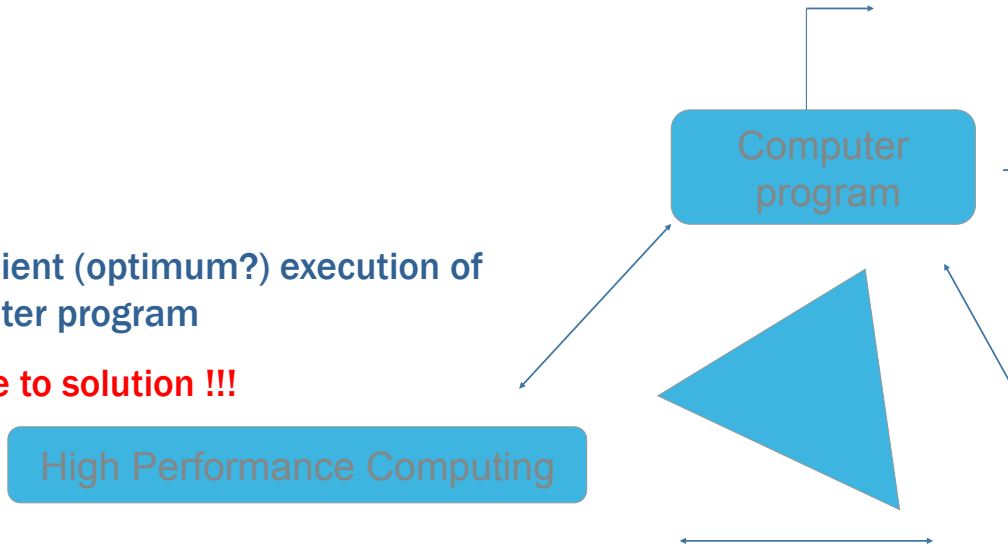
Plan Formation :

- Introduction : Mésocentre CALMIP
- Premier jour
  - Concepts fondamentaux
    - Introduction à l'Architecture des systèmes HPC
      - *Calcul Intensif et Panorama des Systèmes*
      - *Architecture Processeurs/ Accélérateurs*
      - *Présentation système de Calcul CALMIP : OLYMPE*
      - *Visite salle Machine*
    - Introduction programmation sur les systèmes HPC
      - *Programmation Parallèle*
      - *Optimisation de codes*

- Introduction : Mésocentre CALMIP
- Premier jour
  - Concepts fondamentaux
    - Introduction à l'Architecture des systèmes HPC
      - *Calcul Intensif et Panorama des Systèmes*
      - *Architecture Processeurs/ Accélérateurs*
      - *Présentation système de Calcul CALMIP : OLYMPE*
      - *Visite salle Machine*
    - Introduction programmation sur les systèmes HPC
      - *Programmation Parallèle*
      - *Optimisation de codes*

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++

✓ Efficient (optimum?) execution of computer program

✓ Time to solution !!!

✓ 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

calmip

#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	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
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
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

17 MW

8 MW

7 MW

12 MW

Top500 Juin 2018

Rank	Site	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	DOE/SC/Oak Ridge National Laboratory United States	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband IBM	2,282,544	122,300.0	187,659.3	8,806
2	National Supercomputing Center in Wuxi China	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway NRCPC	10,649,600	93,014.6	125,435.9	15,371
3	DOE/NNSA/LLNL United States	**Sierra** - IBM Power System S922LC, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband IBM	1,572,480	71,610.0	119,193.6	
4	National Super Computer Center in Guangzhou China	**Tianhe-2A** - TH-IVB-FEP Cluster, Intel Xeon E5-2692v2 12C 2.2GHz, TH Express-2, Matrix-2000 NUDT	4,981,760	61,444.5	100,678.7	18,482
5	National Institute of Advanced Industrial Science and Technology (AIST) Japan	**AI Bridging Cloud Infrastructure (ABCI)** - PRIMERGY CX2550 M4, Xeon Gold 6148 20C 2.4GHz, NVIDIA Tesla V100 SXM2, Infiniband EDR Fujitsu	391,680	19,880.0	32,576.6	1,649
6	Swiss National Supercomputing Centre (CSCS) Switzerland	**Piz Daint** - Cray XC50, Xeon E5-2690v3 12C 2.6GHz, Aries interconnect , NVIDIA Tesla P100 Cray Inc.	361,760	19,590.0	25,326.3	2,272

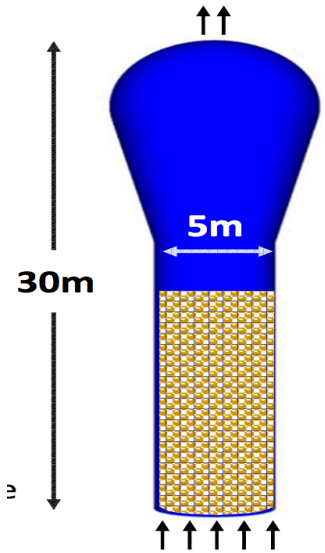
#1 2016

8 MW

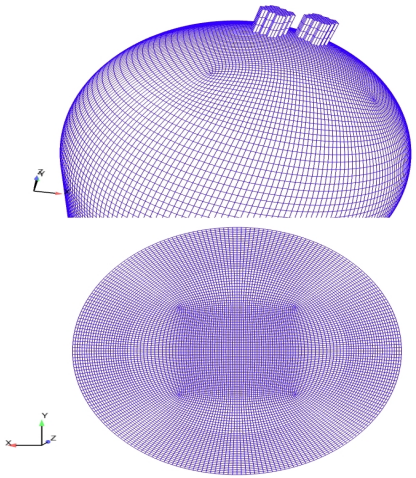
15 MW

Mesochallenges on OLYMPE system

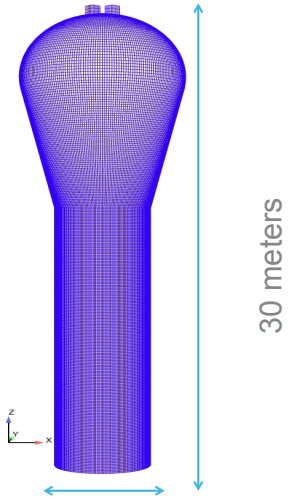
- Fluidized Bed / Particules Injection/Petroleum Industry
- Mesh(problem size) : 1 000 Million of Elements
- Parallel code : Neptune_CFD (EDF)



81 Tons of particles



3D mesh : 100 M elements



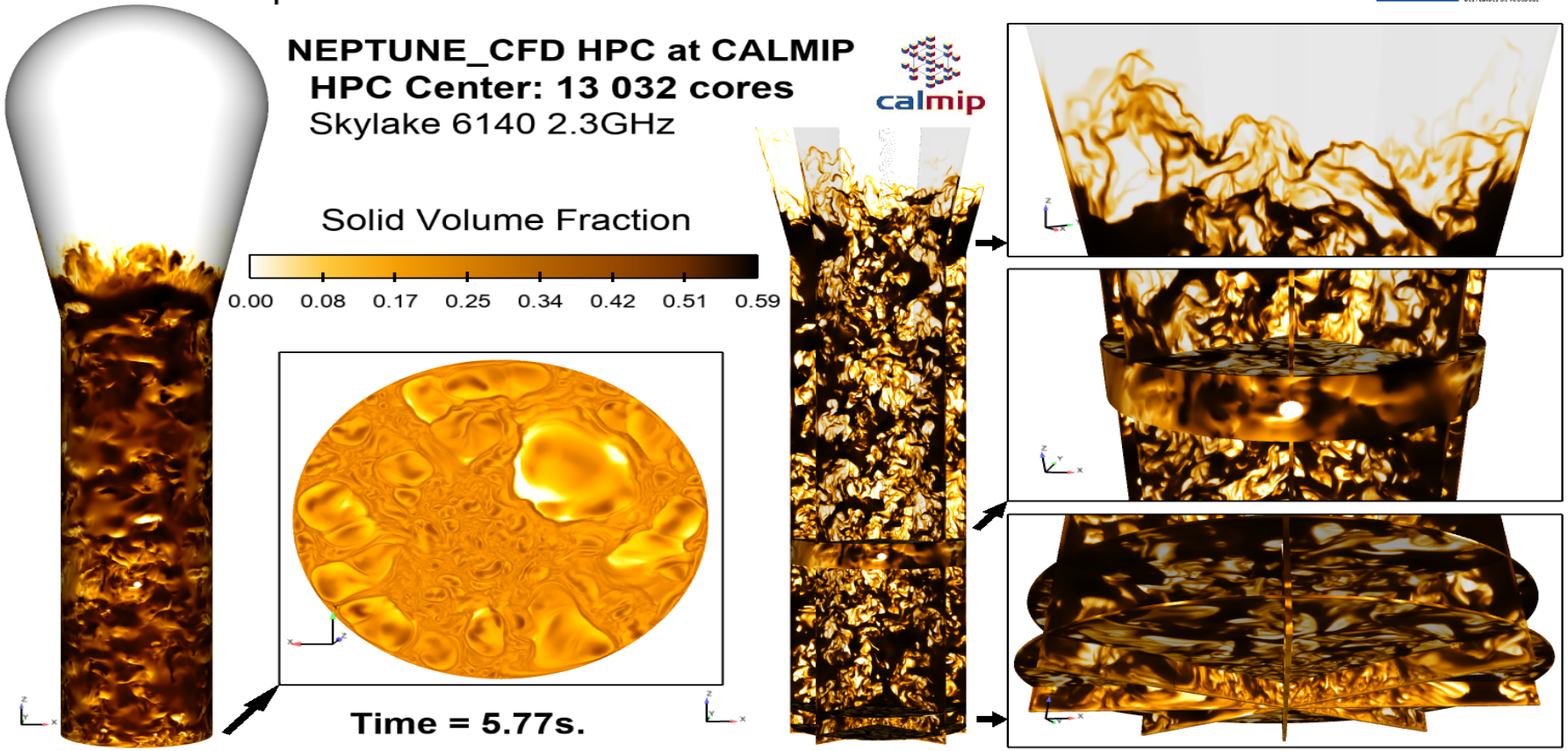
5 meters

Solid Volume Fraction

Courtesy of : H. neau, P. Fede O Simonin, H. Neau - Institut de Mécanique des fluides de Toulouse - Université de Toulouse/ CNRS

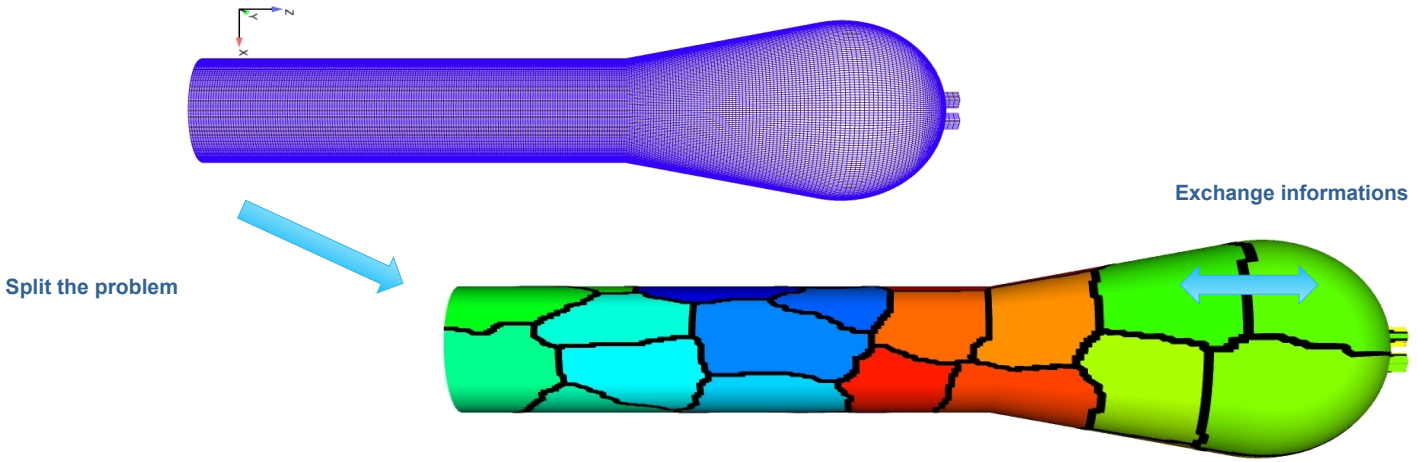
Industrial Scale Bidispersed Reactive Fluidized Bed Reactor

100 tonnes of particles - D~5m - H~30m - Unstructured Mesh: > 1 billion cells



Algorithm Aspect : split the problem

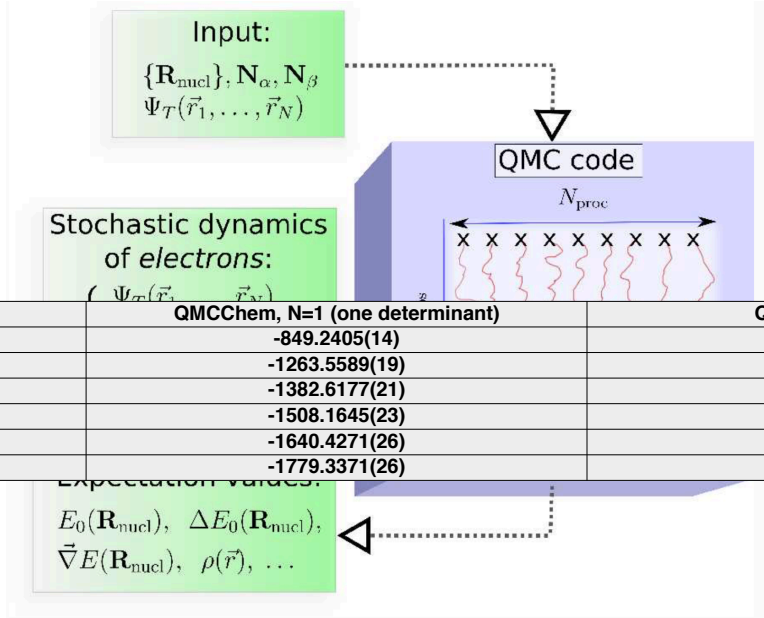
- One Domain (Mesh, i.e. 1 000 000 000 element)
  - Basic Idea : Split in subdomain (create a partition of the inital mesh)
  - Each core/processor work on a sub-domain
    - Namely 100 cores, 10 000 000 element each



- Mesh(domain) partitioned in sub mesh (domain)

Mesochallenges on Eos system

- Computational Quantum Chemistry
- Parallel code : QMC_CHEM > 10,000 cores
- Almost Embarrassingly Parallel / Highly Tuned Code : vectorisation
- Accurate nonrelativistic ground-state energies of 3d transition metal atoms



Atome	Sarsa 2013	QMCChem, N=1 (one determinant)	QMCChem N determinant, N large)
Ti	-849.2492(7)	-849.2405(14)	-849.2841(19) N=14054
Fe	-1263.5607(6)	-1263.5589(19)	-1263.5868(21) N=13171
Co	-1382.6216(8)	-1382.6177(21)	-1382.6377(24) N=15949
Ni	-1508.1743(7)	-1508.1645(23)	-1508.1901(25) N=15710
Cu	-1640.4266(7)	-1640.4271(26)	-1640.4328(29) N=48347
Zn	-1779.3425(8)	-1779.3371(26)	-1779.3386(31) N=44206

- Courtesy of : M. caffarel, T. Applencourt, A. Scemama –LCPQ –IRSAMC - Université de Toulouse/ CNRS

Astrophysique

- Mesure des vitesses horizontales de la surface solaire (Coherent Structure Tracking)T. Roudier (IRAP)
  - Code Fortran, Parallélisé mémoire partagée (OpenMP)
  - Calculs sur 32 cores à 64 cores / 90 000 h_cpu consommées en 2017

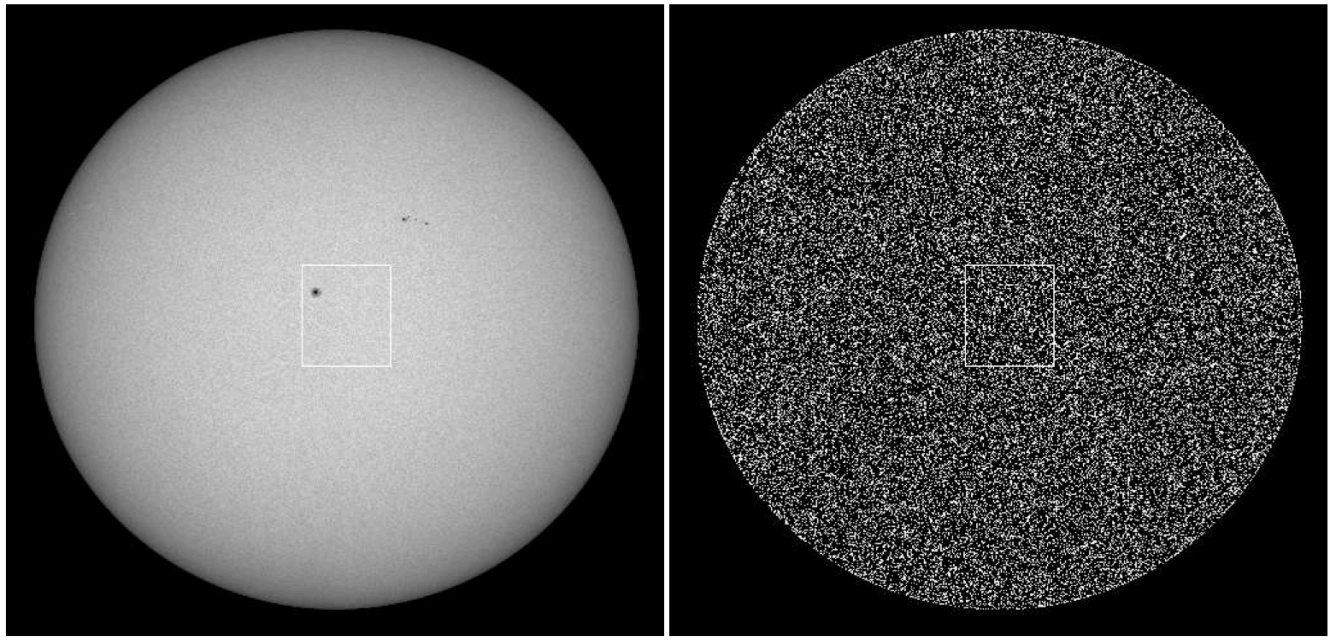


Fig. 1. Full Sun HMI/SDO white Light on August 30, 2010 (left) and the segmented map where around 500 (88 granules are detected right).

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
- ❑ Plusieurs Utilisateurs d'un même serveur
  - ❑ Partage des ressources cohérent : Règles Utilisation
  - ❑ OS performant : Multi-applications, Multi-User

- ❑ Flop : floating operation (mult, add) ⇒ opération sur les nombres à virgule flottantes (nombre réels)
  - ❑ 3,14159
  - ❑ -6,8675456342 E+08

- ❑ 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

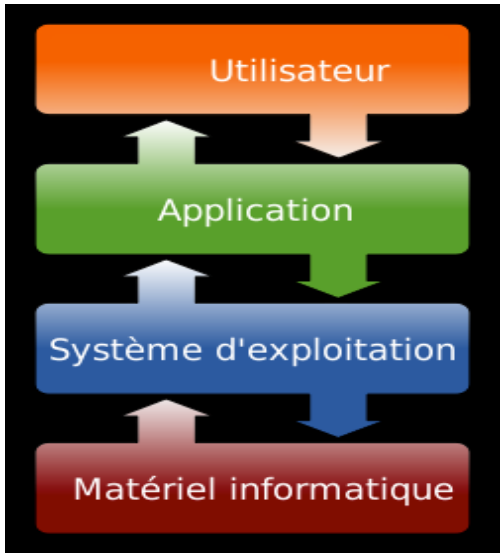
NORMA
no-remote memory access

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.

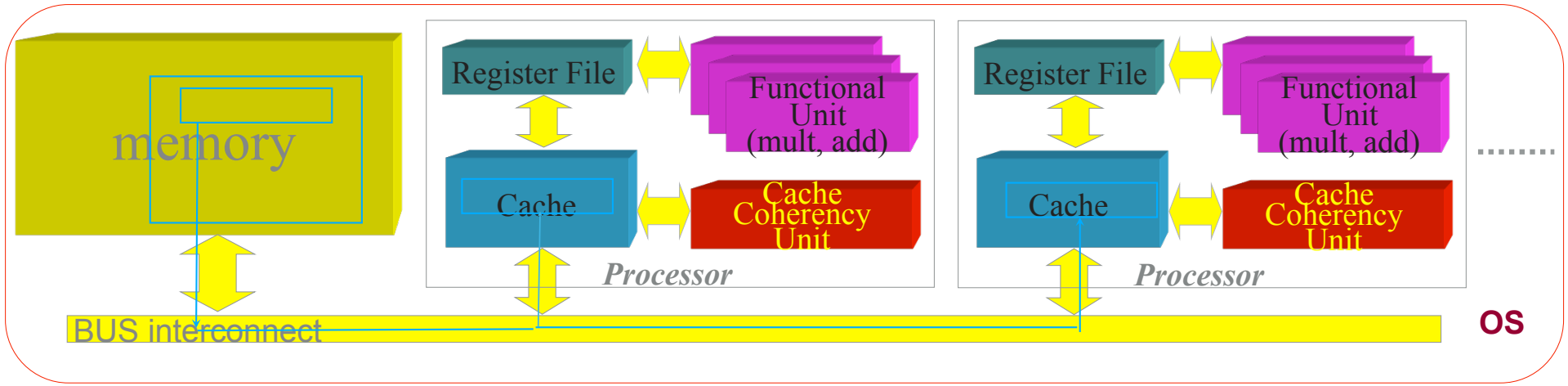
Exemple : Windows, Linux, MacOSX, AIX (IBM), HPUX (HP), SOLARIS (SUN)



(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

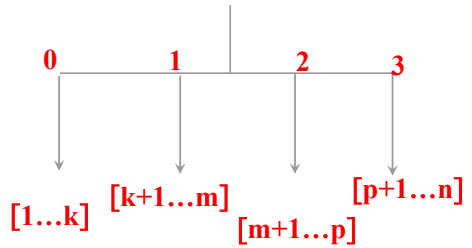


UMA Architecture (Shared Memory) - Multithreading

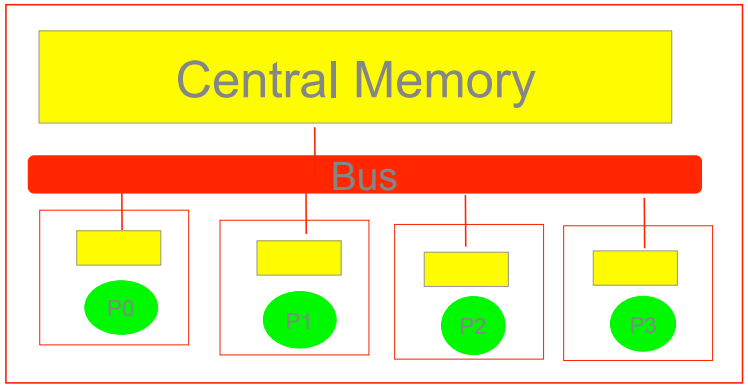
➢ Parallel Programing with Shared Memory Architecture : OpenMP

AUTOMATIC LOOP PARALLELISATION !!!!

```
!$OMP DO PARALLEL
do i = 1, n
a(i) = 92290. + i
end do
!$OMP END D0
```

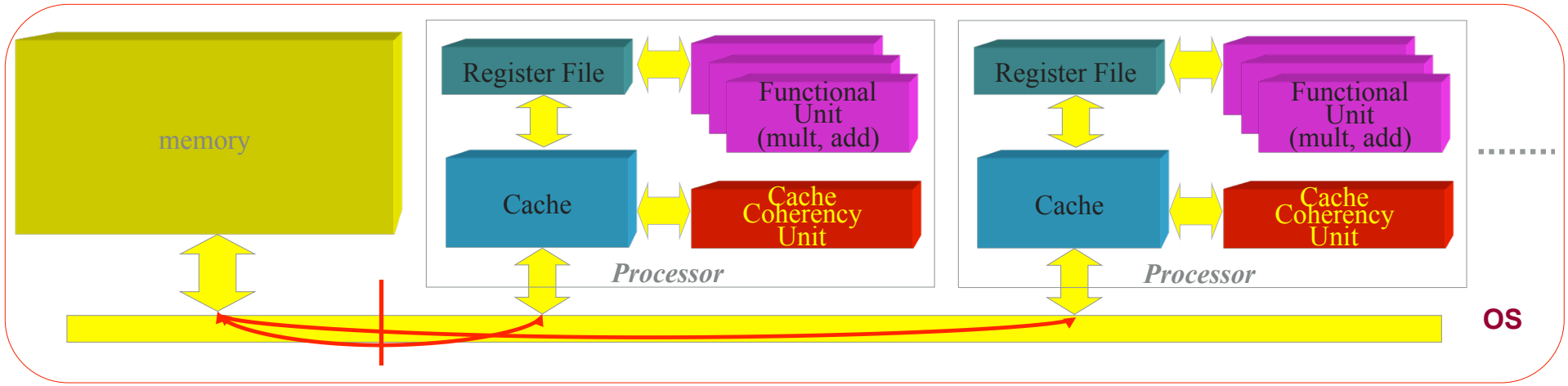


Automatic : spread loop's iterations on cores



UMA Architecture

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



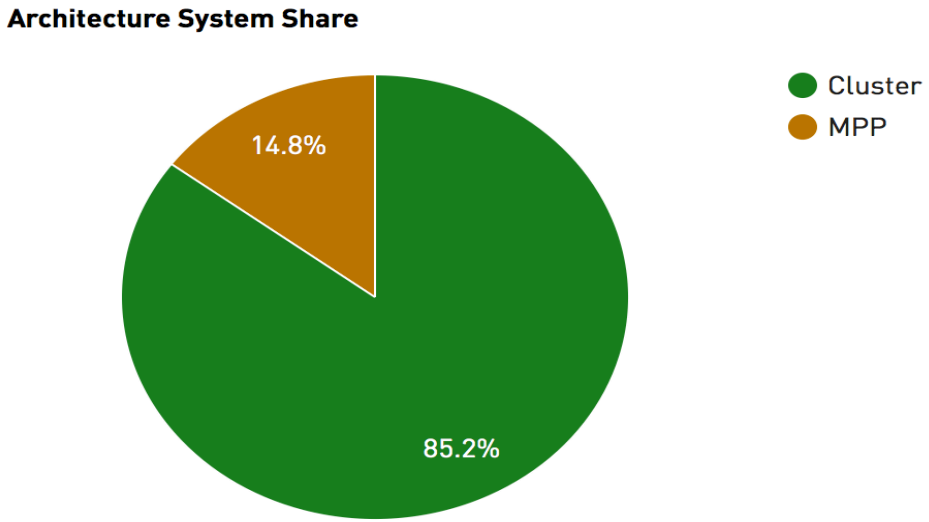
Consequence : few number of processor

•**Another paradigma/option : distribute memory ?**

Distributed memory (NORMA)

calmip

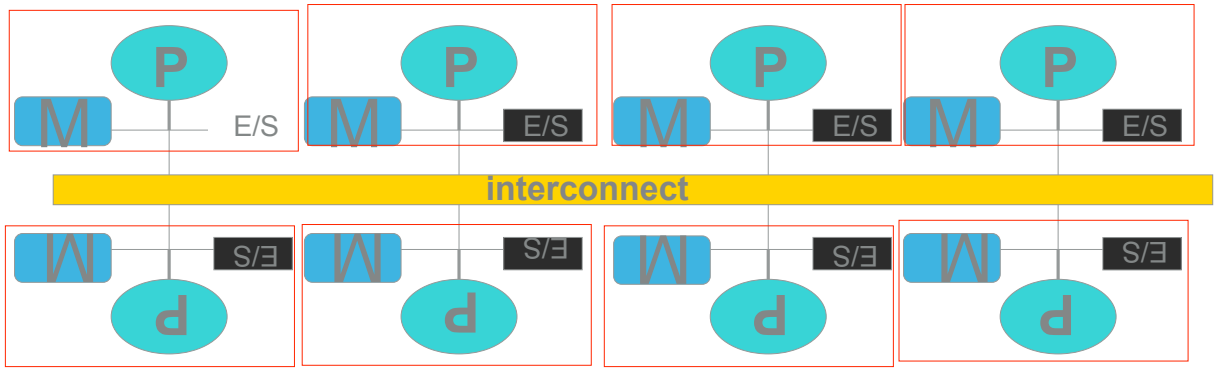
- Processor and memory tighly interconnected
  - MPP : Massively Parallel Processing
  - Cluster : machines (comput nodes) interconnection



Distributed memory

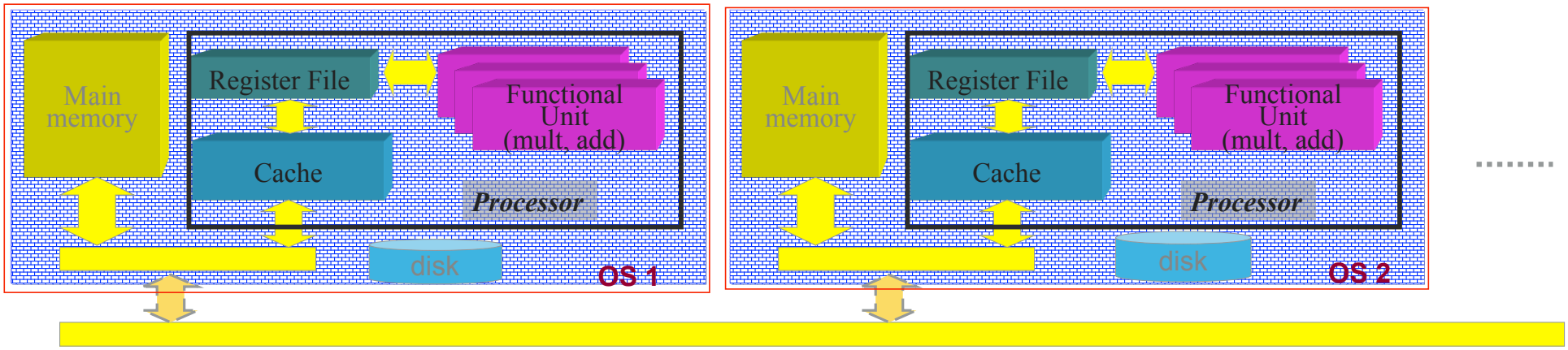
Cluster : massive technology

- A lot of processor ….
- A lot of machine (nodes) interconnected
- *nodes: multi-processors, multi-core*



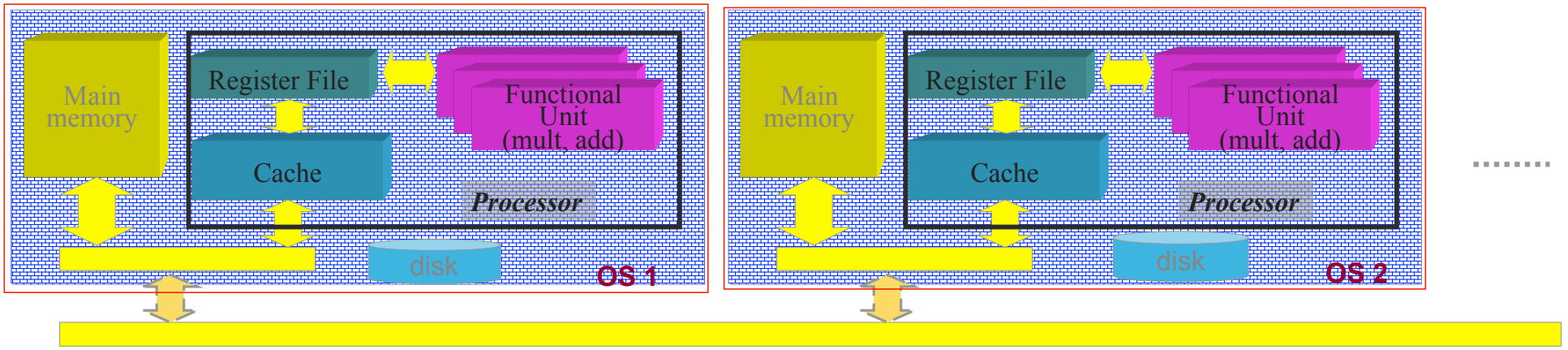
Distributed memory : multi-computer Architecture (Clusters)

- **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



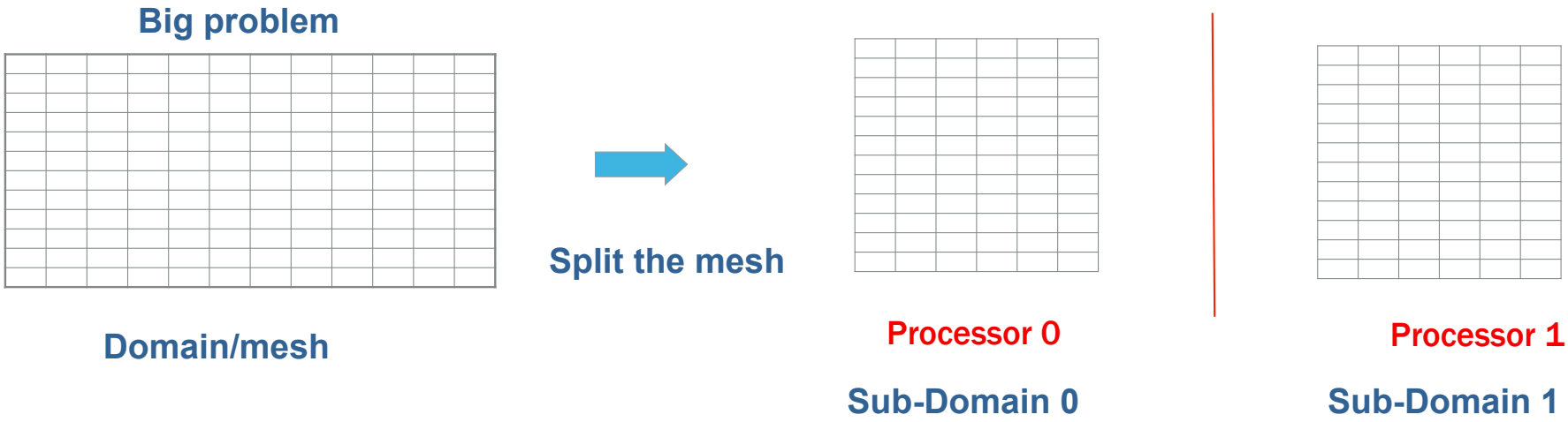
Distributed memory : multi-computer Architecture (Clusters)

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



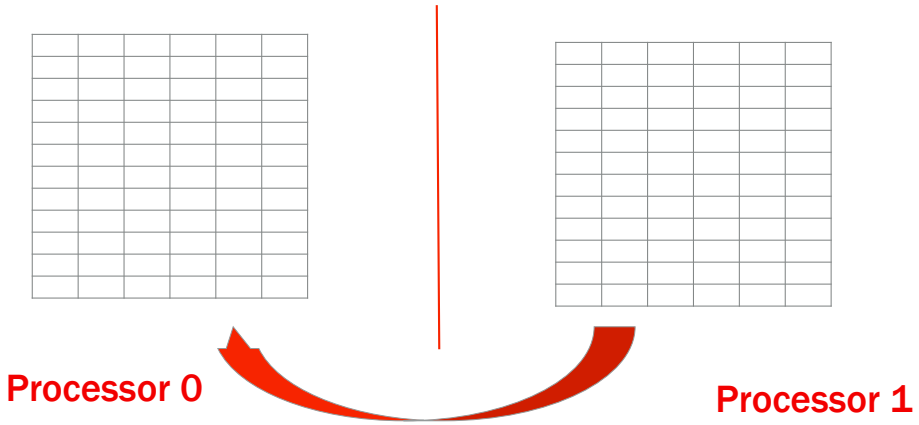
Parallel Programing and Distributed Memory

➢ Distributed Architecture : why are we (must we) exchanging ‘messages’ (data) ?

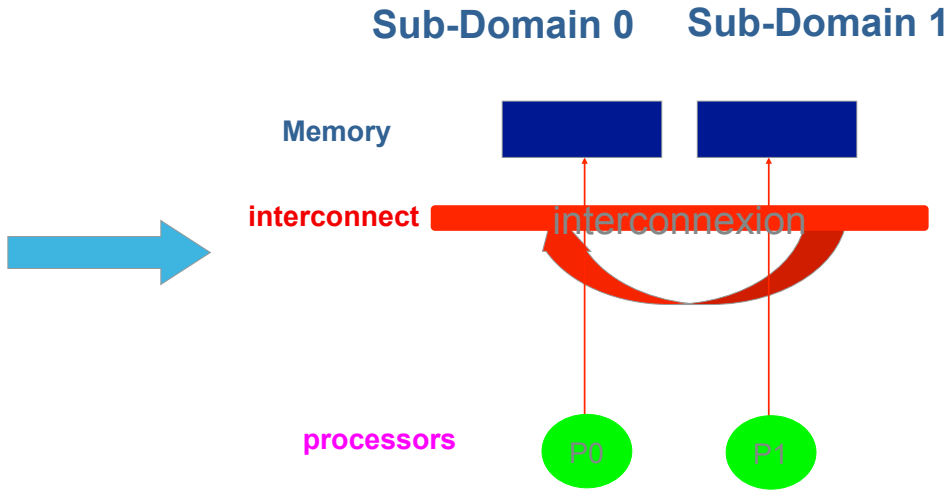


Parallel Programing and Distributed Memory

➢ Why are we exchanging ‘messages’ (data) ?

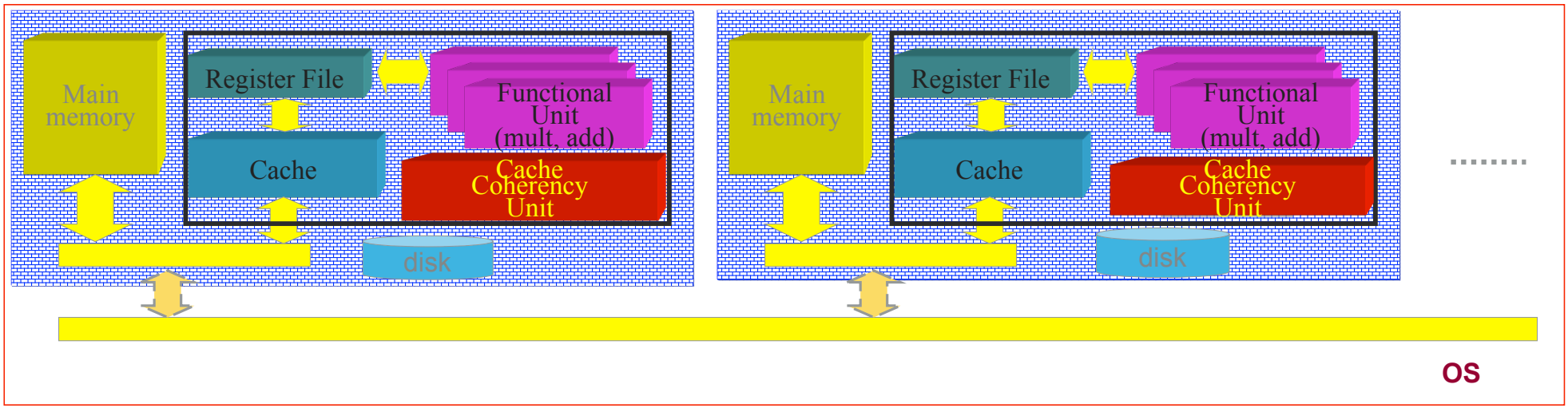


Need Data (‘message’) transfert(‘passing’)



NUMA Architecture : Shared memory, physically distributed

- Non-uniform memory access (NUMA)
  - One (big) space address (shared memory) **but Physically Distributed**
  - Access to local and « distant » memory
  - Local Access Faster than « distant » access
  - Programing in shared or distributed memory



Interconnection : key of parallel machine

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

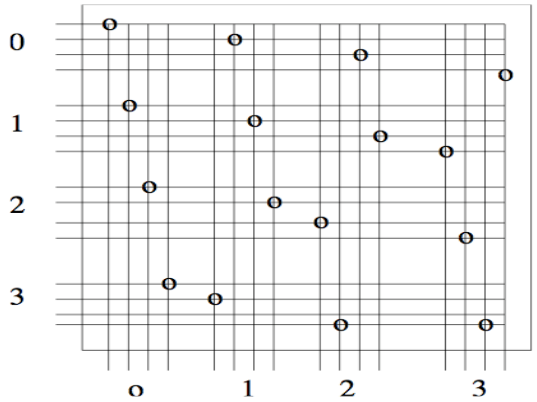
LATENCY		
time	From	to
1 nano-second	Computing units (in the core)	register
100 nano-second	core	RAM
100 Micro-second	Computer	computer
Milli-second	Core	Disk

Bandwidth		
1 GB/s	10^9 bytes/sec	Fast interconnect/
100 GB/s	10^{11} bytes/sec	RAM
1 TB/s	10^{12} bytes/s	Exists ?

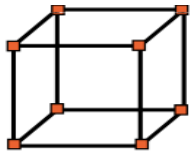
Interconnection

- Interconnection :
  - Topology :
    - Best choice:
      - Each processor to all:
        - price (affordable?), few numbers of cores
        - impossible for large scale (1000 cores…)
    - The least bad:
      - try to avoid bottleneck
      - scalability of the network topology

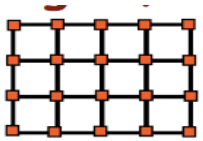
- Different strategies



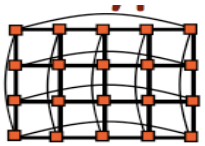
Cross-bar



3-d hypercube

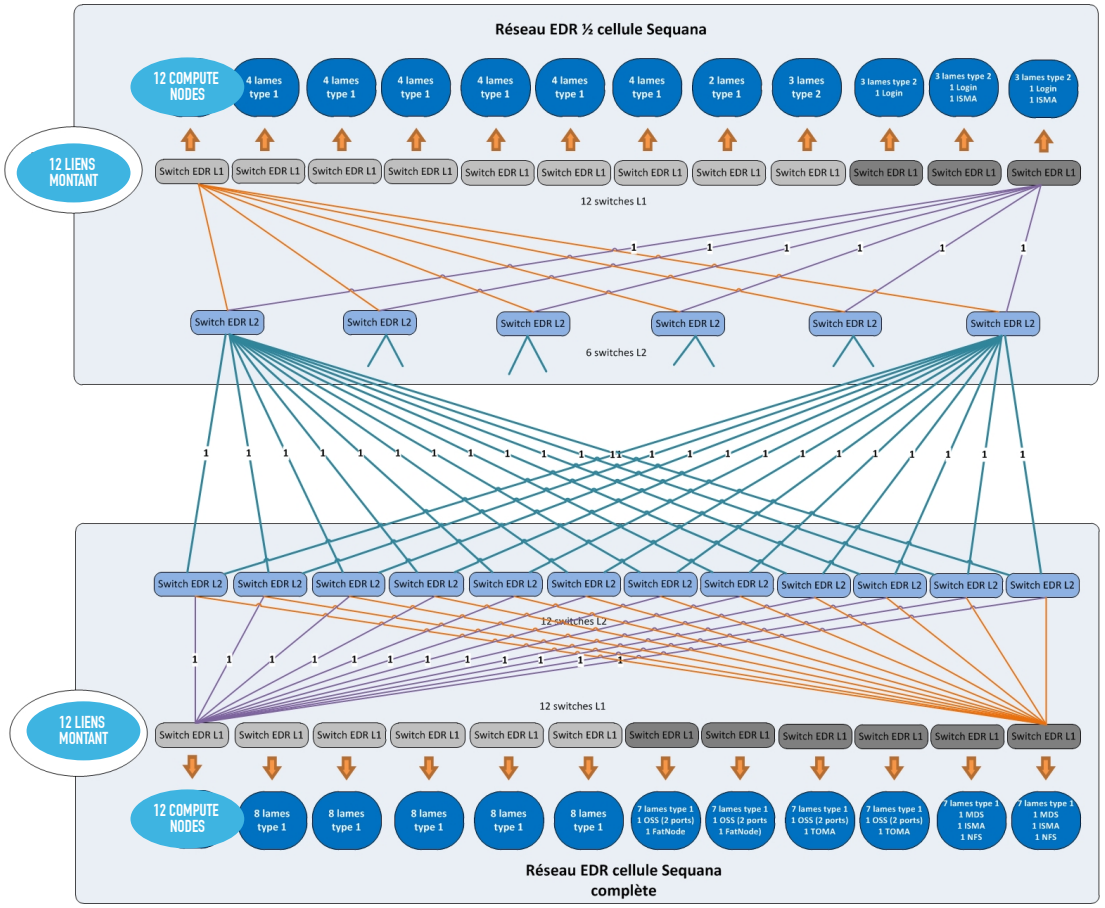


2-d mesh



2-d torus

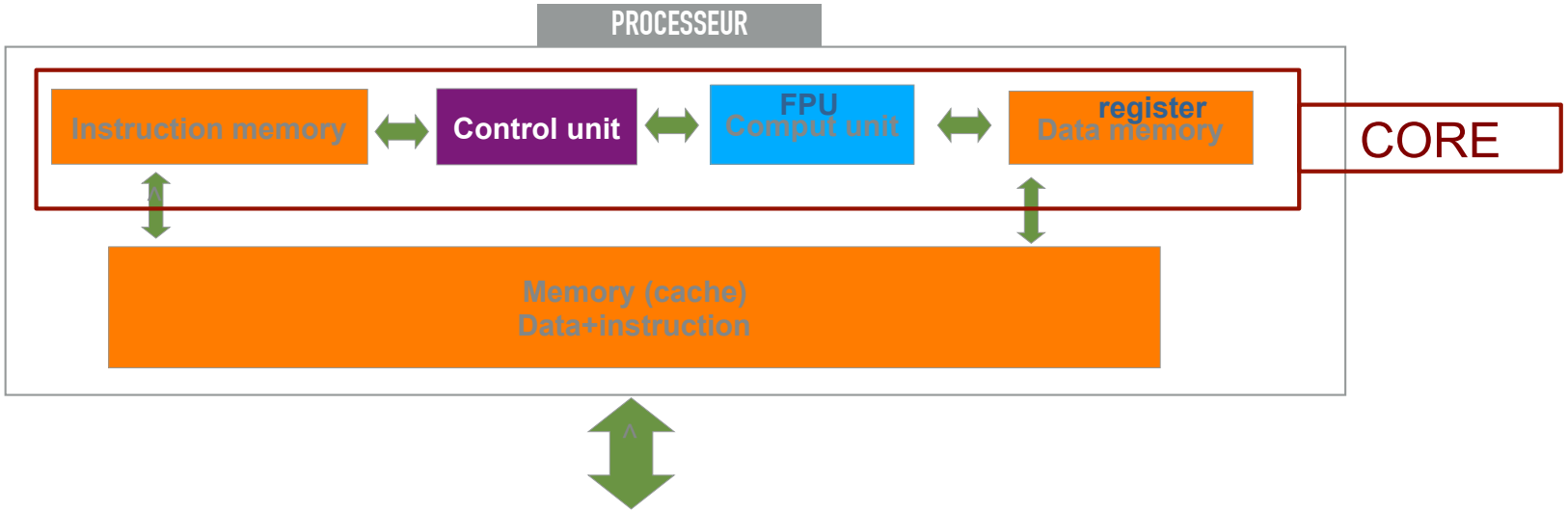
Interconnect : Detail of Olympe fat-tree Interconnect



Plan Formation :

- Introduction : Mésocentre CALMIP
- Premier jour
  - Matin : Concepts fondamentaux
    - Introduction à l'Architecture des systèmes HPC
      - *Calcul Intensif et Panorama des Systèmes*
      - ***Architecture Processeurs***
      - *Présentation système de Calcul CALMIP : EOS*
      - *Visite salle Machine*
    - Introduction programmation sur les systèmes HPC
      - *Programmation Parallèle*
      - *Optimisation de codes*

ARCHITECTURE PROCESSEUR



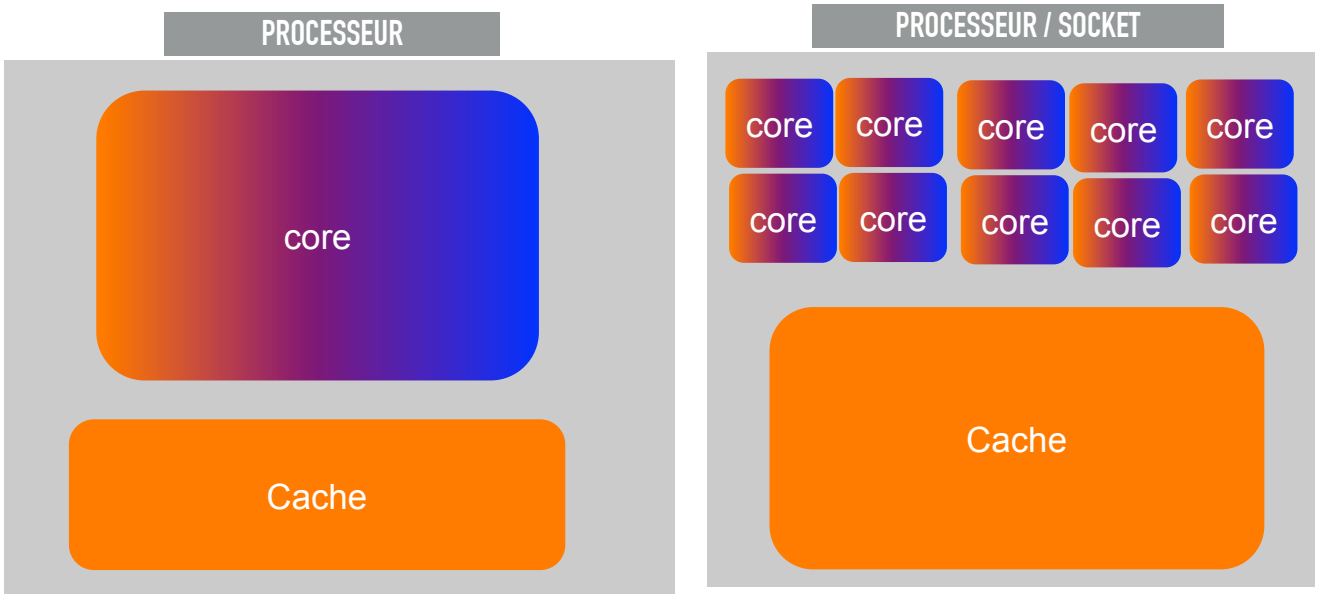
➢ processor cycle

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

  A = B + C :=> 3 phases : Load/Exec/Store :=> 3 cycles

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

ARCHITECTURE PROCESSEUR : MULTI-CORE



Gravure 180 nm
ITANIUM - 2004
6 Gflop/s

X40

Gravure 22 nm
IVYBRIDGE 2012
220 Gflop/s

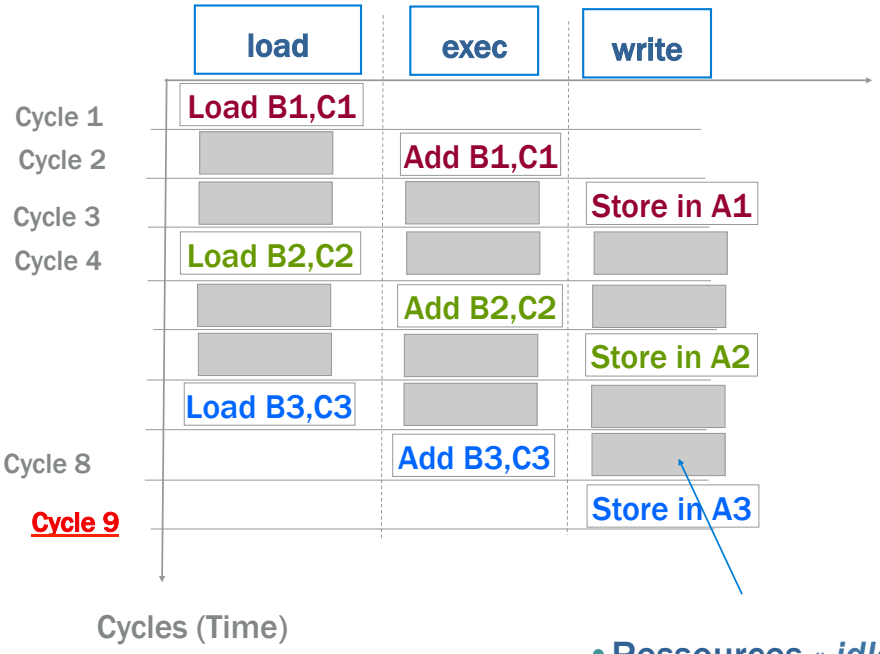
GRAVER plus fin :=> plus de transistor :=> plus de core :=> plus de capacité de calcul

calmip

➢Smart core ? 1 cycle = 1 result ?

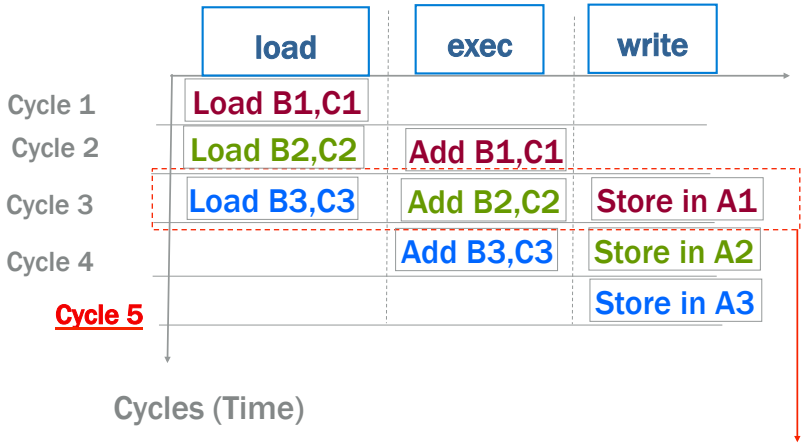
- Independent set of instructions

A1 = B1 + C1
A2 = B2 + C2
A3 = B3 + C3



- Ressources « *idle* »

after 9 cycles, 3 results



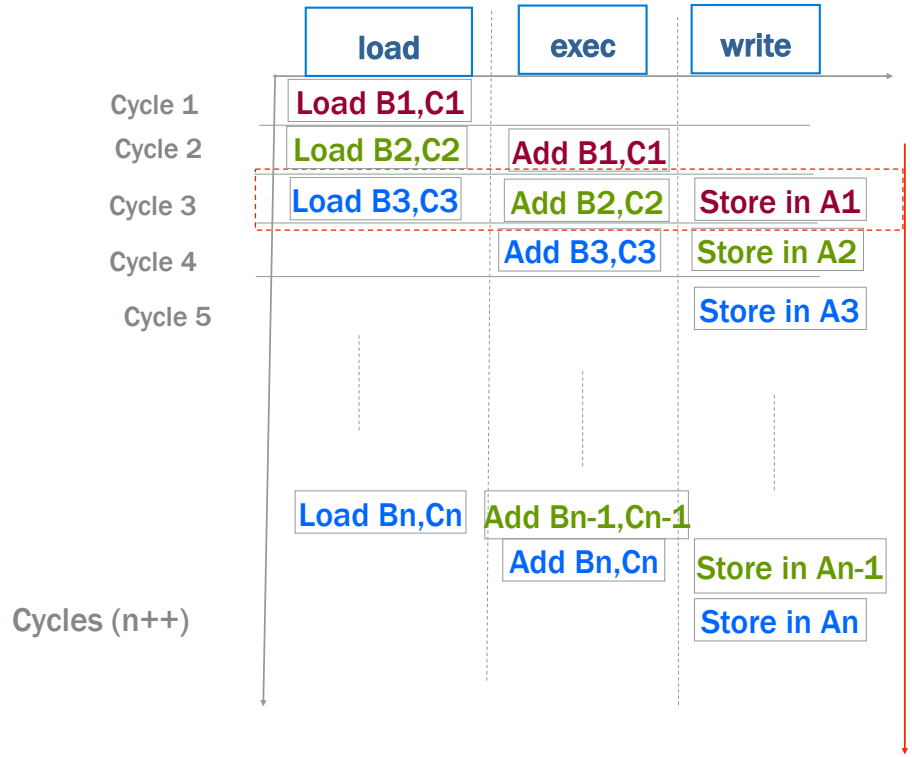
After 5 cycles, 3 results

➢Smart core ? 1 cycle = 1 result

- n independent instructions ($n=10^6$)

A1 = B1 + C1
A2 = B2 + C2
A3 = B3 + C3

An-1 = Bn-1 + Cn-1
An = Bn + Cn



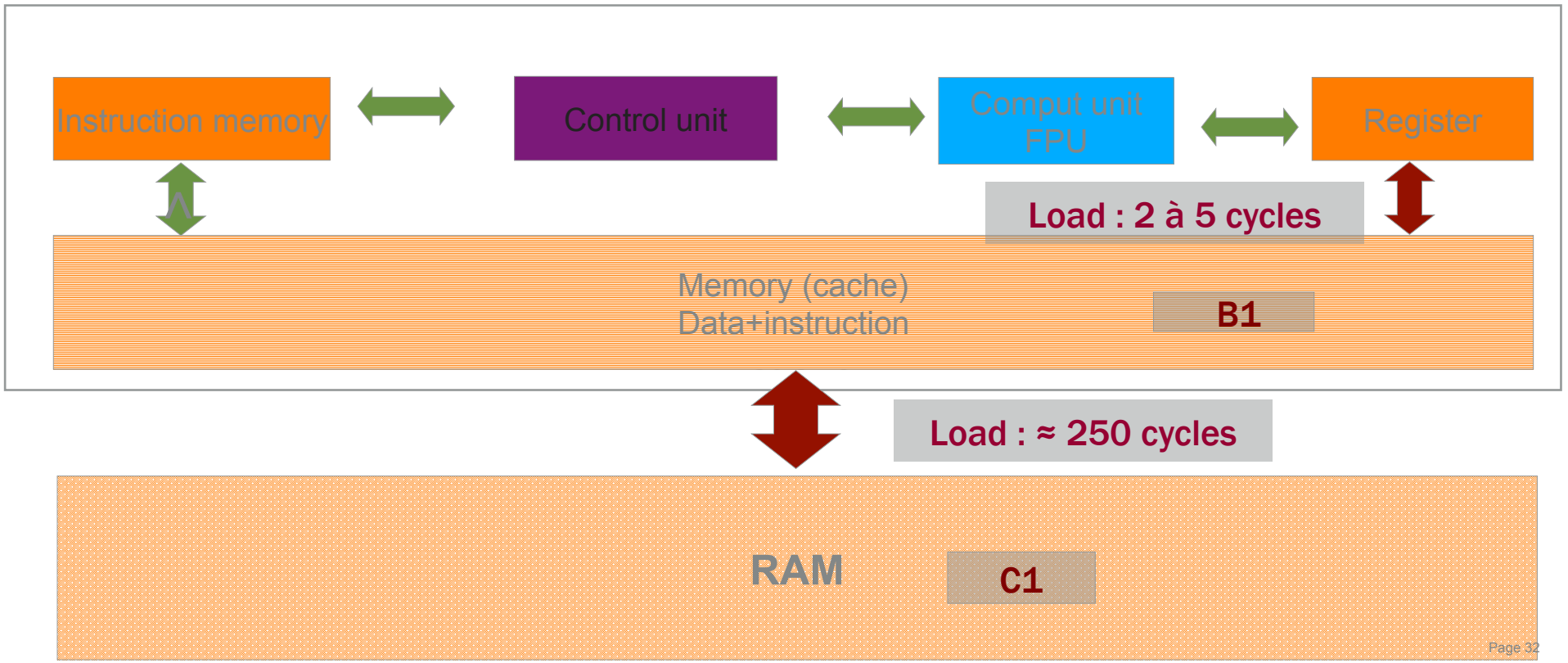
➢Exhibit independent instructions ⇒ Feed Pipeline

PROCESSOR ARCHITECTURE HPC

Hierarchical Memory

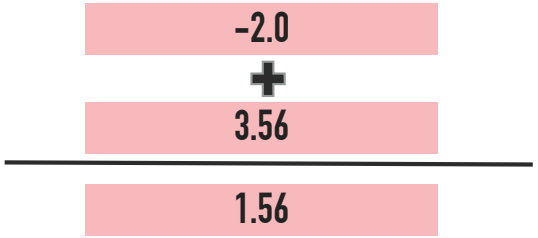
1,5 Ghz => 1 cycle = 0,6 ns

A1 = B1 + C1 ⟶ load | exec | write



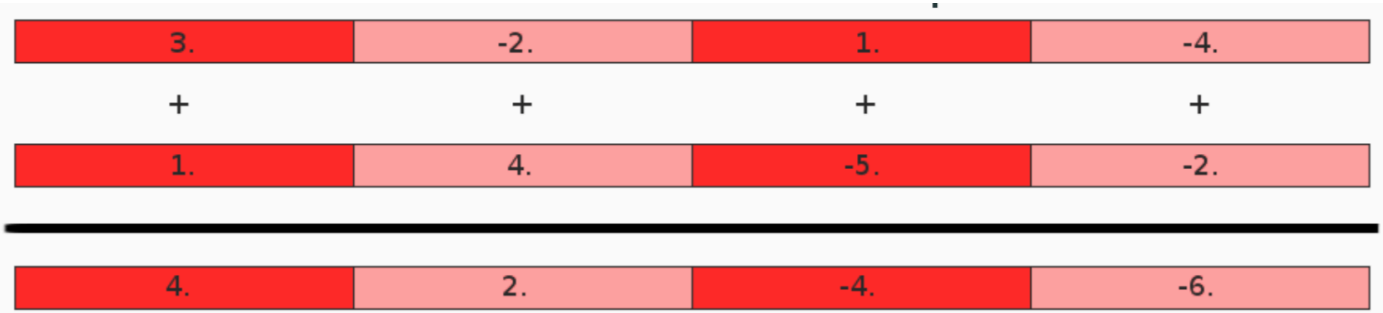
PROCESSOR ARCHITECTURE : SCALAR VS. VECTOR (SIMD)

Core scalar



1 CPU cycle

Core vector



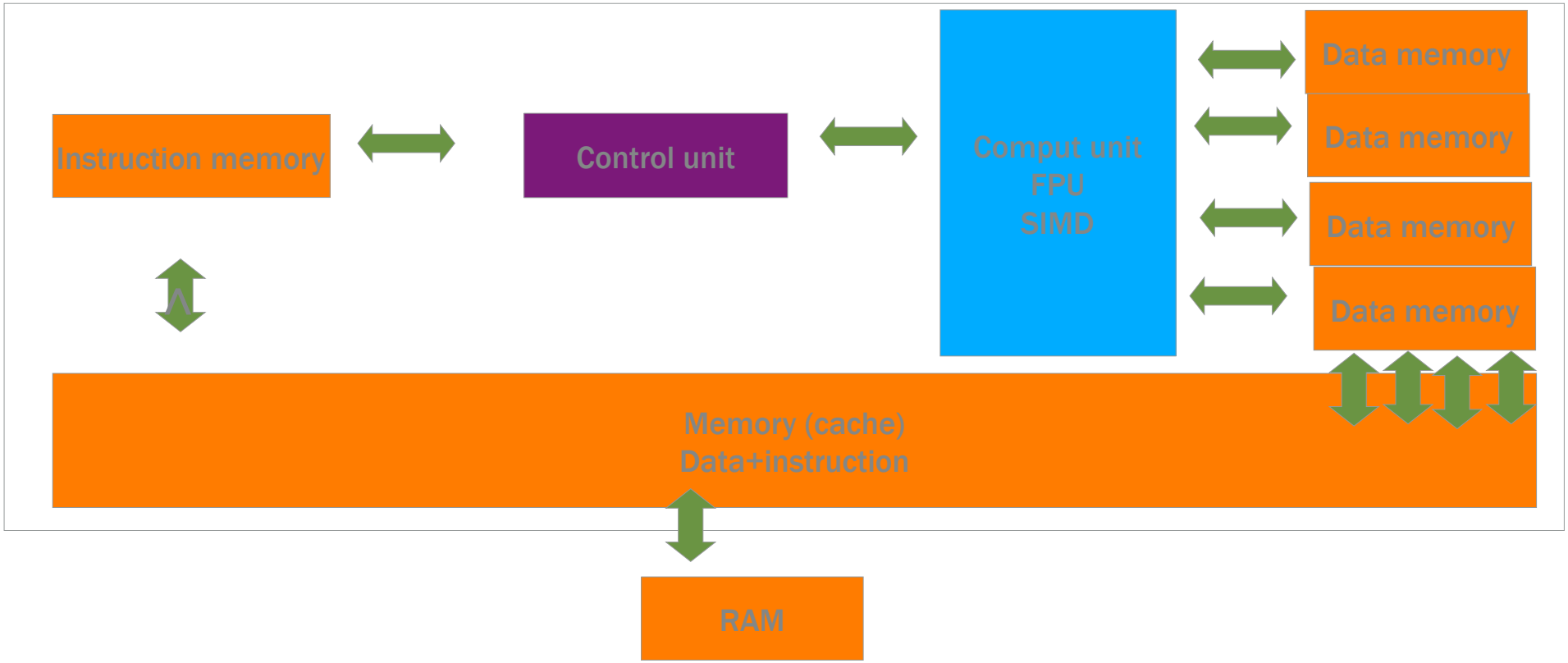
1 CPU cycle

With vector instruction : Flop/s x 4 vs. Scalar instruction

S.I.M.D. : Single Instruction Multiple Data

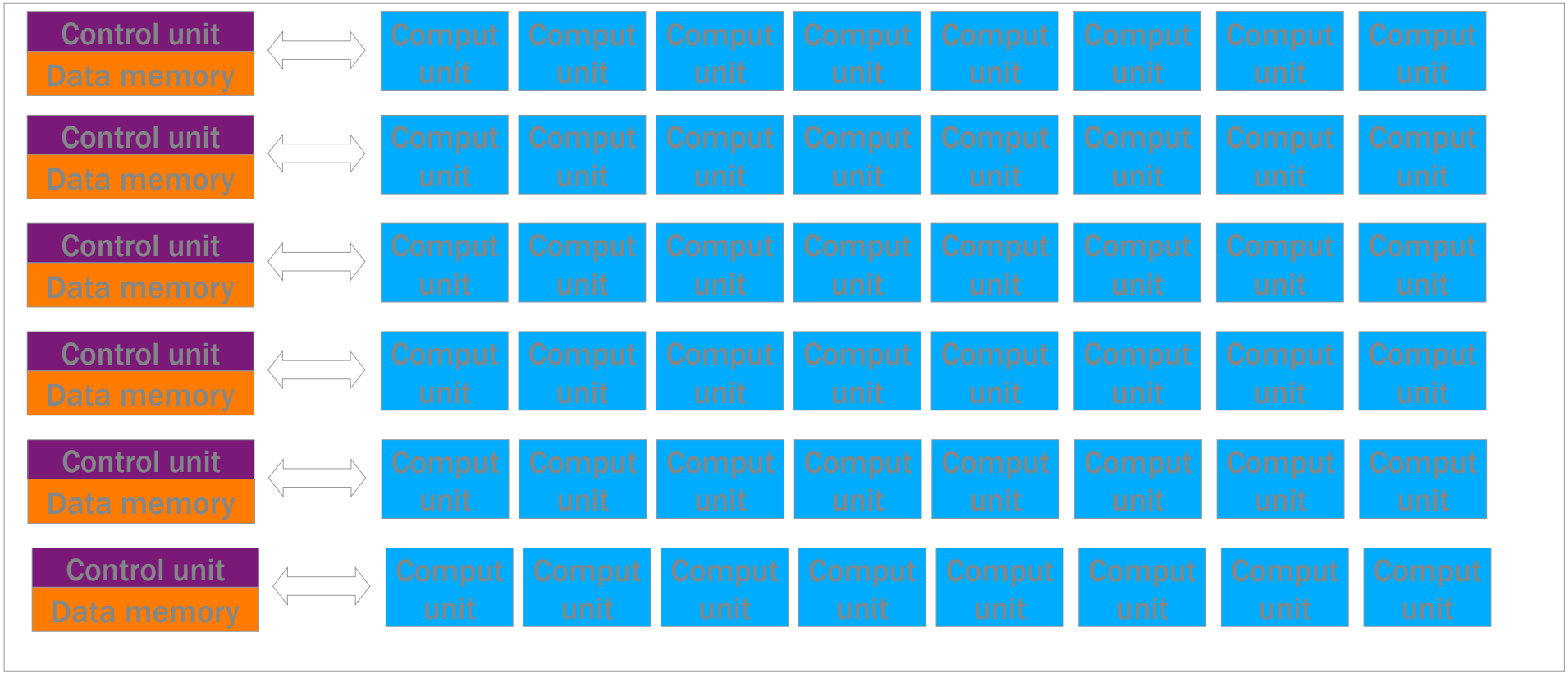
PROCESSORS ARCHITECTURE : SIMD

➢ Processor Architecture

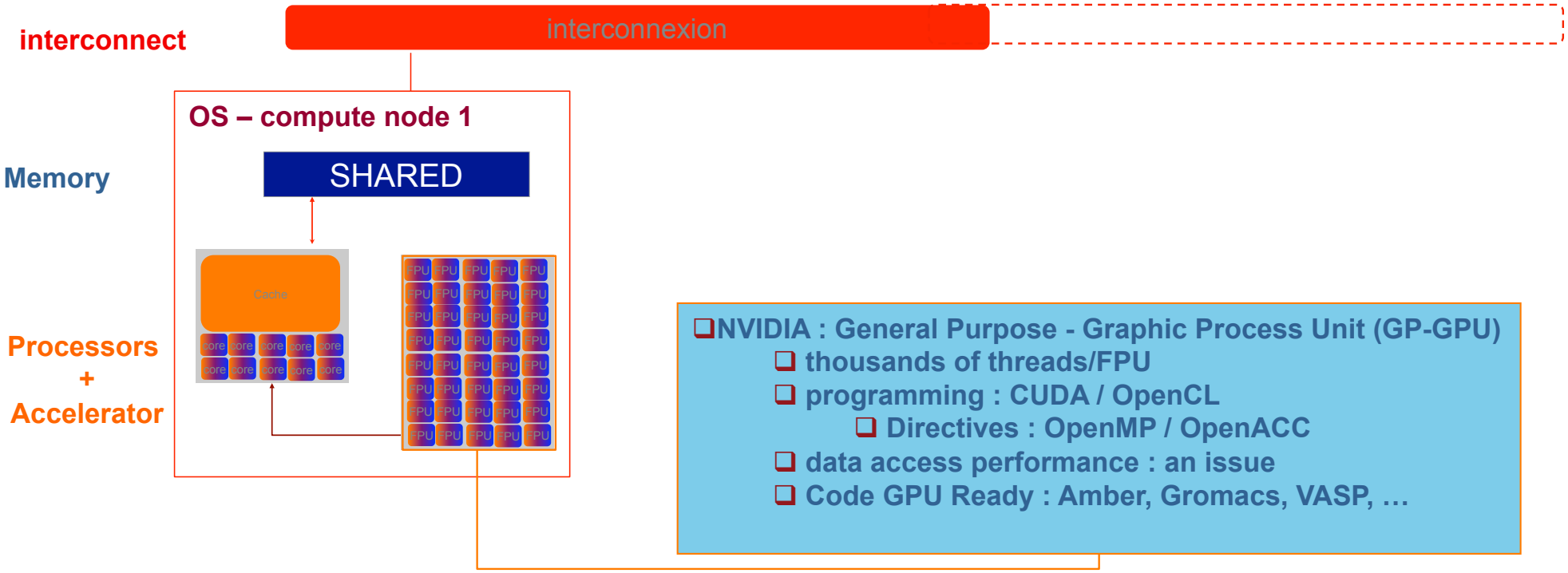


Accelerators

➢ Accelerator Architecture : GP-GPU



Accelerator GP-GPU



- NVIDIA : General Purpose - Graphic Process Unit (GP-GPU)
  - thousands of threads/FPU
  - programming : CUDA / OpenCL
    - Directives : OpenMP / OpenACC
  - data access performance : an issue
  - Code GPU Ready : Amber, Gromacs, VASP, …

Top500 Juin 2018

Rank	Site	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	DOE/SC/Oak Ridge National Laboratory United States	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband IBM	2,282,544	122,300.0	187,659.3	8,806
2	National Supercomputing Center in Wuxi China	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway NRCPC	10,649,600	93,014.6	125,435.9	15,371
3	DOE/NNSA/LLNL United States	**Sierra** - IBM Power System S922LC, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband IBM	1,572,480	71,610.0	119,193.6	
4	National Super Computer Center in Guangzhou China	**Tianhe-2A** - TH-IVB-FEP Cluster, Intel Xeon E5-2692v2 12C 2.2GHz, TH Express-2, Matrix-2000 NUDT	4,981,760	61,444.5	100,678.7	18,482
5	National Institute of Advanced Industrial Science and Technology (AIST) Japan	**AI Bridging Cloud Infrastructure (ABCI)** - PRIMERGY CX2550 M4, Xeon Gold 6148 20C 2.4GHz, NVIDIA Tesla V100 SXM2, Infiniband EDR Fujitsu	391,680	19,880.0	32,576.6	1,649
6	Swiss National Supercomputing Centre (CSCS) Switzerland	**Piz Daint** - Cray XC50, Xeon E5-2690v3 12C 2.6GHz, Aries interconnect , NVIDIA Tesla P100 Cray Inc.	361,760	19,590.0	25,326.3	2,272

Accélerateurs GPU

8 MW

15 MW

Plan Formation :

- Introduction : Mésocentre CALMIP
- Premier jour
  - Matin : Concepts fondamentaux
    - Introduction à l'Architecture des systèmes HPC
      - *Calcul Intensif et Panorama des Systèmes*
      - *Architecture Processeurs/ Accélérateurs*
      - ***Présentation système de Calcul CALMIP OLYMPE(+Film) + Visite salle Machine***
    - Introduction programmation sur les systèmes HPC
      - *Programmation Parallèle*
      - *Optimisation de codes*