CentraleSupélec

SG6: High Performance Computing

Introduction to HPC hardware issues

Stéphane Vialle

CentraleSupélec université PARIS-SACLAY École doctorale Sciences et technologies de l'information et de la communication (STIC) RISEGrid Grand Est Alsace Champagne-Ardenne Lorraine

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

CentraleSupélec

What is « HPC » ?

High performance hardware

Parallel algorithmic & programming

Numerical algorithms

Intensive computing & large scale applications

High Performance Computing

TeraFlops
PetaFlops
ExaFlops

HPDA

Big Data

TeraBytes
PetaBytes
ExaBytes

2

CentraleSupélec

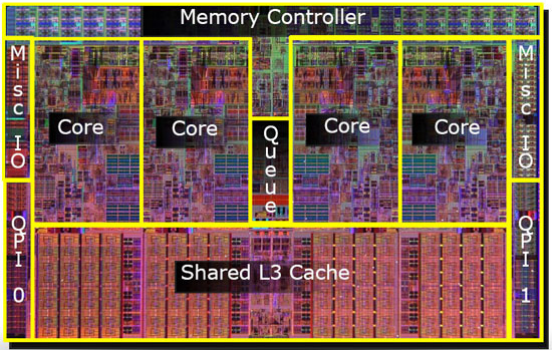
High Performance Hardware

Inside ….

… high performance hardware

CentraleSupélec

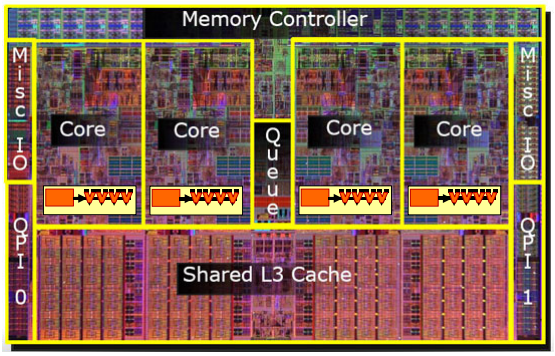
From core to SuperComputer



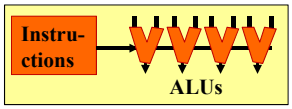
Computing cores

High Performance Hardware

From core to SuperComputer



Computing cores
with vector units

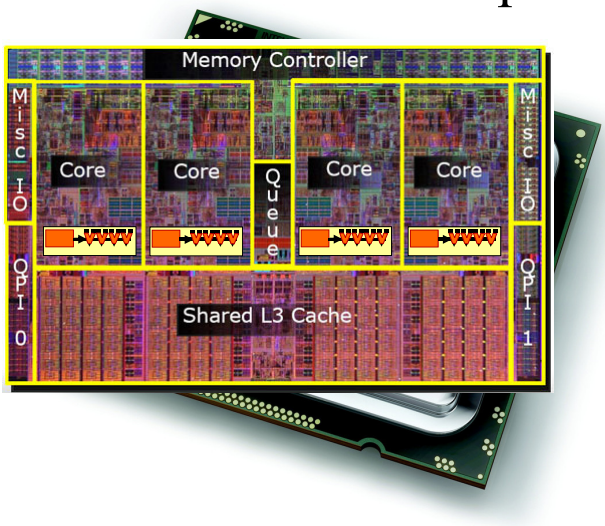


Unités AVX

5

High Performance Hardware

From core to SuperComputer



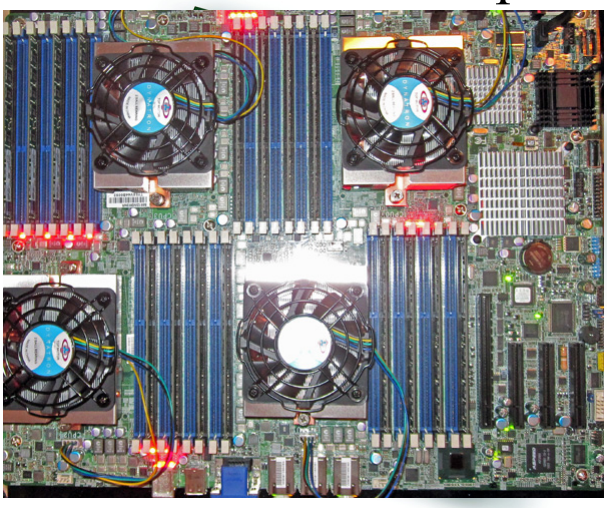
Computing cores
with vector units

Multi-core
processor

6

High Performance Hardware

From core to SuperComputer



Computing cores with vector units

Multi-core processor

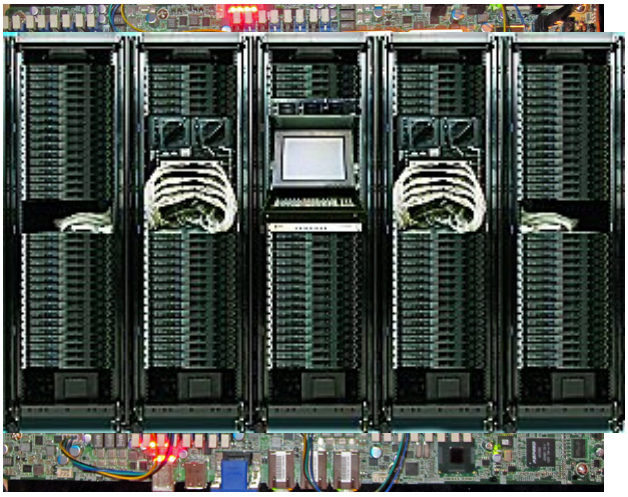
Multi-core PC/node

7

CentraleSupélec

High Performance Hardware

From core to SuperComputer



Computing cores with vector units

Multi-core processor

Multi-core PC/node

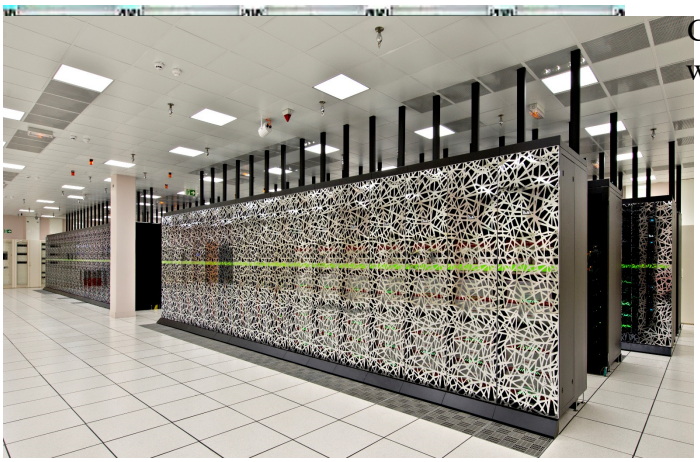
Multi-core PC cluster

8

CentraleSupélec

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

CentraleSupélec

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

CentraleSupélec

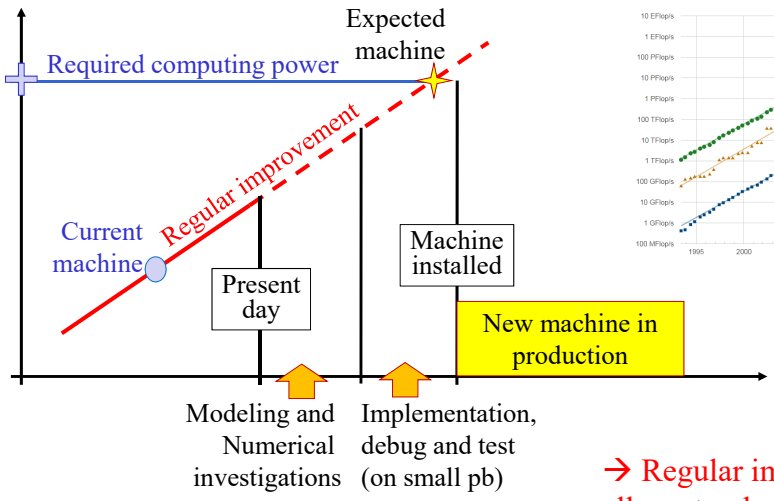
High Performance Hardware

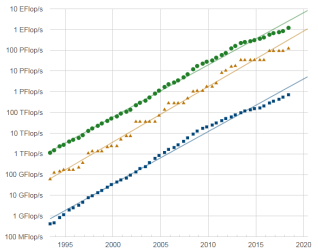
Top500 site: HPC roadmap

CentraleSupélec

High Performance Hardware

Top500 site: HPC roadmap





→ Regular improvement allows to plan the future

12

CentraleSupélec

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**

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

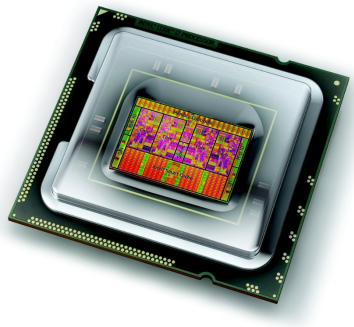
Some HPC or Large Scale PC-clusters exist in some Clouds
But no *SuperComputer* available in a cloud

13

CentraleSupélec

Architecture issues

Why multi-core processors ?



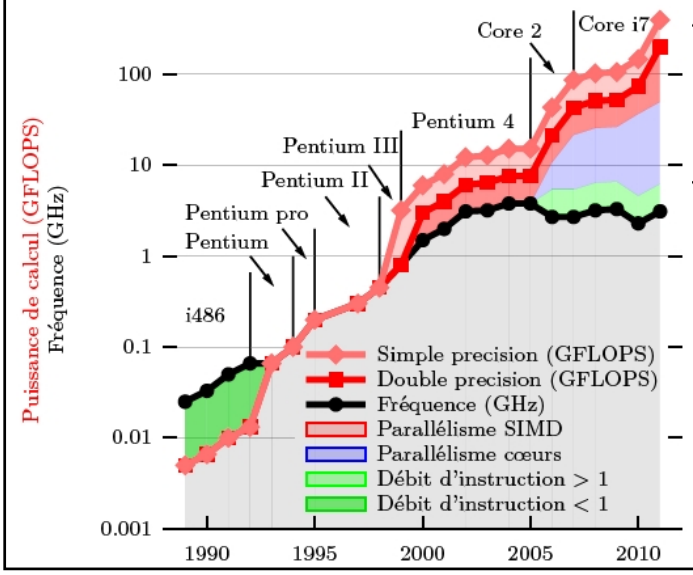
Shared or distributed memory ?

14

Architecture issues

Re-interpreted Moore's law

Processor performance increase due to parallelism since 2004:



CPU performance increase is due to (different) parallel mechanisms since several years…

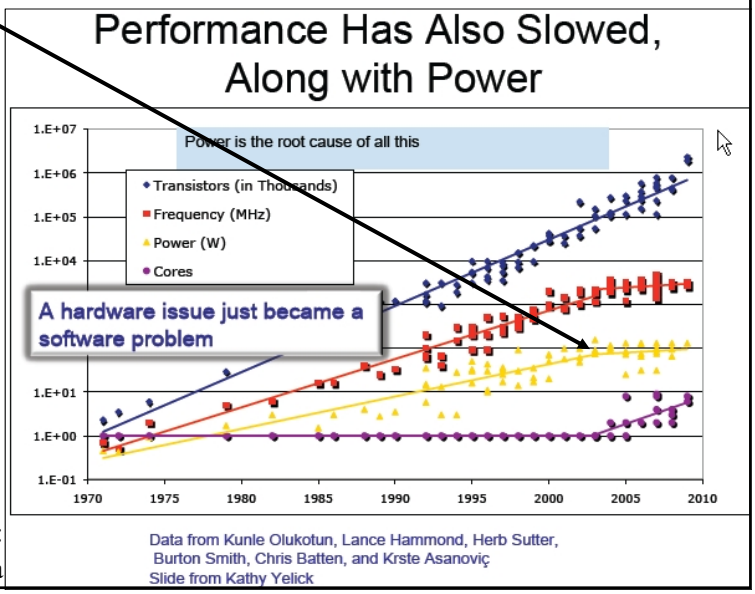
… and they require explicit parallel programming

W. Kirshenmann EDF & EPI AlGorille, d'après une première étude menée par SpiralGen

Architecture issues

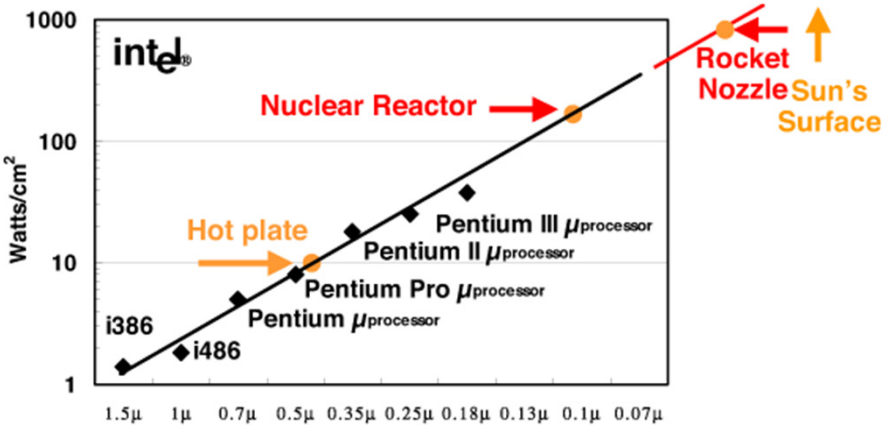
Re-interpreted Moore's law

Impossible to dissipate so much electrical power in the silicium



Auteur : Jack Dongara

Re-interpreted Moore's law



It became impossible to dissipate the energy consummed by the semiconductor when the frequency increased !

17

Architecture issues

Re-interpreted Moore's law

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

$2 \times 0.75^3 = 0.8$

Auteur : Jack Dongara

Power Cost of Frequency

- Power ∝ Voltage² x Frequency (V²F)
- Frequency ∝ Voltage
- Power ∝ Frequency³

	Cores	V	Freq	Perf	Power	PE (Bops/watt)
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

(Bigger # is better)

50% more performance with 20% less power

Preferable to use multiple slower devices, than one superfast device

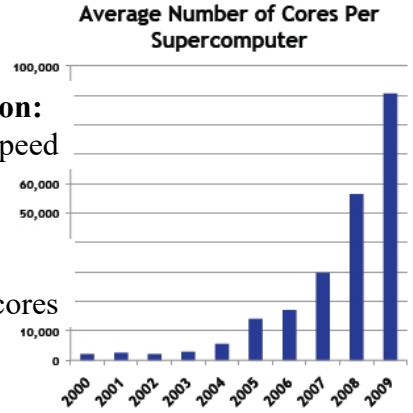
Architecture issues

Re-interpreted Moore's law

Initial (electronic) Moore's law:
each 18 months → x2 number of transistors per µm²

Previous computer science interpretation:
each 18 months → x2 processor speed

New computer science interpretation:
each 24 months → x2 number of cores

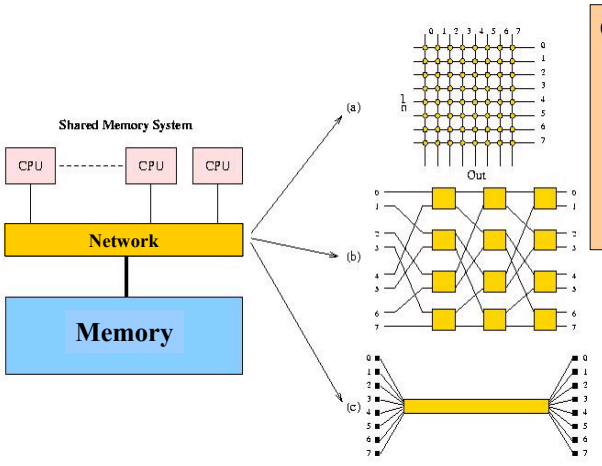


Leads to a massive parallelism challenge:
to split many codes in 100, 1000, ….. 10^6 threads … 10^7 threads!!

Architecture issues

3 classic parallel architectures

Shared-memory machines (*Symetric MultiProcessor*):

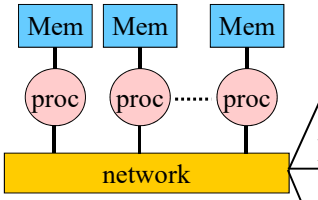


One principle:
- several implementations,
- different costs,
- different speeds.

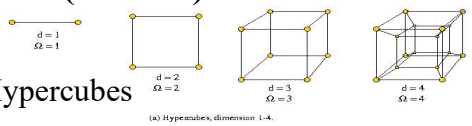
Overview *of Recent Supercomputers*
Aad J. van der Steen
Jack J. Dongarra20

3 classic parallel architectures

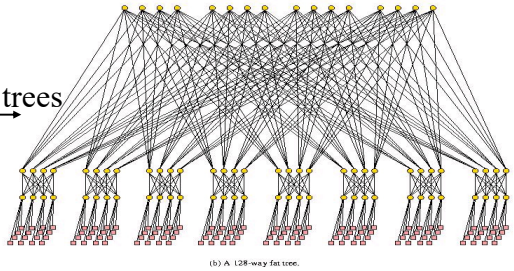
Distributed-memory machines (*clusters*):



Hypercubes



Fat trees



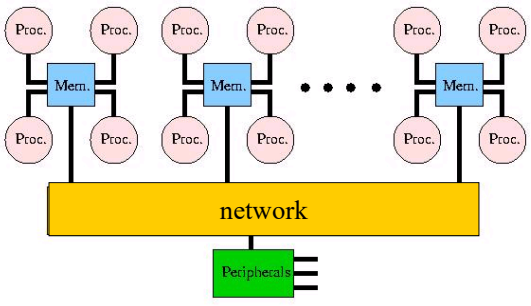
Cluster basic principles, but cost and speed depend on the interconnection network !

Gigabit Ethernet

Highly scalable architecture

3 classic parallel architectures

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



cache coherence Non Uniform Memory Architecture (**cc**NUMA)

Extends the cache mechanism

Up to 1024 nodes
Supports global multithreading

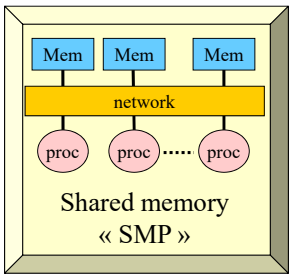
Overview *of Recent Supercomputers*
Aad J. van der Steen
Jack J. Dongarra

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

CentraleSupélec

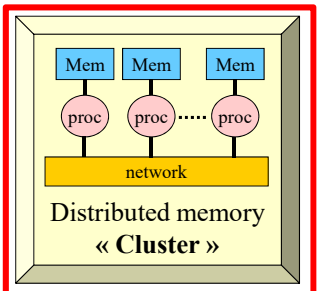
Architecture issues

3 classic parallel architectures



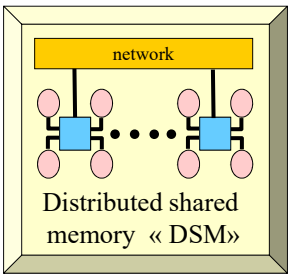
Shared memory
« SMP »

- 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.

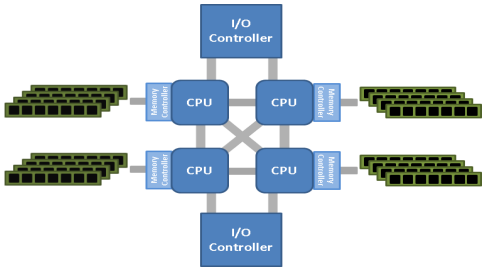
2019 : all supercomputers are **clusters** with a **hierarchical architecture**

23

CentraleSupélec

Architecture issues

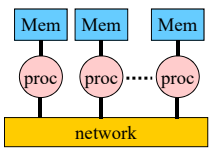
Modern clusters



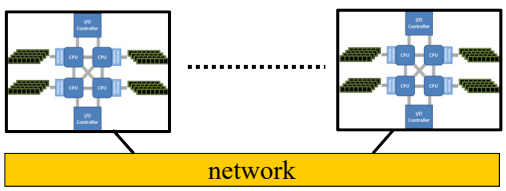
One PC cluster

One Non-Uniform Memory Architecture node

Cluster of nodes



Cluster of NUMA nodes

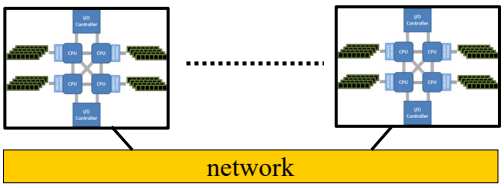


CentraleSupélec

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

CentraleSupélec

Architecture issues

Hierarchical & hybrid architecture

Cluster of multi-processor NUMA nodes with hardware vector accelerators

- **Hierarchical and hybrid** architectures
- **Multi-paradigms** programming

Developement 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>>>(…)

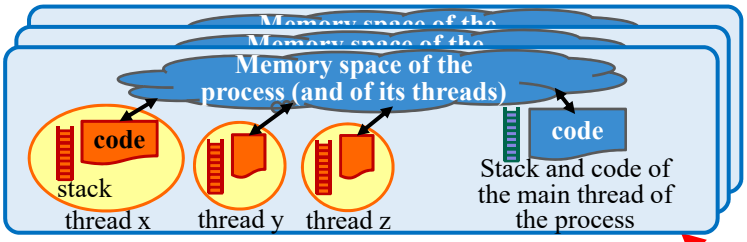
+ *checkpointing*

26

Architecture issues

Hierarchical & hybrid architecture

Distributed Software Architecture



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

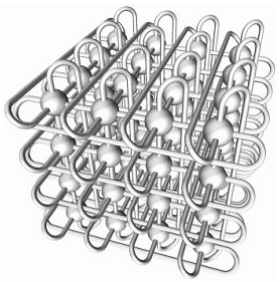
- One process per node ? per processor ? per core ?
- How many threads per process ? per core ?
- How to distributed the processes and threads to load balance the computations ?
- How to group intensively communicating processes ?
- How to co-localize data and process/threads ?

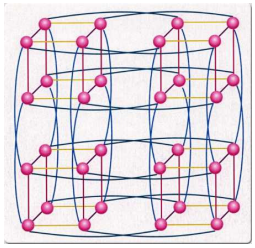
Deployment problem

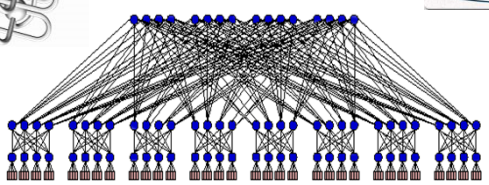
Distributed Hardware Architecture

« Interconnect » (cluster interconnection network)

What is a good interconnection network for a parallel computer ?







28

Main features of a computing cluster interconnect

Criteria are differents of LAN criteria

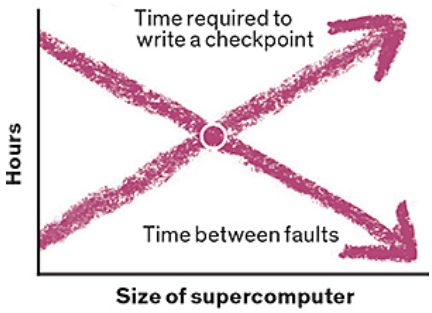
- Bandwidth
- **Latency**
- **Contention & saturation resilience**
  many algorithms are synchronous ones: all nodes compute, and then enter a communication step at the same time
- Performances of *Point-to-Point Communications*
- Performances of *Collective Communications*
  broadcast, scatter, gather, 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

10/25-Gigabit Ethernet ⇨ **Infiniband** ⇨ **Proprietary networks**

Fault Tolerance in HPC

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

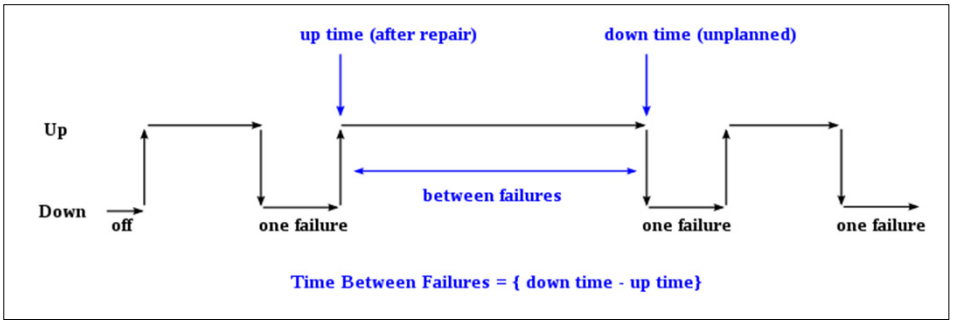
Can a one-million core parallel program run during one week ?



30

Mean Time Between Failures

MTBF definition:



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

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

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)

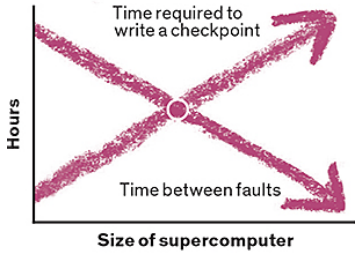
⇩

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 ?**



CentraleSupélec

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

CentraleSupélec

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 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)

34

Energy consumption

From Petaflops to Exaflops

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)

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 !!!!!!

Related challenges:
- Elegant & efficient programming
- Training of large programmer teams

35

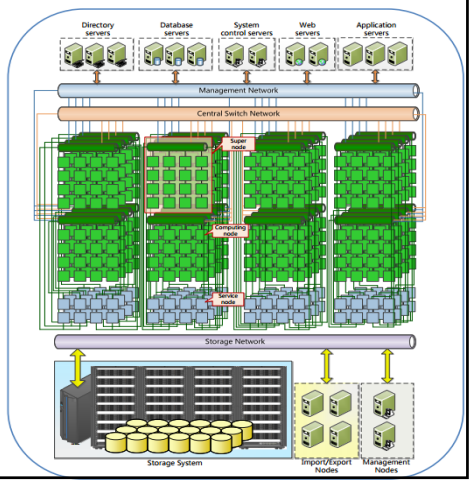
Energy consumption

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



CentraleSupélec

Energy consumption

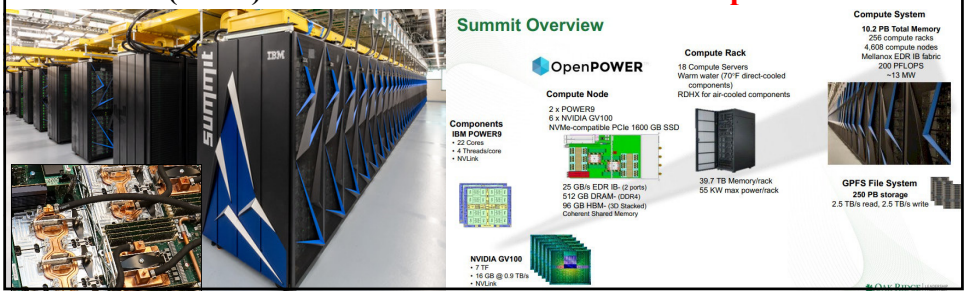
Summit - USA: N°1 June 2018

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

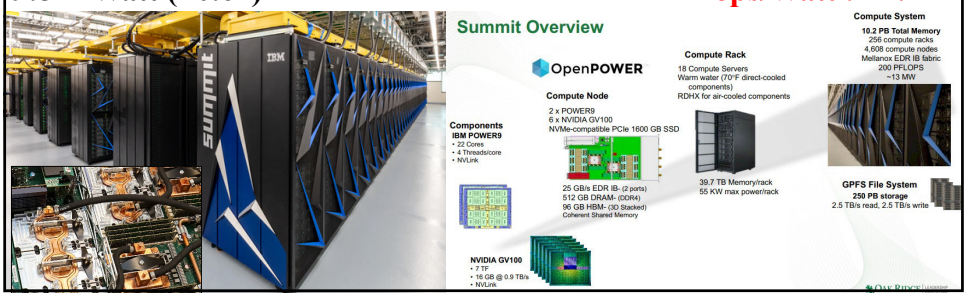
Summit - USA: N°1 November 2018

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

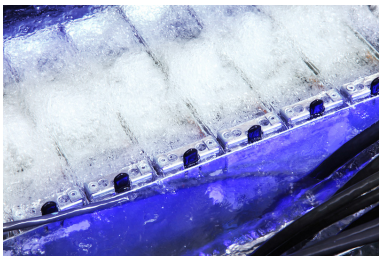
Appendix

39

CentraleSupélec

Cooling

Cooling is close to 30% of the energy consumption



Optimization is mandatory!

40

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 m² au sol)

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

Souvent estimé à 30% de la dépense énergétique!

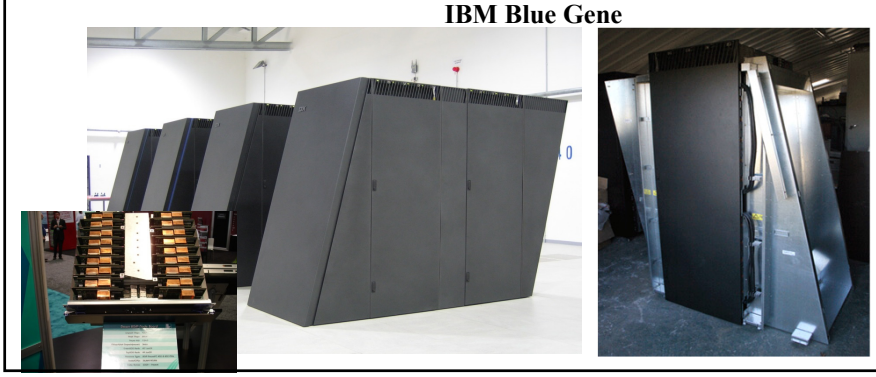
Cooling

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**

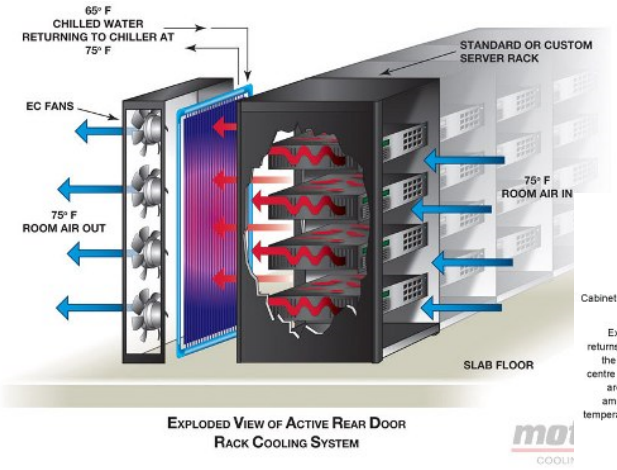
IBM Blue Gene



42

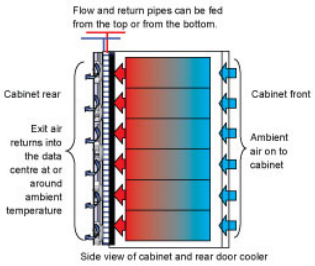
Cold doors (air+water cooling)

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



EXPLODED VIEW OF ACTIVE REAR DOOR
RACK COOLING SYSTEM

Le refroidissement se concentre sur l'armoire.



Side view of cabinet and rear door cooler

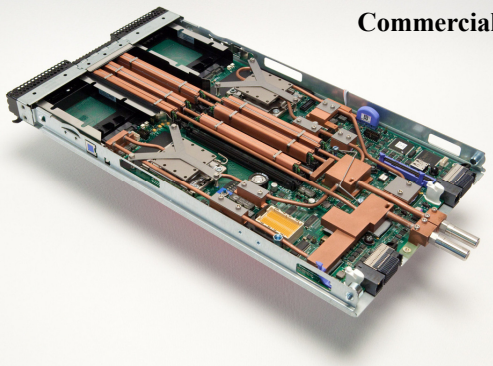
Direct liquid cooling

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)

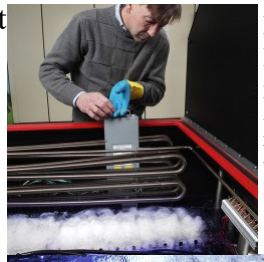
Lame de calcul IBM en 2012 Commercialisée



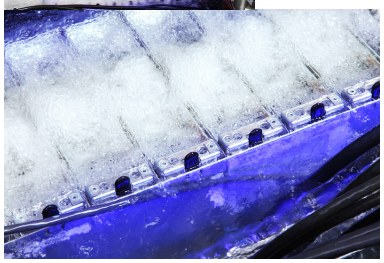
Cooling

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

Cooling

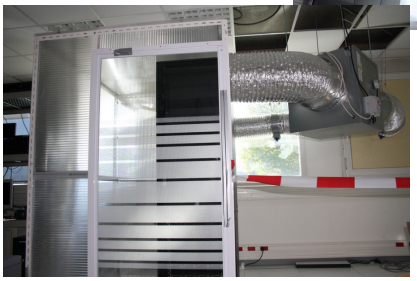
Extreme air cooling

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

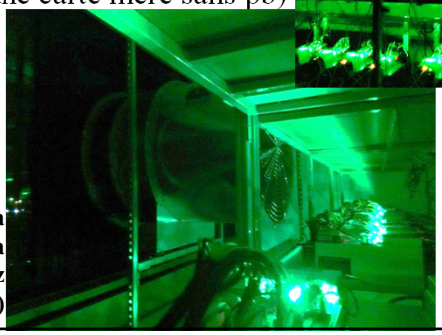
Extreme air cooling

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 Iliium 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