**1.00 Exaflops : 2018-2020
2020-2022**

25 Tb/s (IO)
20/35 MWatt max....

122 Petaflops : juin 2018

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



1.03 Petaflops : June 2008

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

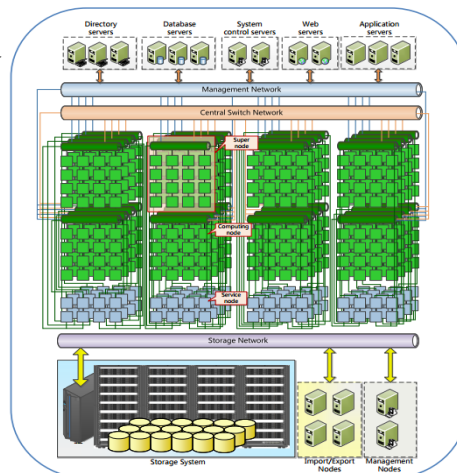
- How to program these machines ?
- How to train large programmer teams ?

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



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

Summit Overview

OpenPOWER

Compute Node
2 x POWER9
6 x NVIDIA GV100
NVMe-compatible PCIe 1600 GB SSD

Components IBM POWER9
• 22 Cores
• 4 Threads/core
• NVLink

25 GB/s EDR (8-12 ports)
512 GB DRAM (CCB4)
96 GB HBM (3D Stacked)
Coherent Shared Memory

NVIDIA GV100
• 712
• 16 GB @ 0.9 TB/s
• NVLink

Compute Rack
18 Compute Servers
Warm water (70°F direct-cooled components)
RDHX for air-cooled components

39.7 TB Memory/rack
55 kW max power/rack

Compute System
10.2 PB Total Memory
256 compute racks
4,608 compute nodes
Mellanox EDR IB fabric
200 PFLOPS
~13 MW

GPFS File System
250 PB storage
2.5 TB/s read, 2.5 TB/s write

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

Summit Overview

OpenPOWER

Compute Node
2 x POWER9
6 x NVIDIA GV100
NVMe-compatible PCIe 1600 GB SSD

Components IBM POWER9
• 22 Cores
• 4 Threads/core
• NVLink

25 GB/s EDR (8-12 ports)
512 GB DRAM (CCB4)
96 GB HBM (3D Stacked)
Coherent Shared Memory

NVIDIA GV100
• 712
• 16 GB @ 0.9 TB/s
• NVLink

Compute Rack
18 Compute Servers
Warm water (70°F direct-cooled components)
RDHX for air-cooled components

39.7 TB Memory/rack
55 kW max power/rack

Compute System
10.2 PB Total Memory
256 compute racks
4,608 compute nodes
Mellanox EDR IB fabric
200 PFLOPS
~13 MW

GPFS File System
250 PB storage
2.5 TB/s read, 2.5 TB/s write

What is the sustainable architecture ?

Différentes stratégies s'affrontent dans le Top500 :

- La performance à tous prix avec de gros CPUs très gourmands
Cray XT6 : 1.7 Pflops, 6.9 Mwatts
K-Computer : 10.5 Pflops, 12.6 MWatts
- Beaucoup de processeurs moyennement puissants et peu gourmands
IBM Blue Gene (gamme terminée)
- Utilisation d'accélérateurs matériels : GPU, ~~Xeon-phi~~, ...
→ machines hybrides : CPU + accélérateurs
→ difficiles à programmer et pas adaptées à tous les problèmes

Quel est le(s) bon(s) choix pour atteindre l'Exaflops ?

Quel est le choix pertinent pour de « plus petits » clusters ?

47

Cooling

Cooling is close to 30% of the energy consumption



Optimization is mandatory!

48

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 m² au sol)

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

Souvent estimé à 30% de la dépense énergétique!

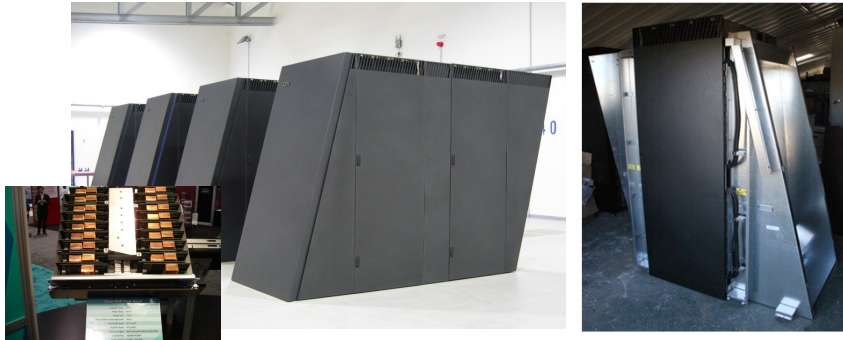


Optimized air flow

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

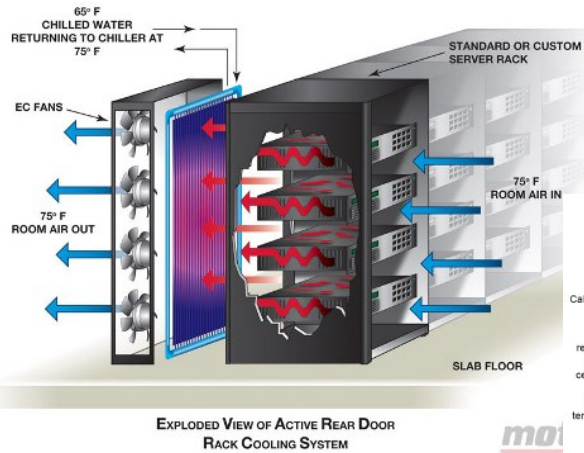
- Architecture Blue Gene : haute densité de processeurs
- Objectif d'encombrement minimal (au sol) et de consommation énergétique minimale
- **Formes triangulaires ajoutées pour optimiser le flux d'air**

IBM Blue Gene

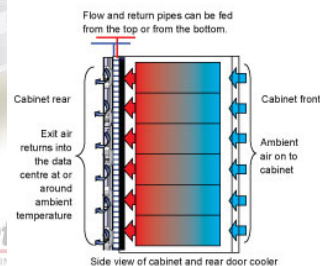


Cold doors (air+water cooling)

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



Le refroidissement se concentre sur l'armoire.



Direct liquid cooling

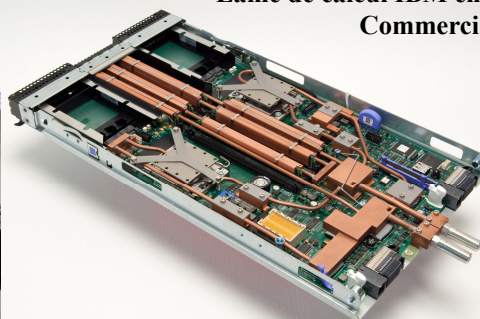
On amène de l'eau froide directement sur le point chaud, mais l'eau reste isolée de l'électronique.

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

Carte expérimentale IBM en 2009 (projet Blue Water)

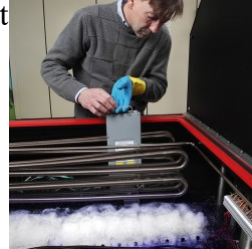


Lame de calcul IBM en 2012 Commercialisée



Liquid and immersive cooling

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

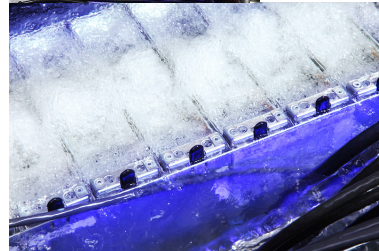


Refroidissement liquide par immersion testé par SGI & Novec en 2014

Cray 2 (1985)

- 4 processeurs
- 1.9 Gflops
- FLuorocarbon

Refroidissement liquide par immersion sur le CRAY-2 en 1985



53

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) !

54

Extreme air cooling

Refroidissement avec de l'air à température ambiante :

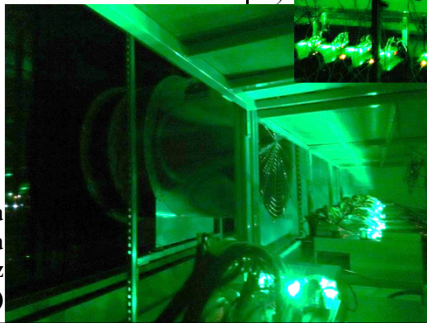


- circulant à grande vitesse
- circulant à gros volume

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

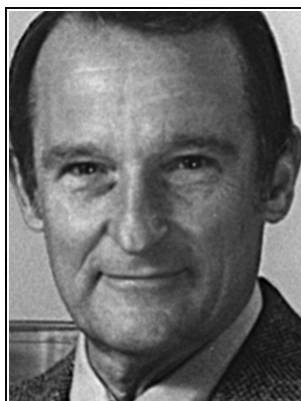
→ Il n'y a pas de refroidissement du flux d'air.

Installation Ilium à CentraleSupélec à Metz (blockchain - 2018)



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

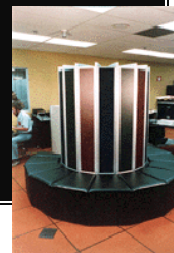
Interesting history of CRAY company



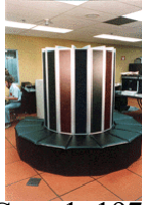
If you were plowing a field, which would you rather use? Two strong oxen or 1024 chickens?

— Seymour Cray —

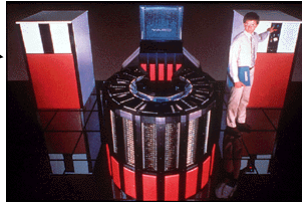
AZ QUOTES



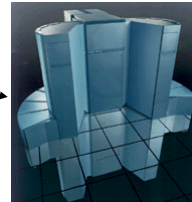
Architecture des machines parallèles Histoire des ordinateurs CRAY



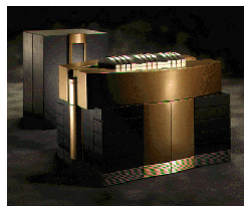
Cray-1, 1976
133Mflops



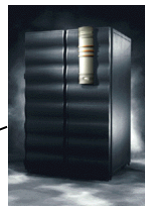
Cray-2, 1985
1.9 gigaflops



Cray-YMP, 1988



Cray-T90,
60 gigaflops

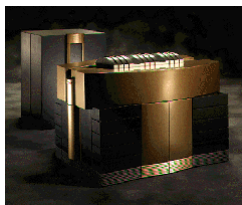


Cray-J90



Cray-C90, 1991
16 gigaflops

Architecture des machines parallèles Histoire des ordinateurs CRAY



Cray-T90,
60 gigaflops



Cray-SV1
1 téraflop

NEC (SX)

Accord
d'importation



Cray-SX-6

Cray est démembré et
semble avoir disparu.

Puis en 2002 un
événement survient

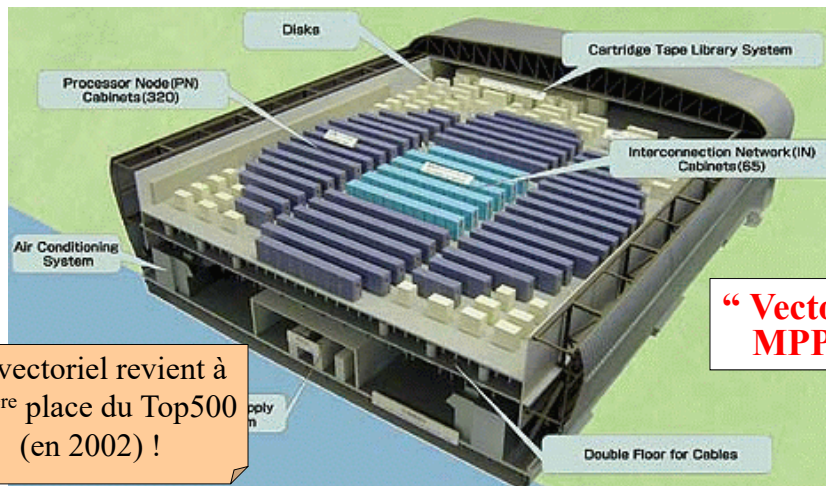


Cray-SV2

Architecture des machines parallèles Histoire des ordinateurs CRAY

Apparition du *Earth Simulator* : gros cluster vectoriel **NEC** :

- 640-nœuds de 8 processeurs : **5120 processeurs**
- 40 Tflops crête, a atteint les **35 Tflops** en juin 2002



“ Vector MPP ”

Le vectoriel revient à la 1^{ère} place du Top500 (en 2002) !

Architecture des machines parallèles Histoire des ordinateurs CRAY

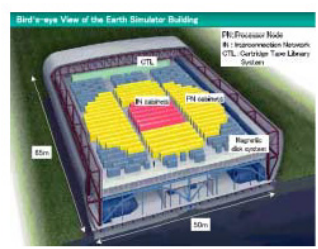


Japan's Impressive Earth Simulator Is As Fast As the Top 20 U.S. Supercomputers Combined

The New York Times
Japanese Computer Is World's Fastest, as U.S. Falls Back
 By John Markoff
 SAN FRANCISCO, April 19, 2002 — A Japanese laboratory has built the world's fastest computer, a machine so powerful that it matches the raw processing power of the 20 fastest American computers combined and far outstrips the previous leader.

"The first results from the Earth Simulator are stunning. The Earth Simulator will put U.S. scientists at a 10-100 fold disadvantage vis-à-vis their colleagues in Japan. The U.S. has lost the lead in climate science. If we allow the Japanese to deploy [this] strategy uncontested, we will surely lose the lead in other computational disciplines."

Reasserting U.S. Leadership in Scientific Computation, U.S. Department of Energy - Office of Science (June 2002)

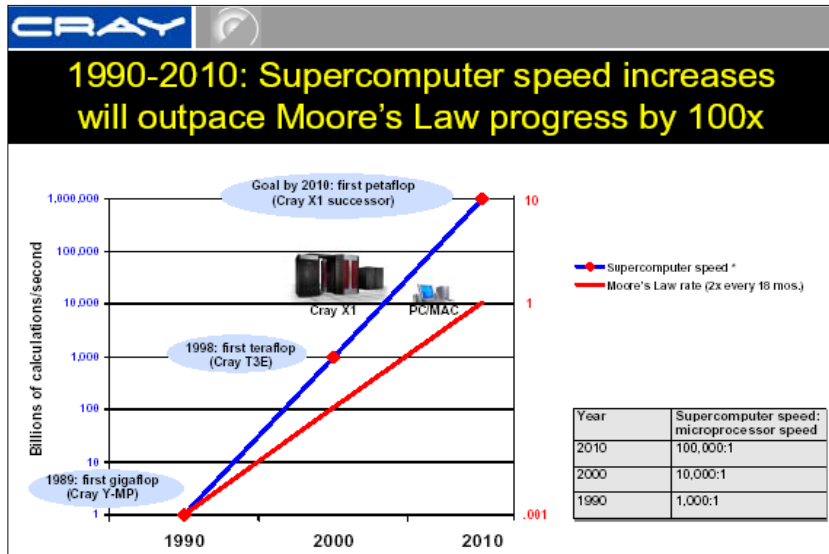


"The balance of the computer makes it easier to use."

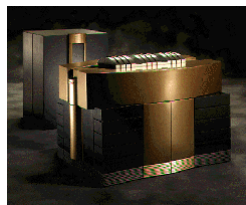
Forte inquiétude des USA !

Architecture des machines parallèles Histoire des ordinateurs CRAY

CRAY était de nouveau là, avec de grandes ambitions:



Architecture des machines parallèles Histoire des ordinateurs CRAY



Cray-T90,
60 gigaflops



Cray-SV1
1 téraflop



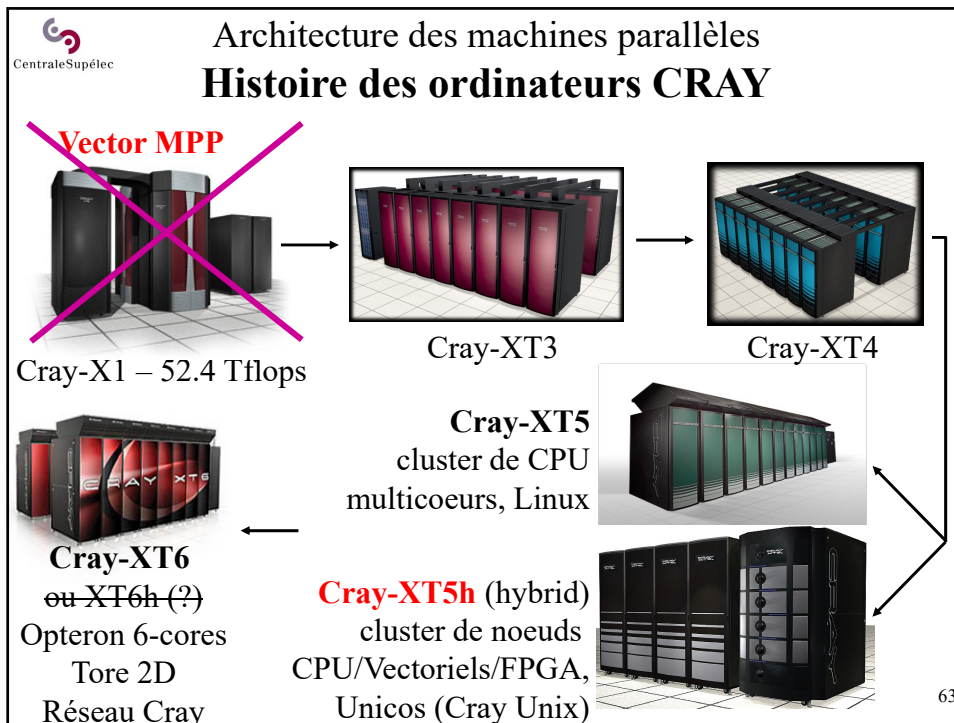
Cray-X1 – 52.4 Tflops



~~Cray-SV2~~



Cray-SX-6



CentraleSupélec

Architecture des machines parallèles Histoire des ordinateurs CRAY

Cray XT6 :
 1^{er} au top500 en novembre 2009 : 1.7 Pflops avec 6.9 M watt
 Architecture : réseau d'interconnexion propriétaire + Opteron 6-cœurs
 → Architectures traditionnelles et très consommatrices d'énergie
 mais très efficace et sous Linux (logiciels disponibles)

Machine dénommée « Jaguar »

**Cray de nouveau à la 1^{ère} place
en nov 2009 avec des Opteron**

Architecture des machines parallèles

Histoire des ordinateurs CRAY

Cray XK7 : entrée des accélérateurs dans la gamme CRAY

1^{er} au top500 en novembre 2012 : 17.6 Pflops avec 8.2 Mwatt

Architecture :

- réseau d'interconnexion propriétaire
- chaque nœud : Opteron 16-cœurs + GPU NVIDIA Tesla K20
- 18688 nœuds → 299008 CPU cores + 18688 GPU K20

**Cray à la 1^{ère} place en nov
2012 avec Opteron + GPU**



Architecture des machines parallèles

Histoire des ordinateurs CRAY

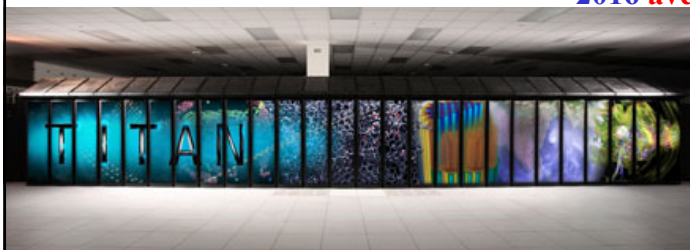
Cray XK7 : entrée des accélérateurs dans la gamme CRAY

1^{er} au top500 en novembre 2012 : 17.6 Pflops avec 8.2 Mwatt

Architecture :

- réseau d'interconnexion propriétaire
- chaque nœud : Opteron 16-cœurs + GPU NVIDIA Tesla K20
- 18688 nœuds → 299008 CPU cores + 18688 GPU K20

**Cray à la 3^{ème} place en juin
2016 avec Opteron + GPU**



A short overview of
High Performance Computing

Questions ?