Plan Formation Démarrage Rapide:

- Introduction : Mésocentre CALMIP
  - Concepts fondamentaux
    - Introduction à l'Architecture des systèmes HPC
      - *Système à mémoire partagée / distribuée*
      - *Architecture Processeurs/ Accélérateurs*
      - *Présentation système de Calcul CALMIP : EOS*
      - *Visite salle Machine*

Scientific computation

❑ numerical simulation thanks to Scientific computation :

- ❑ Quantum Chemistry: Gaussian, deMon
- ❑ Atomic scale Material : VASP,Wien2K, CP2K
- ❑ Molecular Dynamic : Amber, GROMACS, ...
- ❑ Genomic : Blast, vina, (CHDB)
- ❑ Crash : LS-Dyna, Radioss
- ❑ Fluid Dynamic : Jadim, Neptune_CFD, OpenFOAM, Fluent, STARCCM+
- ❑ Mechanics : ABAQUS, GETFEM++

Computer program

✓ Efficient (optimum?) execution of computer program

✓ Time to solution !!!

High Performance Computing

Mathematics

✓ Discretisation : spatial (FEM, Finite Volume, time (Explicit, implict scheme, ...

✓ Numerical Methods (Solve Linear Systems, Algebra, ...)

❑ Scientific Computation : transverse (all disciplines)....

Top500 November 2014

Calcul Haute Performance : TOP 500 List

#1 2014
#1 2013
#1 Nov. 2012
#1 2012
#1 2011
Moore's Law ?

17 MW
8 MW
7 MW
12 MW

RANK	SITE	SYSTEM	CORES	RMAX (TFLOP/S)	RPEAK (TFLOP/S)	POWER (KW)
1	National Super Computer Center in Guangzhou China	**Tianhe-2 (MilkyWay-2)** - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31S1P NUDT	3,120,000	33,862.7	54,902.4	17,808
2	DOE/SC/Oak Ridge National Laboratory United States	**Titan** - Cray XK7 , Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x Cray Inc.	560,640	17,590.0	27,112.5	8,209
3	DOE/NNSA/LLNL United States	**Sequoia** - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom IBM	1,572,864	17,173.2	20,132.7	7,890
4	RIKEN Advanced Institute for Computational Science (AICS) Japan	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect Fujitsu	705,024	10,510.0	11,280.4	12,660
5	DOE/SC/Argonne National Laboratory United States	**Mira** - BlueGene/Q, Power BQC 16C 1.60GHz, Custom IBM	786,432	8,586.6	10,066.3	3,945
6	Swiss National Supercomputing Centre (CSCS) Switzerland	**Piz Daint** - Cray XC30, Xeon E5-2670 8C 2.600GHz, Aries interconnect , NVIDIA K20x Cray Inc.	115,984	6,271.0	7,788.9	2,325
7	Texas Advanced Computing Center/Univ. of	**Stampede** - PowerEdge C8220, Xeon	462,462	5,168.1	8,520.1	4,510
26	Grand Equipement National de Calcul Intensif - Centre Informatique National de l'Enseignement Suprieur (GENCI-CINES) France	**Occigen** - bullx DLC, Xeon E5-2690v3 12C 2.6GHz, Infiniband FDR Bull SA	50,544	1,628.8	2,102.6	935
	United States	1.600GHz, Custom Interconnect IBM				
10	Government United States	Cray CS-Storm, Intel Xeon E5-2660v2 10C 2.2GHz, Infiniband FDR, Nvidia K40 Cray Inc.	72,800	3,577.0	6,131.8	1,499

http://www.calmip.univ-toulouse.fr/

Top500 November 2015

#1 2015
#1 2014
#1 2013
#1 Nov. 2012
#1 2012

RANK	SITE	SYSTEM	CORES	RMAX (TFLOP/S)	RPEAK (TFLOP/S)	POWER (KW)
1	National Super Computer Center in Guangzhou China	**Tianhe-2 (MilkyWay-2)** - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31S1P NUDT	3,120,000	33,862.7	54,902.4	17,808
2	DOE/SC/Oak Ridge National Laboratory United States	**Titan** - Cray XK7 , Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x Cray Inc.	560,640	17,590.0	27,112.5	8,209
3	DOE/NNSA/LLNL United States	**Sequoia** - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom IBM	1,572,864	17,173.2	20,132.7	7,890
4	RIKEN Advanced Institute for Computational Science (AICS) Japan	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect Fujitsu	705,024	10,510.0	11,280.4	12,660
5	DOE/SC/Argonne National Laboratory United States	**Mira** - BlueGene/Q, Power BQC 16C 1.60GHz, Custom IBM	786,432	8,586.6	10,066.3	3,945
6	DOE/NNSA/LANL/SNL United States	**Trinity** - Cray XC40, Xeon E5-2698v3 16C 2.3GHz, Aries interconnect Cray Inc.	301,056	8,100.9	11,078.9	
7	Swiss National Supercomputing Centre (CSCS) Switzerland	**Piz Daint** - Cray XC30, Xeon E5-2670 8C 2.600GHz, Aries interconnect , NVIDIA K20x Cray Inc.	115,984	6,271.0	7,788.9	2,325
8	HLRS - Höchstleistungsrechenzentrum Stuttgart Germany	**Hazel Hen** - Cray XC40, Xeon E5-2680v3 12C 2.5GHz, Aries interconnect Cray Inc.	185,088	5,640.2	7,403.5	
9	King Abdullah University of Science and Technology Saudi Arabia	**Shaheen II** - Cray XC40, Xeon E5-2698v3 16C 2.3GHz, Aries interconnect Cray Inc.	196,608	5,537.0	7,235.2	2,834
10	Texas Advanced Computing Center/Univ. of Texas United States	**Stampede** - PowerEdge C8220, Xeon E5-2680 8C 2.700GHz, Infiniband FDR, Intel Xeon Phi SE10P Dell	462,462	5,168.1	8,520.1	4,510

Espace Clément Ader
http://www.calmip.univ-toulouse.fr/

Calcul Intensif et Système HPC

➢ Calcul Intensif : Principes des systèmes HPC

- Hardware et Software => Performance Calcul Flottants
  - Flop/s + Mémoire (RAM et espace fichier)
  - Stockage, I/O
  - €++ => Mutualisation
- Flop : floating operation (mult, add) ⇒ opération sur les nombres à virgule flottantes (nombre réels)
  - 3,14159
  - -6,8675456342 E+08
- Plusieurs Utilisateurs d'un même serveur
  - Partage des ressources cohérent : Règles Utilisation
  - OS performant : Multi-applications, Multi-User
- Serveur Totalement Dédié au Calcul
  - Applications Scientifiques Calcul uniquement
  - Sauvegarde
  - Espace Fichier / Stockage
  - Accès distant
- contrainte d'hébergement lourdes :
  - Electricité, (secouru)
  - Refroidissement
  - Poids
  - Sécurité
  - ....

Panorama Systèmes HPC

Machine à Mémoire Partagée :

- Multiprocesseurs
- Un seul espace d'adressage
- Mémoire partagée

Symetric Multi-Processing
SMP

UMA
Accès Mémoire Uniforme

NUMA
Accès Mémoire Non Uniforme

PROGRAMME

Machine à Mémoire Distribuées :

- Multi-Ordinateurs
- Espace d'Adressage Multiple

Massively Parallel Processing
MPP
Clusters

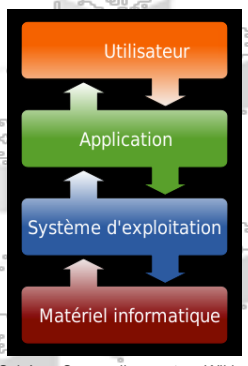
NORMA
no-remote memory access

Système d'Exploitation : Définition

Système d'Exploitation (ou Operanding System « OS » en anglais) :

Un ensemble de logiciels ou programmes qui permet d'unifier les ressources matérielles, pour quelles soient utilisables par l'utilisateur.

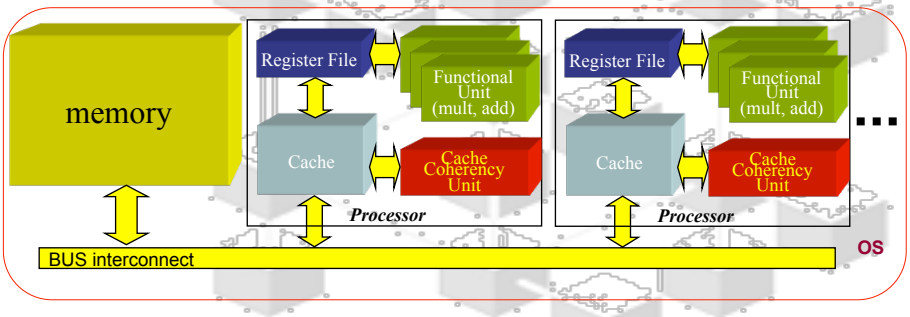
Exemple : Windows, Linux, MacOSX, AIX (IBM), HPUX (HP), SOLARIS (SUN)



(Schéma Operanding system Wikipedia)

UMA Architecture (Shared Memory)

- Machine side: SMP Symmetric MultiProcessor (SMP)
  - Bus Interconnexion between memory and processors
  - Central memory and I/O : shared by all processors
  - Processors access to the same memory(adress space)
- User side:
  - A single machine (single OS) – several processors – one single space memory adress
  - How to program : extension of sequential programmation

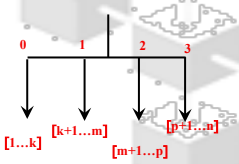


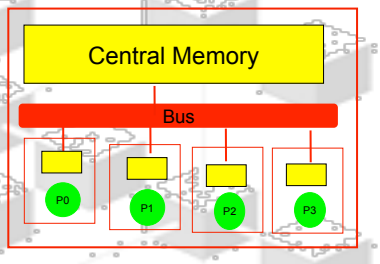
UMA Architecture (Shared Memory) - Multithreading

➢ Parallel Programming with Shared Memory Architecture : OpenMP

AUTOMATIC LOOP PARALLELISATION !!!!

```
!SOMP DO PARALLEL
do i = 1, n
a(i) = 92290. + real(i) ;
end do
!SOMP END D0
```

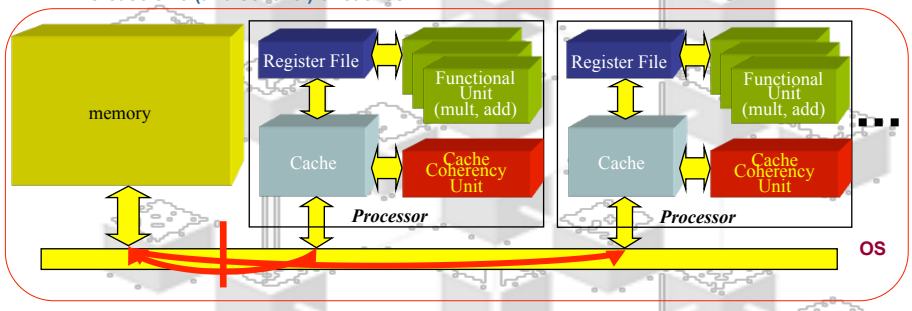




Automatic : spread loop's iterations on cores

UMA Architecture

- Memory access:
  - Concurrent access to central memory => bottleneck
    - Time access increase.
  - Increase size (and so level) of caches

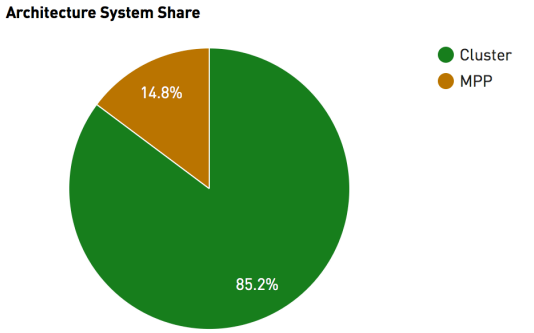


Consequence : few number of processor

•Another paradigma/option : distribute memory ?

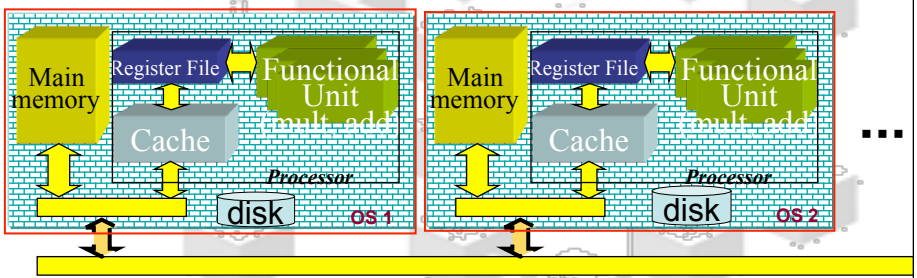
Distributed memory(NORMA)

- Processor and memory tighly interconnected
  - MPP : Massively Parallel Processing
  - Cluster : machines(comput nodes) interconnection



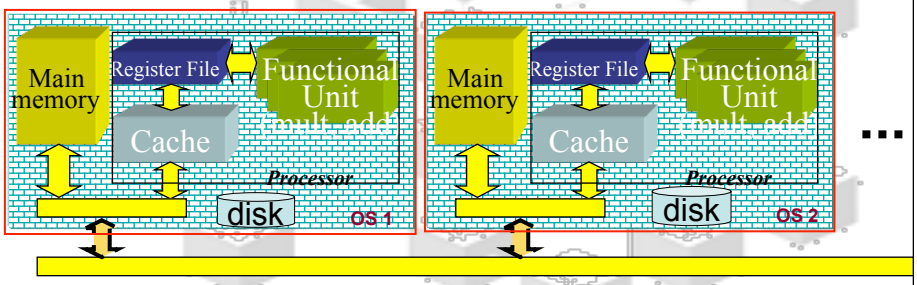
Distributed memory : multi-computer Architecture (Clusters)

- Machine side:
  - Massive technology
  - Process access to its own (local) memory space
  - Interconnect nodes :
    - Like internet (ethernet)…
    - need much faster (bandwidth and latency)
    - process to process communication



Distributed memory : multi-computer Architecture (Clusters)

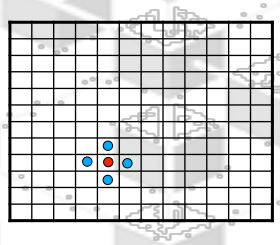
- User side:
  - *n* different nodes (*n* OS) interconnected, 1 (or +) processor per node.
  - Parallel programming ⇨ Message Passing Interface (exchange messages, work done by developper …you?)
  - Need efficient tools to properly access computing ressources



Parallel Programing

➢ Why are we exchanging 'messages' (data) ?

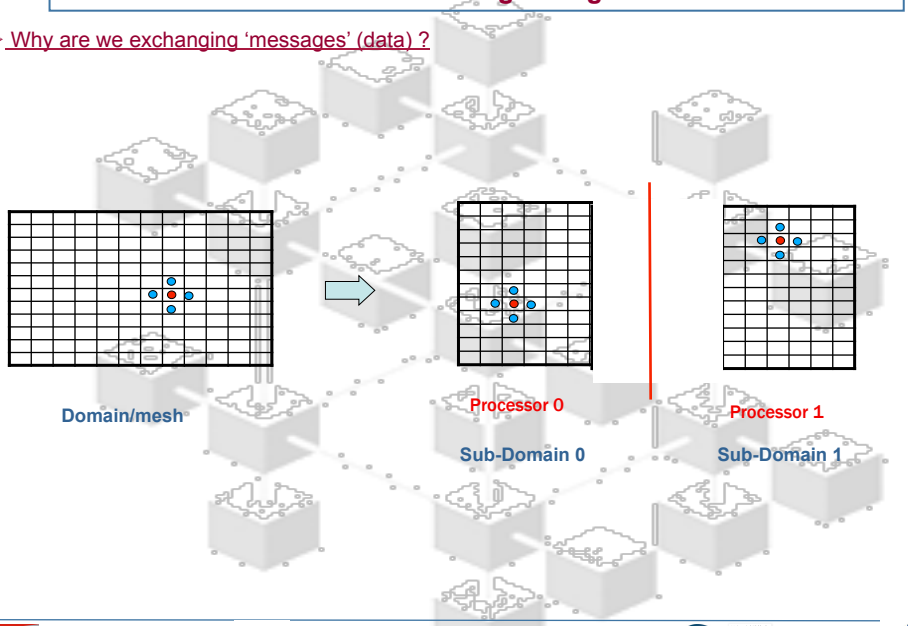
Compute the Red (i,j) point thanks to values of blue points (neighbors)



Domain/mesh/grid

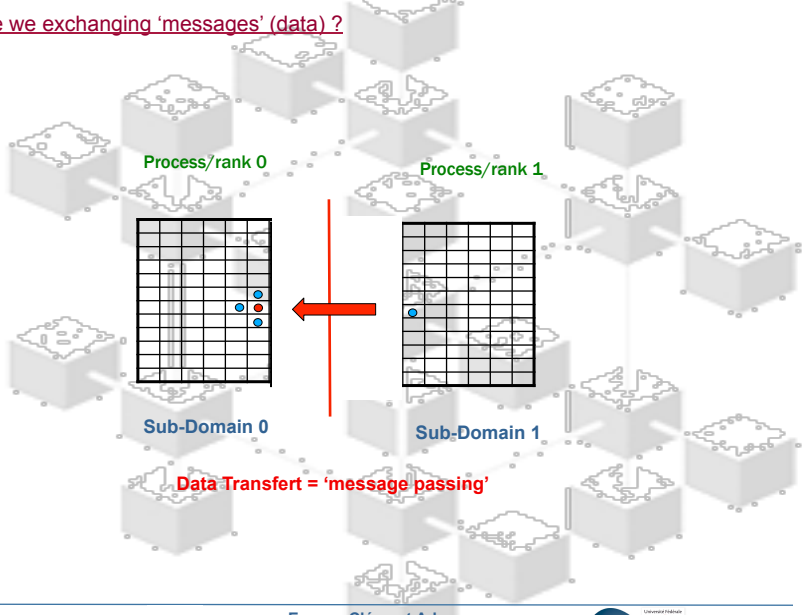
Parallel Programing

➢ Why are we exchanging 'messages' (data) ?

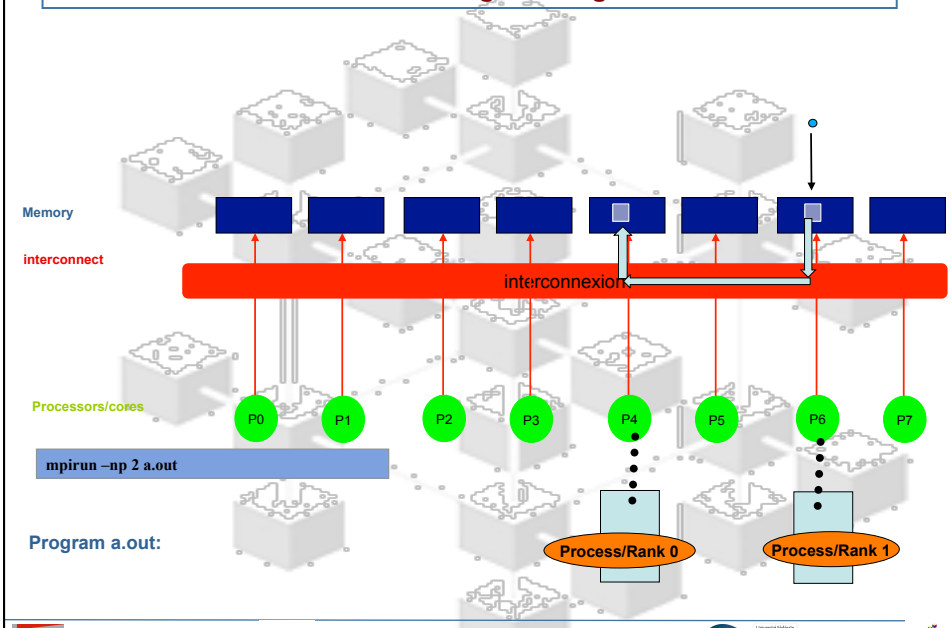


Parallel Programing and Distributed Memory

➢ Why are we exchanging 'messages' (data) ?



Parallel Programming



Parallel Programming

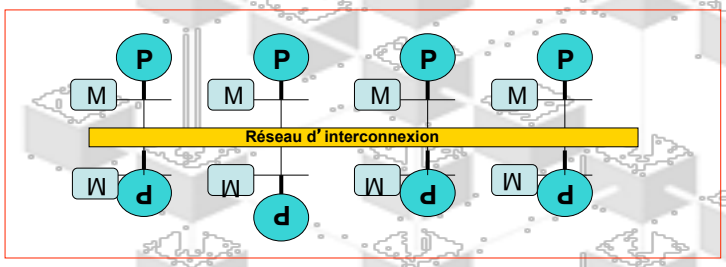
➢SEND/RECV routine

```
program point_to_point
USE MPI

call MPI_INIT(code)
call MPI_COMM_RANK(MPI_COMM_WORLD,rank)

 if (rank == 6) then

   value=1000

   call MPI_SEND(value,1,MPI_INTEGER,4,...,MPI_COMM_WORLD)

 elseif (rank == 4) then

   call MPI_RECV(value,1,MPI_INTEGER,6,...,MPI_COMM_WORLD)

   print *,'I, process',4, 'I recieved', value,' from process 6'

 end if
 call MPI_FINALIZE(code)
end program point_to_point
```

Architecture à mémoire partagée physiquement distribuée

Machine NUMA : Non-Uniform Memory Access

- Mixage SMP/cluster
- Souplesse d'utilisation et performance
- Dépendance « spatiale » par rapport aux ressources utilisées

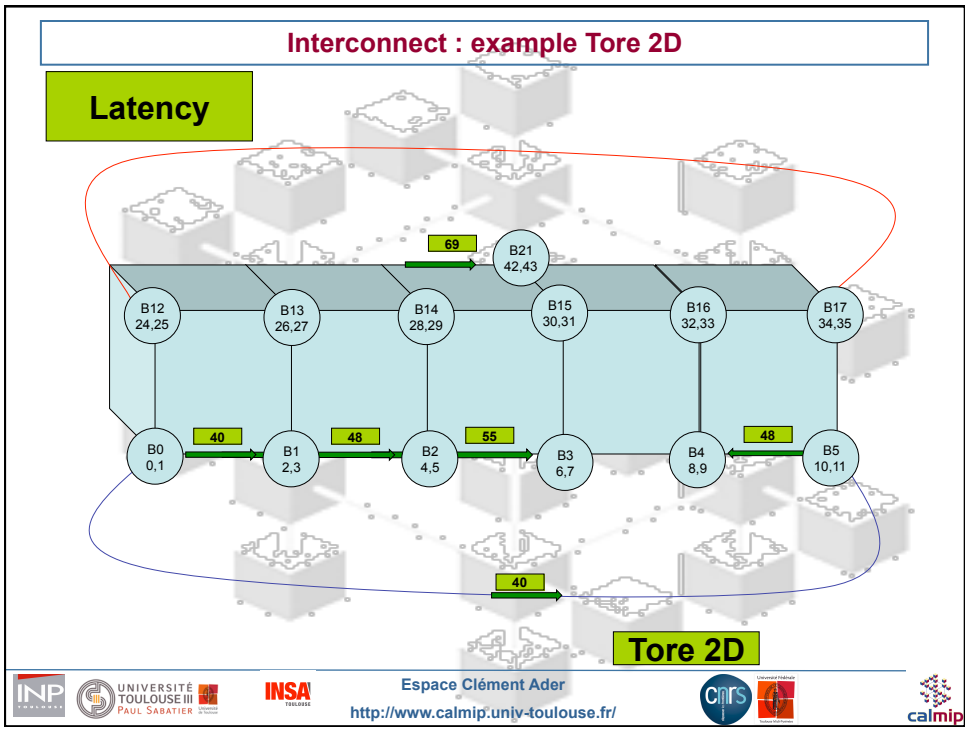


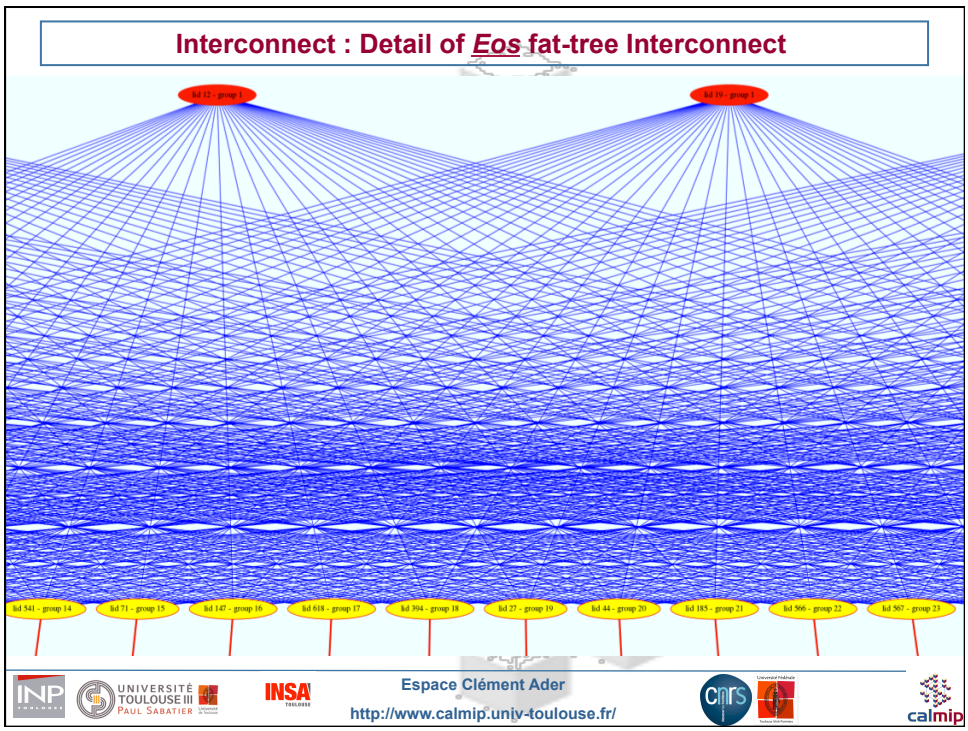
Interconnection : key of parallel machine

- Interconnexion :
  - Latency : how much time to be connected ? Order of microsecond
  - bandwidth (throughput): rate of data transfer ? Mbytes/sec
  - Topology : how many path from a point to another

- In parallel machine hardware (processor / memory) have to be connected
- specific (fast) protocols
  - (cluster) infiniband, proprietary
  - topology:
    - ring, hypercube, Torus, fat-tree, …

1 Mo/s	1 Megaoctet/s	10^6 octets/sec
1 Go/s	1 Gigaoctet/s	10^9 octets/sec
1 To/s	1 Téraoctet/s	10^{12} octets/s

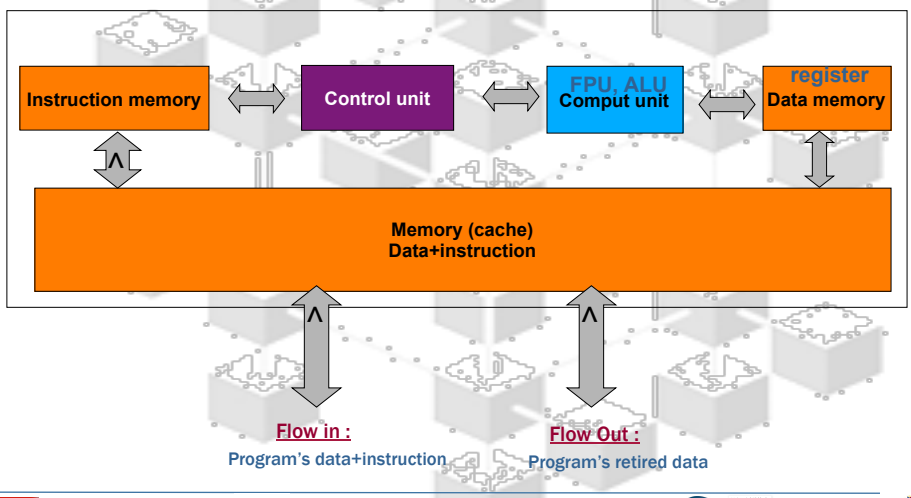

Interconnect : example Tore 2D
Latency
69
B21
42,43
B12
24,25
B13
26,27
B14
28,29
B15
30,31
B16
32,33
B17
34,35
40
48
55
48
B0
0,1
B1
2,3
B2
4,5
B3
6,7
B4
8,9
B5
10,11
40
Tore 2D
Espace Clément Ader
http://www.calmip.univ-toulouse.fr/


Interconnect : Detail of Eos fat-tree Interconnect
Espace Clément Ader
http://www.calmip.univ-toulouse.fr/

- Introduction : Mésocentre CALMIP
  - Concepts fondamentaux
    - Introduction à l'Architecture des systèmes HPC
      - Système à mémoire partagée / distribuée
      - *Architecture Processeurs/ Accélérateurs*
      - Présentation système de Calcul CALMIP : EOS
      - Visite salle Machine

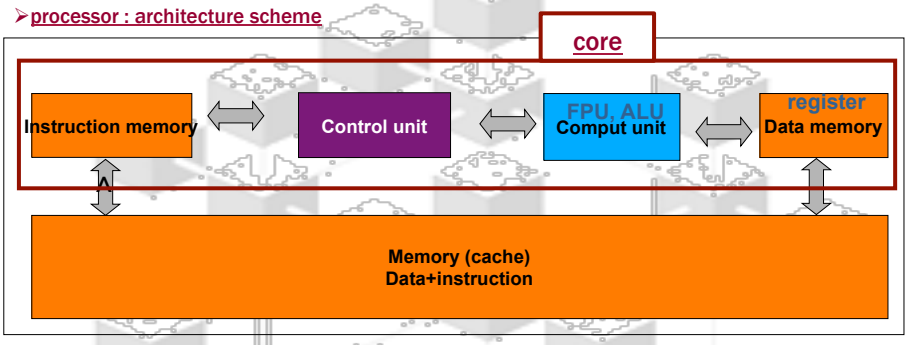
High Performance Computing

➢processor : architecture scheme



High Performance Computing

➢ processor : architecture scheme

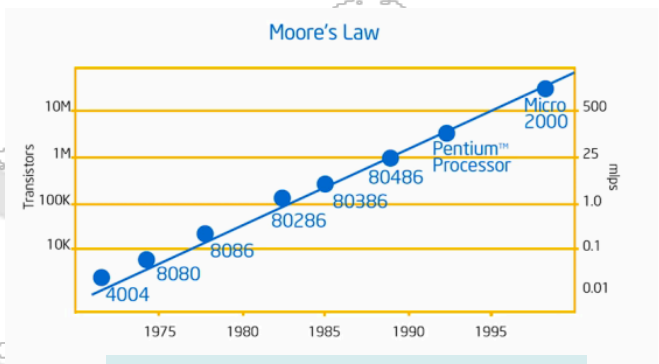


➢ processor cycle

- Clock frequency = number of pulse per second
- 1,5 Ghz ⇒ 1,5 Billion cycles per second
- 1 cycle : perform one « atomic » processor instruction

1,5 Ghz => 1 cycle = 0,6 ns

High Performance Computing



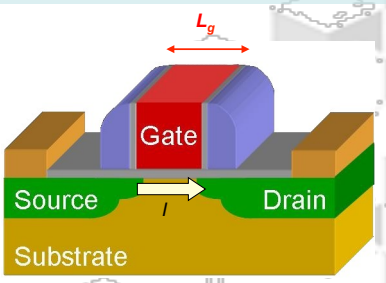
- ❑ « something » double each 18 month
- ❑ « something » = transistors
- ❑ « something » related to « performance »
- ❑ this is an exponentiel Law
  - ❑ could it last ? (True for 30 years)
  - ❑ emperical anyway

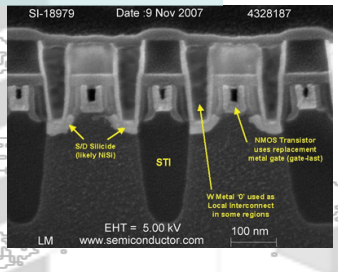
Moore(Intel), Moore'Law is not a Physical law, but depend a lot on … Physics!

http://www.nature.com/nature/journal/v512/n7513/full/nature13570.html

High Performance Computing

❑ A matter of Engraving a piece of Silicon :





❑ Energy Problem

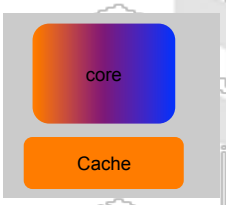
$W = CV^2$ Power = W * F

Consequence : End of the Race for Frequency

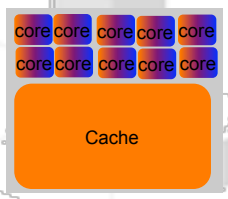
Processors Architecture HPC

➢ Trends nowdays : multi-core

•Fondeurs : Intel, AMD, IBM, FUJITSU(Sparc)



Engrave Shrink :
45 nm, 32nm, 22, nm, 14nm, 7 nm ?



➢ 1 processeur mono-core

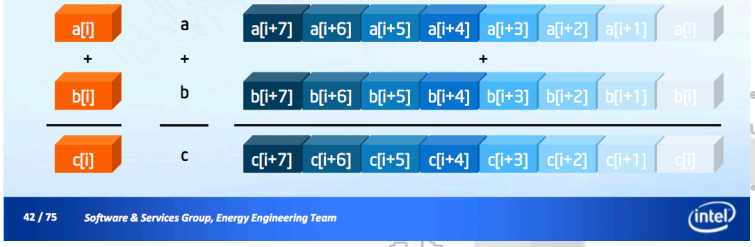
➢ 1 processor (or socket) multi-core
(multi = 2, 4, 6, 8, 10, 12, 20)

➢ More 'RAW' power (×2, 4, 6, 8) :=> parallelism
➢ better ratio flop/watt, flop/m²
➢ same frequency or lower !

Processor Architecture : SIMD

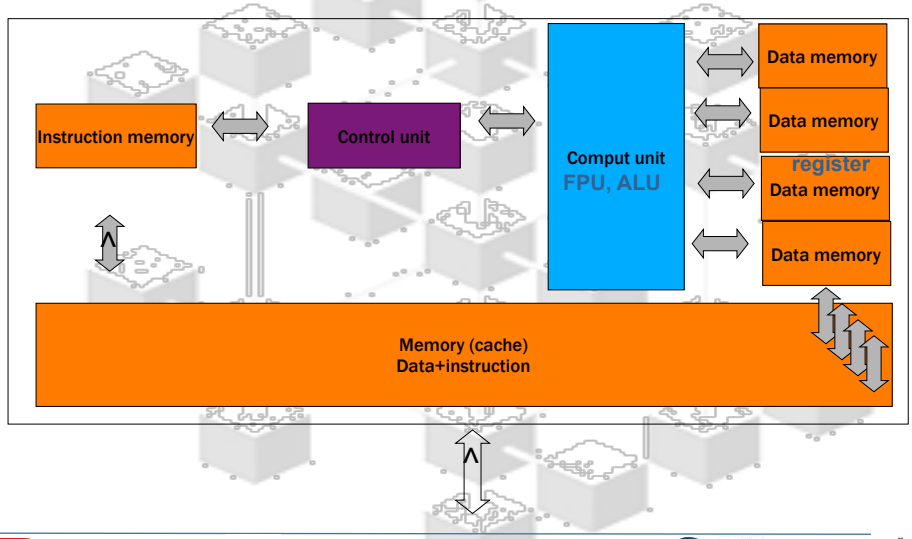
- Vector Processor
  - From computing scalar to comuting Vector
  - SIMD : SINGLE INSTRUCTION MULTIPLE DATA
  - each processor cycle *n* scalar retired operations

```
for(i = 0; i <= MAX;i++)
    c[i] = a[i] + b[i];
```



Processors Architecture : SIMD

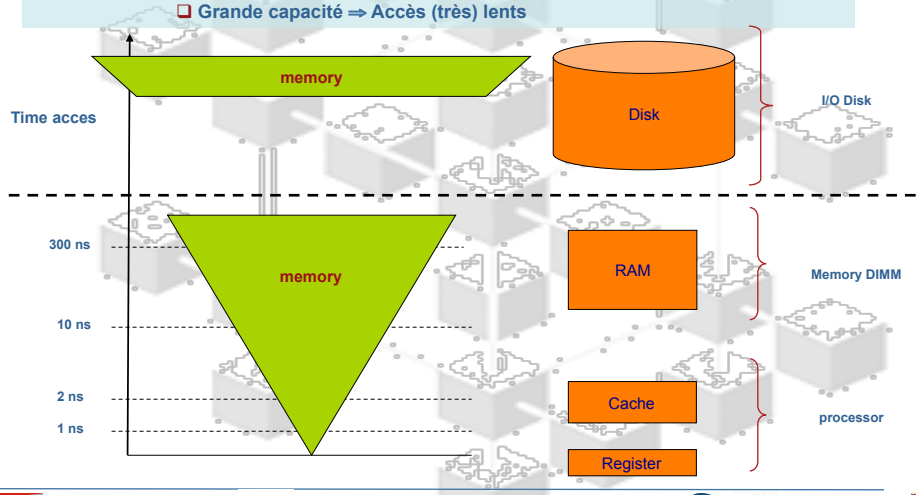
➢ Processor Architecture



High Performance Computing

➢ Mémoire Hierarchique [Hierarchical Memory]

❑ Principe :
- ❑ Faible capacité ⇒ Accès rapides
- ❑ Grande capacité ⇒ Accès (très) lents



Top500 Juin 2015

Calcul Haute Performance : TOP 500 List

#1 2015
#1 2014
#1 2013
#1 Nov. 2012
#1 2012
#1 2011
Moore's Law ?

RANK	SITE	SYSTEM	CORES	RMAX (TFLOP/S)	RPEAK (TFLOP/S)	POWER (KW)
1	National Super Computer Center in Guangzhou China	**Tianhe-2 (MilkyWay-2)** - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31S1P NUDT	3,120,000	33,862.7	54,902.4	17,808
2	DOE/SC/Oak Ridge National Laboratory United States	**Titan** - Cray XK7 , Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x Cray Inc.	560,640	17,590.0	27,112.5	8,209
3	DOE/NNSA/LLNL United States	**Sequoia** - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom IBM	1,572,864	17,173.2	20,132.7	7,890
4	RIKEN Advanced Institute for Computational Science (AICS) Japan	**K computer** SPARC64 VIIIfx 2.0GHz, Tofu interconnect Fujitsu	705,024	10,510.0	11,280.4	12,660
5	DOE/SC/Argonne National Laboratory United States	**Mira** - BlueGene/Q, Power BQC 16C 1.60GHz, Custom IBM	786,432	8,586.6	10,066.3	3,945
6	Swiss National Supercomputing Centre (CSCS) Switzerland	**Piz Daint** - Cray XC30, Xeon E5-2670 8C 2.600GHz, Aries interconnect, NVIDIA K20x Cray Inc.	115,984	6,271.0	7,788.9	2,325
7	King Abdullah University of Science and Technology Saudi Arabia	**Shaheen II** - Cray XC40, Xeon E5-2698v3 16C 2.3GHz, Aries interconnect Cray Inc.	196,608	5,537.0	7,235.2	2,834
8	Texas Advanced Computing Center/Univ. of Texas United States	**Stampede** - PowerEdge C8220, Xeon E5-2680 8C 2.700GHz, Infiniband FDR, Intel Xeon Phi SE10P Dell	462,462	5,168.1	8,520.1	4,510

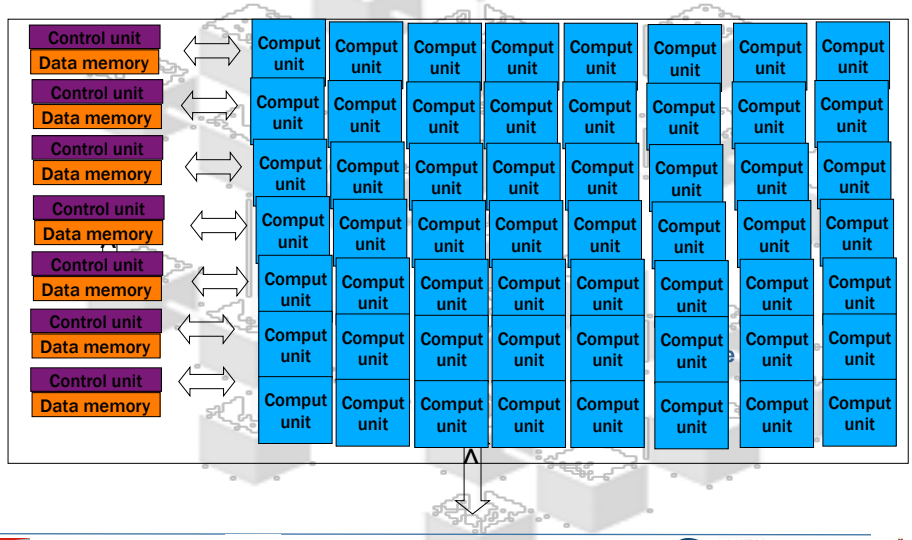
17 MW
8 MW
7 MW
12 MW

Accélérateurs

Nuclear Plant between 40 MW and 1450 MW

Accelerators

➢Accelerator : GP-GPU



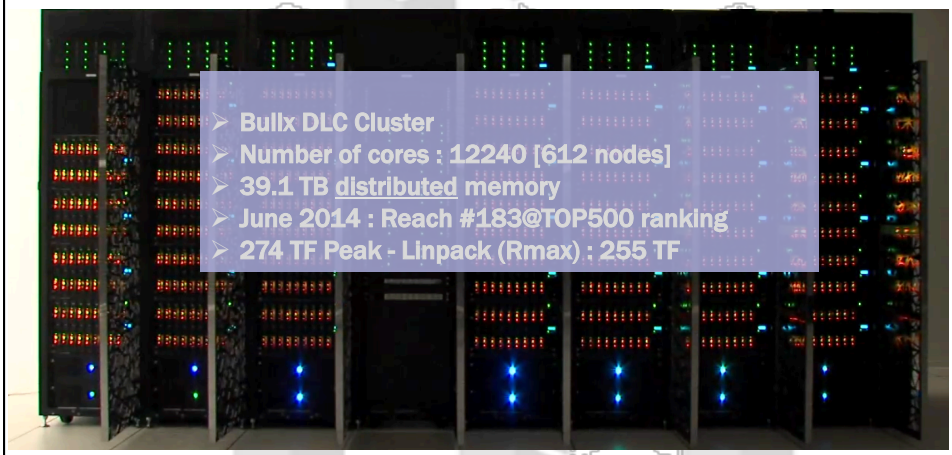
Accelerator : Xeon Phi

- Accelerator 2
  - Intel : Xeon Phi ; 1.074 TF (DP)
  - >50 cores ! MIMD (x86) cores (interprocessor network)
  - Standard Programming languages
  - 1.1 Ghz
  - 512 bit SIMD vectorisation
  - 22 nm ingraving

Plan Formation Démarrage Rapide:

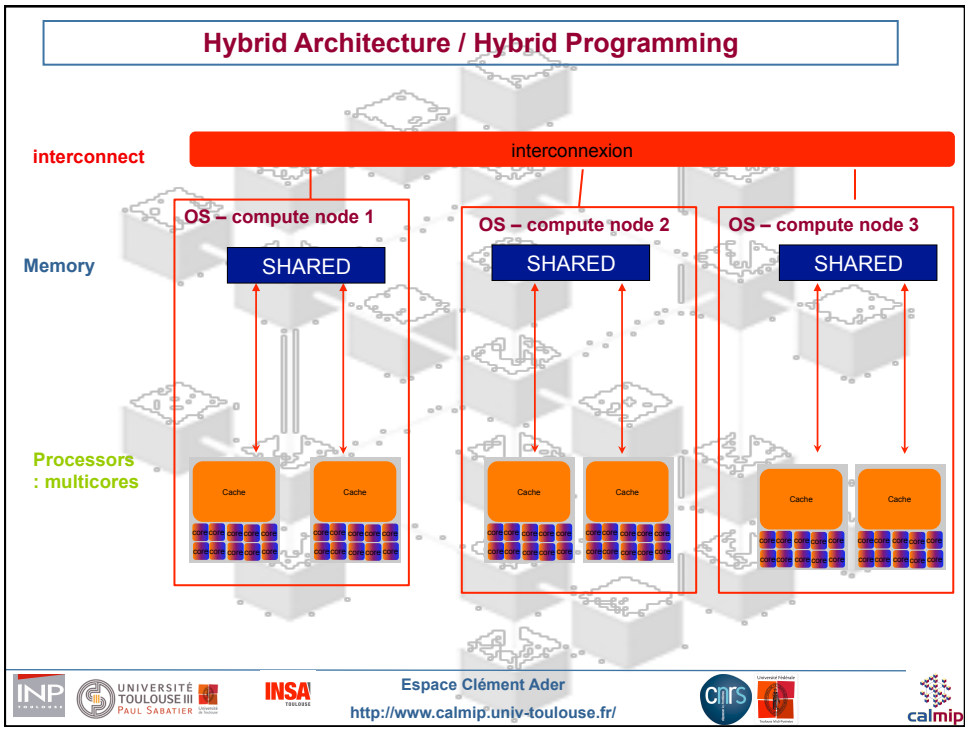
- Introduction : Mésocentre CALMIP
  - Concepts fondamentaux
    - Introduction à l'Architecture des systèmes HPC
      - *Système à mémoire partagée / distribuée*
      - *Architecture Processeurs/ Accélérateurs*
      - *Présentation système de Calcul CALMIP EOS - Visite salle Machine*

(Brand new) CALMIP's Supercomputer : EOS



- Bullx DLC Cluster
- Number of cores : 12240 [612 nodes]
- 39.1 TB distributed memory
- June 2014 : Reach #183@TOP500 ranking
- 274 TF Peak - Linpack (Rmax) : 255 TF

Page web associée :
http://www.calmip.univ-toulouse.fr/spip/spip.php?article388


Hybrid Architecture / Hybrid Programming
interconnect
interconnexion
Memory
OS – compute node 1
SHARED
OS – compute node 2
SHARED
OS – compute node 3
SHARED
Processors
: multicores
Cache
Espace Clément Ader
http://www.calmip.univ-toulouse.fr/