Hybrid Parallelism on “real” applications and simulations

Jean-François Méhaut

Université Joseph Fourier/LIG/INRIA/Mescal

With A. Carissimi, L. Fernandes, T. Deutsch, L. Genovese, M. Ospici

M. Castro, C. Pousa, F. Dupros, H. Aochi, D. Komatitsch

Outline of Presentation

- Introduction
  - General description of « real » applications
- Simulations for nanosciences
  - Ab initio methods
  - BigDFT code
  - Hybrid clusters
- Seismic simulations
  - Wave propagation
  - Memory Affinity

Introduction

Introduction (1)

- « Real » applications
  - Societal and industrial impact
    - Geosciences
    - Nanotechnology
    - ...
- Multidisciplinarity
  - Physics, Chemistry, Geophysics, ...
  - Applied Mathematics
  - Computer Science
- Regional, National and International Projects

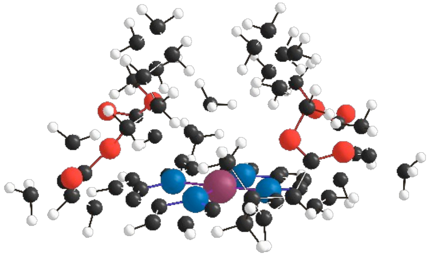
Introduction (2)

- « Real » Applications
  - Old Fashion Software
    - Algorithms
    - Programming Languages (Fortran,...)
  - Intensive computations
  - Huge volume of data
    - Input : storing data in memory
    - Output : Analysis, data mining, visualization...
- Need to anticipate for the using of new and future computing platforms
  - Multicore, GPU, accelerator
  - Memory hierarchy
  - Minimizing the software intrusion!

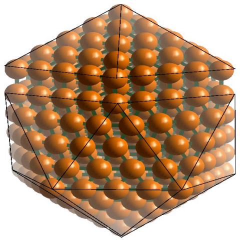
Simulations for nanosciences and nanotechnologies

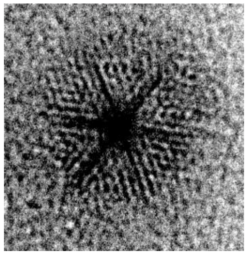
Basic components in nanosciences

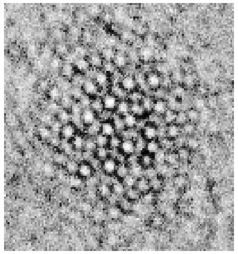
- **Molecules :**



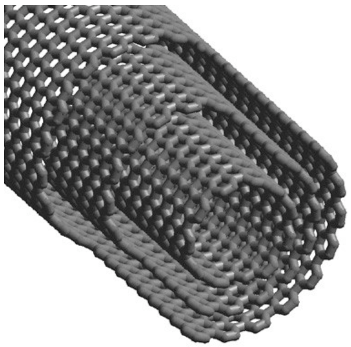
- **Nanocrystal :**

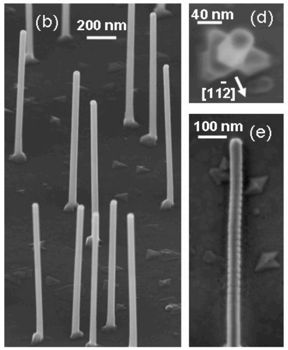






- **Nanotubes of carbon, nanowires**

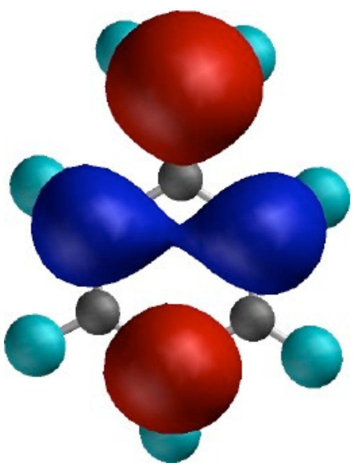




ab initio methods

- Density Functional Theory (DFT)
  - Computing the total energy of a solid or a molecule in its fundamental state (Hohenber & Kohn 1964)
  - Soliving the Schrodinger equation

$$-\frac{1}{2m}\nabla^2\psi(\mathbf{r}) + V_{\text{ions}}(\mathbf{r})\psi(\mathbf{r}) + V_{\text{hxc}}[\psi](\mathbf{r})\psi(\mathbf{r}) = \varepsilon\psi(\mathbf{r})$$



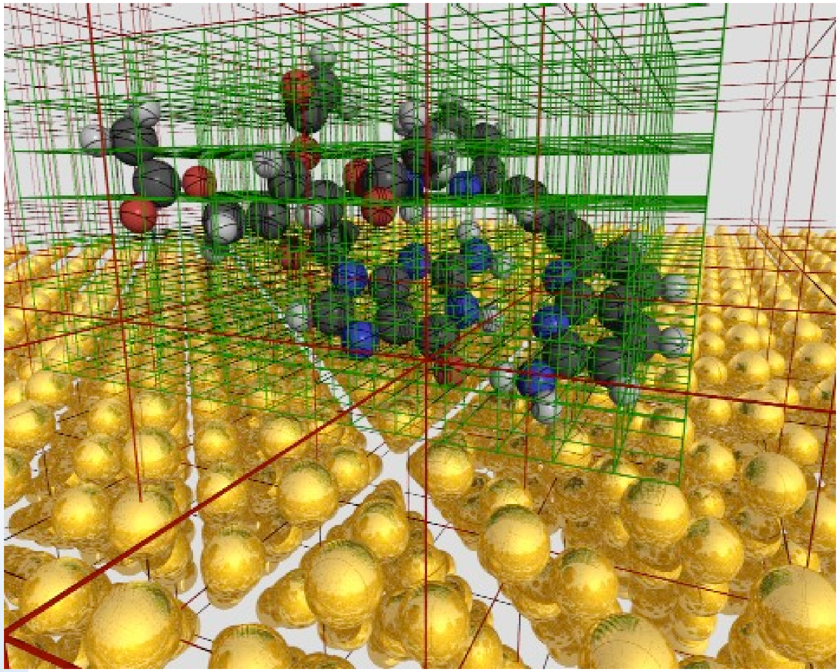
For each electron of system, a wave function ψ Is applied

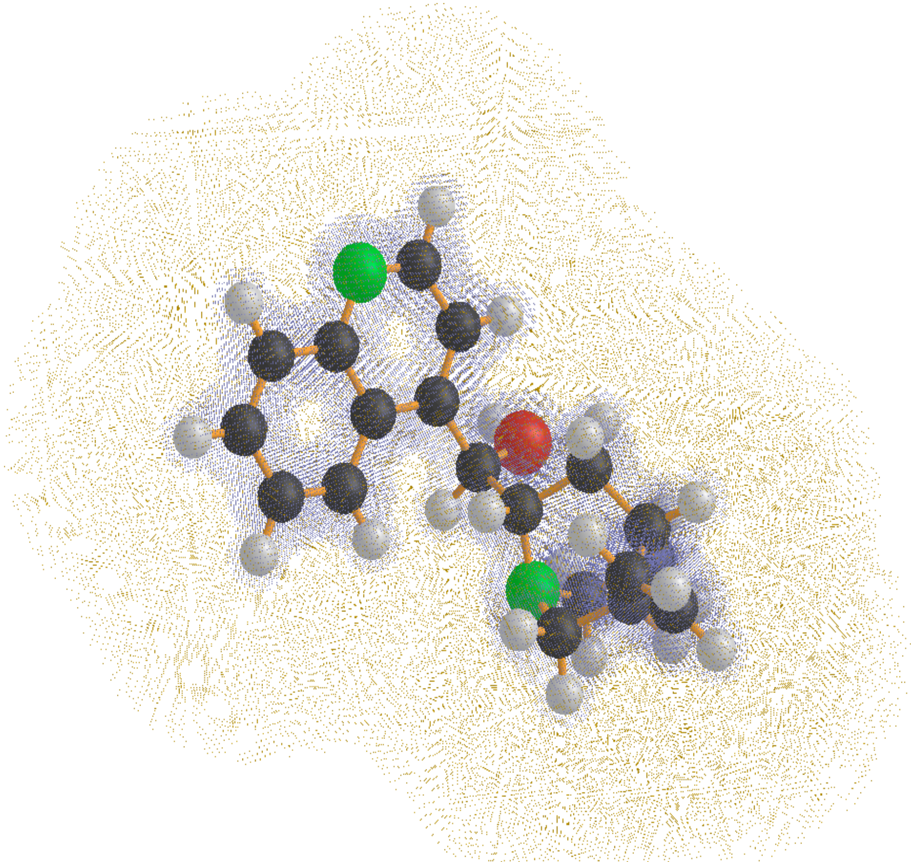
Intensive computations :

- From 10 to 1 000 atoms
- 40 à 10000 wave functions

Adaptive mesh

Molecule with 44 atoms and its adaptive mesh





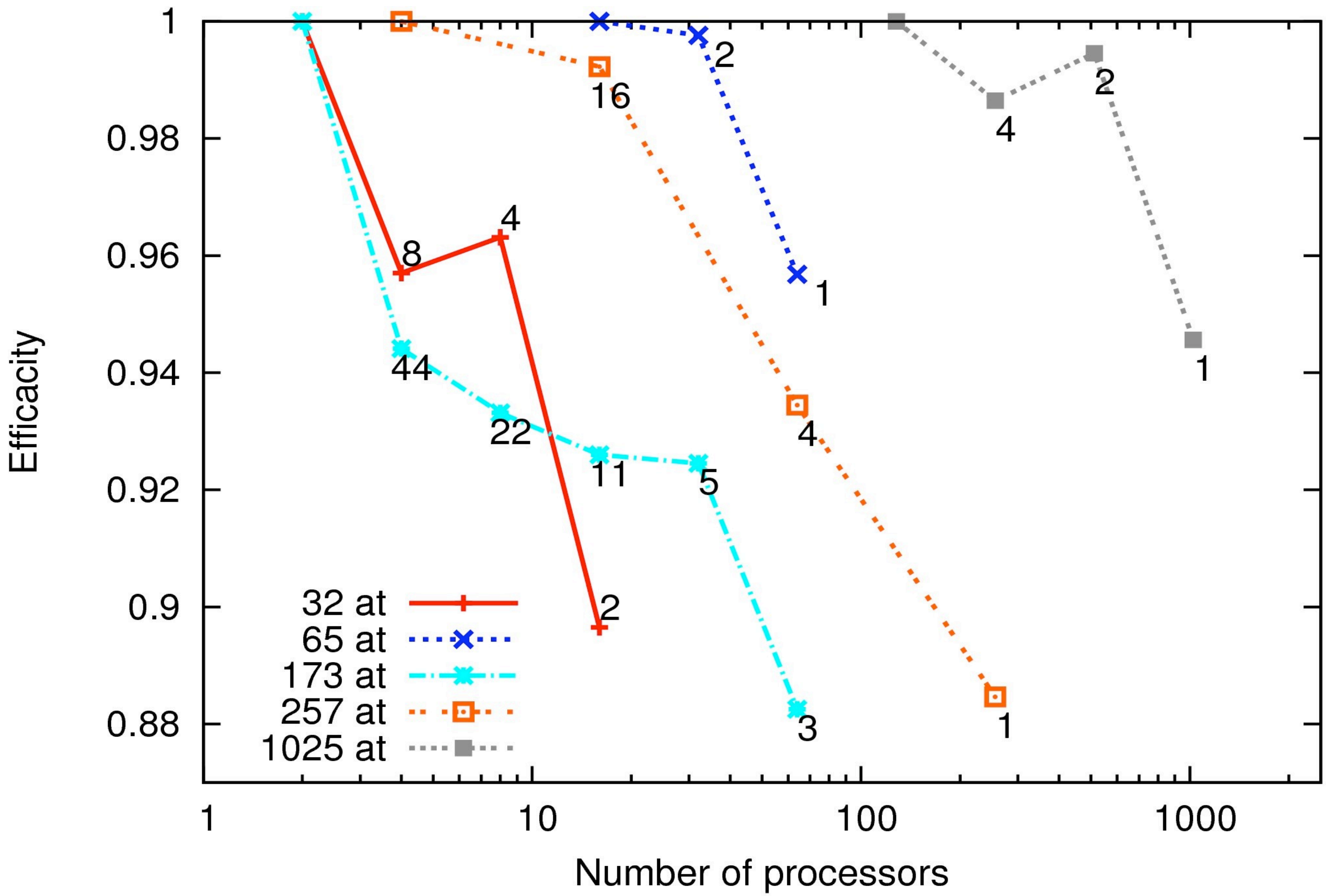
Automaticaly adjust the wave function depending of the properties to study

A basis for nanosciences
The BigDFT Project

- STREP European Project: BigDFT (2005-2008)
  - 4 partners, 20 contributors
    - CEA-INAC Grenoble, U. Basel, U. Louvain-la-Neuve, U. Kiel
- Aim: to develop an ab inito code DFT based on Daubechies Wavelets, to be integrated in AB INIT distributed freely to the scientific community
  - http://www.abinit.org
- Result:
  - The BigDFT code is fully operational, stable and robust with excellent performance

Already in use with several applications

BigDFT performance on homogeneous clusters



BigDFT and hybrid clusters

- Hybrid clusters (CINES, GENCI-CEA)
  - Nodes with Intel processors (2 x 4 cores) and GPU (Nvidia Tesla S1070)
  - Infiniband network
- BigDFT code
  - Fortran code, MPI
  - CUDA
  - S_GPU : Sharing GPUs between many CPU cores
    - Overlap Memory Transferts
  - Wavelet : convolution products executed on GPU
- Preliminar results
  - Around 7 faster with GPUs and Intel Xeon (CINES)
  - Around 4 faster with GPUs and Intel Nehalem (GENCI-CEA)

Financé par ANR

ProHMPT project

- BuLL: French Computer Manufacturer
  - Hybrid cluster
- CAPS Entreprise:
  - Hybrid Multicore Parallel Programming
- CEA: Atomic Reseach Institute
  - INAC (Nanosciences institute)
  - DAM/CESTA (Military division)
- UVSQ: University of Versailles
- INRIA: Computer Science Research
  - Runtime-Bordeaux, Mescal-Grenoble

Seismic Simulations

Goals

- Simulate earthquakes on clusters of NUMA and multi-core nodes
- Improve SeISmic models, define new case tests
- Increase the performance of systems and runtime of NUMA machines (multithreading, memory affinity)
- Goal: A better understanding of earthquakes and increasing of simulation power

- NUMASIS: A multidisciplinary project
  - Geophysician (models of wave propagation)
  - Computer Science (operating system and runtimes, numerical algorithmic))

Financé par ANR

Consortium

NUMASIS

- BuLL: French Computer Manufacturer
  - NUMA Architecture: Novascale
- BRGM: Geoscience Institute
  - Sismic application (ONDES3D)
- CEA: Atomic Reseach Institute
- TOTAL: Oil and gas exploration
  - Knowledge of geological layers
- INRIA: Computer Science Research
  - Magique3D, ScAlApplix, Runtime, Mescal, Paris

Earthquake Hazard Assessment

Use parallel computing to simulate earthquakes

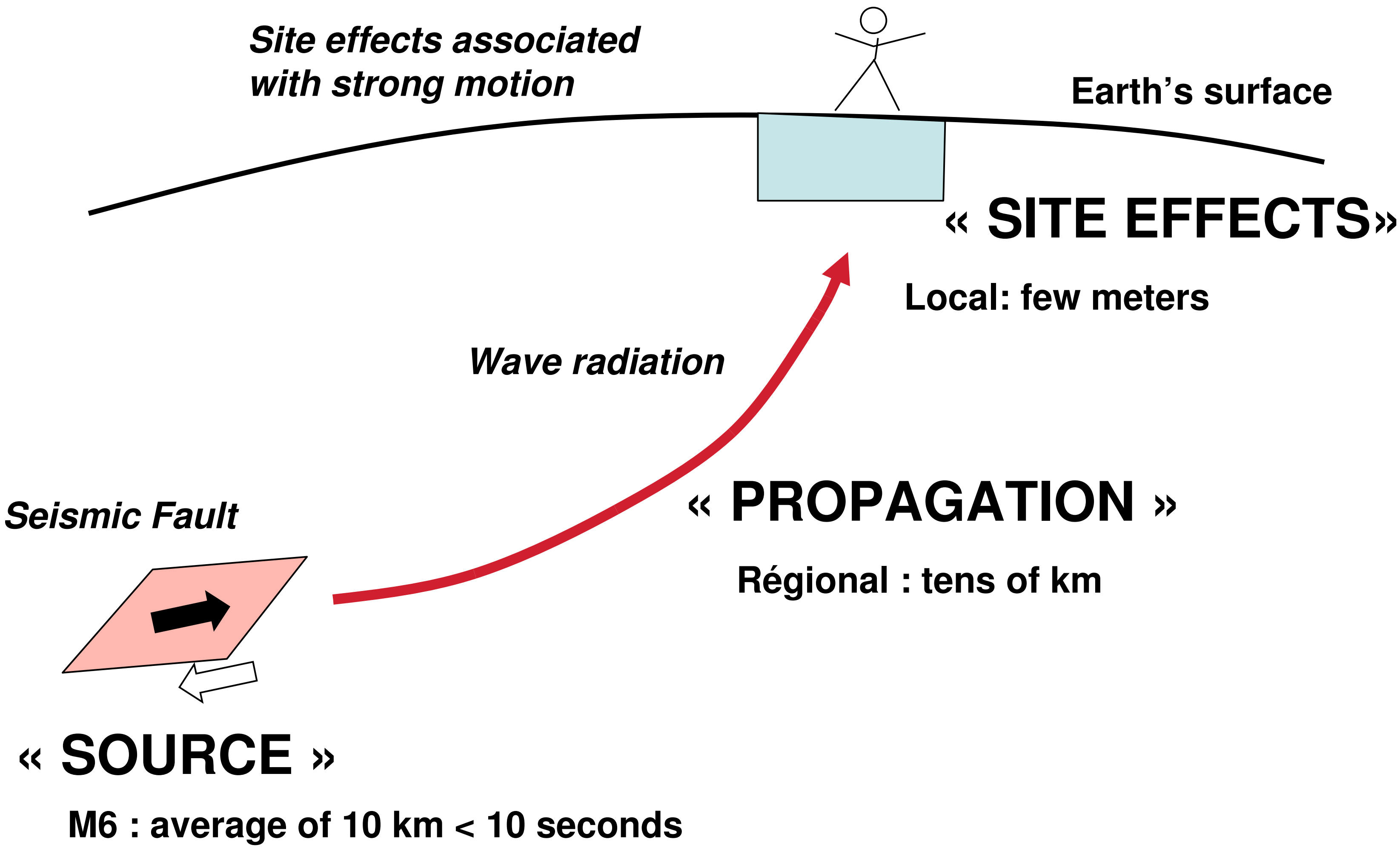
Learn about structure of the Earth based upon seismic waves (tomography)

Produce seismic hazard maps (local/regional scale) e.g. Los Angeles, Tokyo, Nice, Grenoble

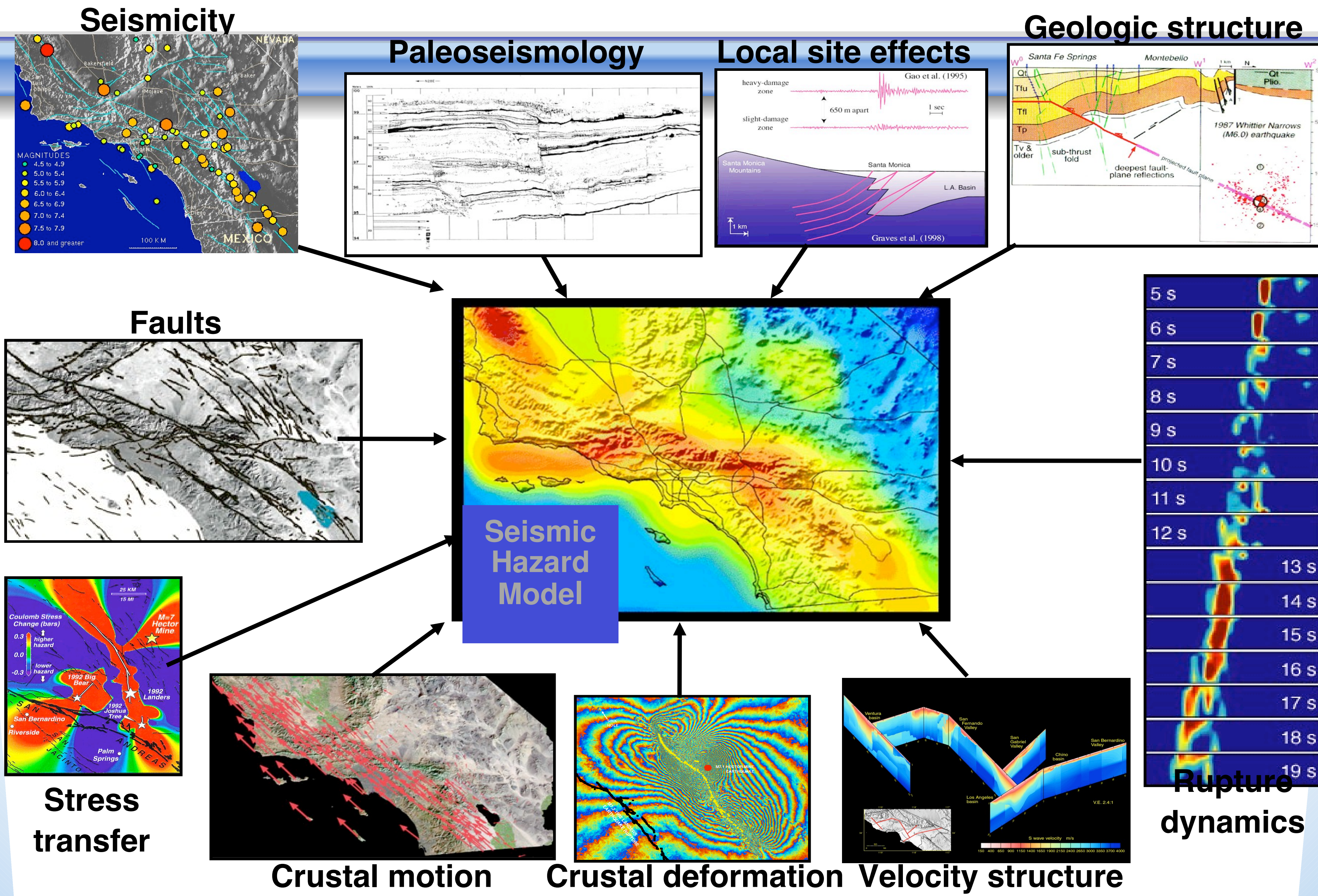
2001 Gujarati (M 7.7) Earthquake, India

20,000 people killed
167,000 injured
≈ 339,000 buildings destroyed
783,000 buildings damaged

Different scales for the simulation and observation of seismic phenomena



Ingredients for a Seismic Hazard Calculation



NUMASIS applications

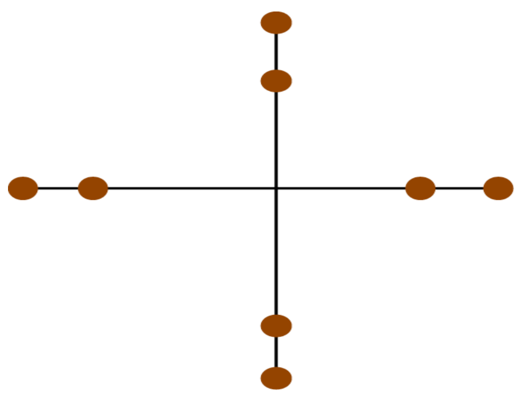
- ONDES3D (BRGM, Ecole Centrale de Paris)
  - OpenMP
  - MPI
  - Hybrid : OpenMP + MPI

- PRODIF (CEA, TOTAL)

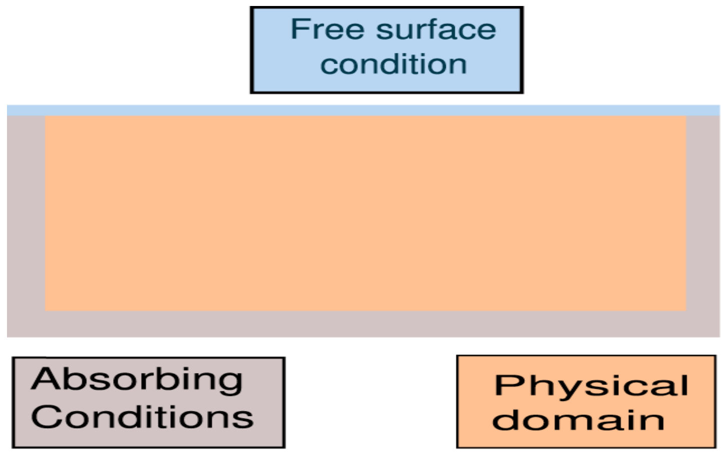
- SPECFEM3D (Magique3D, Caltech)

  http://www.gps.caltech.edu/~jtromp/research/downloads.html

ONDES3D: Numerical Modeling



Stencil computation



Initialization

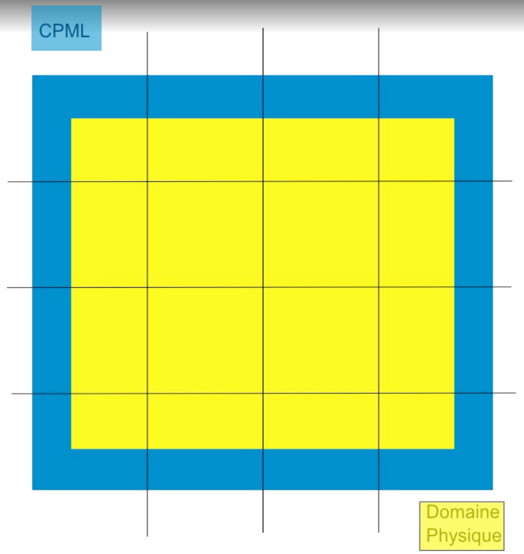
```
For i=1,Nx
 For j=1,Ny
  For k=1,Nz
   init()
```

Time step loop

```
For i=1,Nx
 For j=1,Ny
  For k=1,Nz
   compute_velocity()
```

```
For i=1,Nx
 For j=1,Ny
  For k=1,Nz
   compute_stress()
```

ONDES3D: Memory mapping and multithreading scheduling



Exploit first-touch linux strategy with OpenMP ?
Strong link between initialization (allocation) phase and execution phase. Replay the same mapping of threads

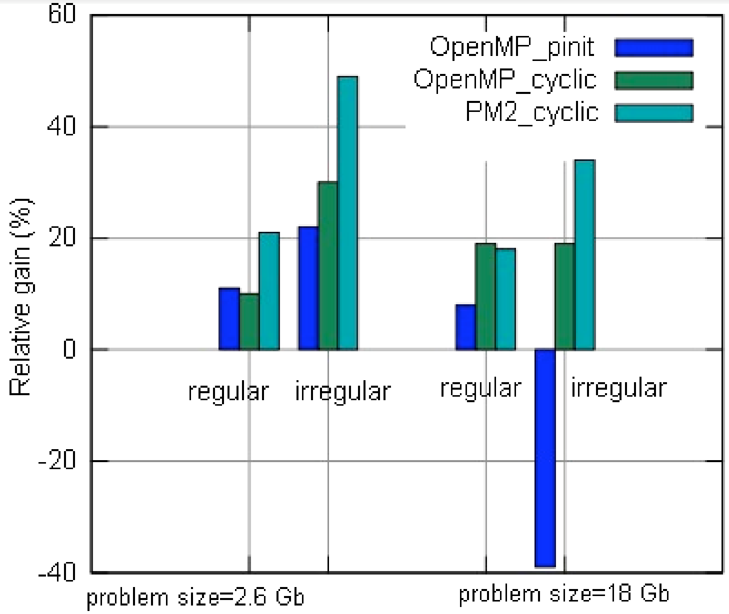
Memory migration ?
Performance depend on the dynamical behavior of the application

Flexible memory policies : MAI library (INRIA Mescal)

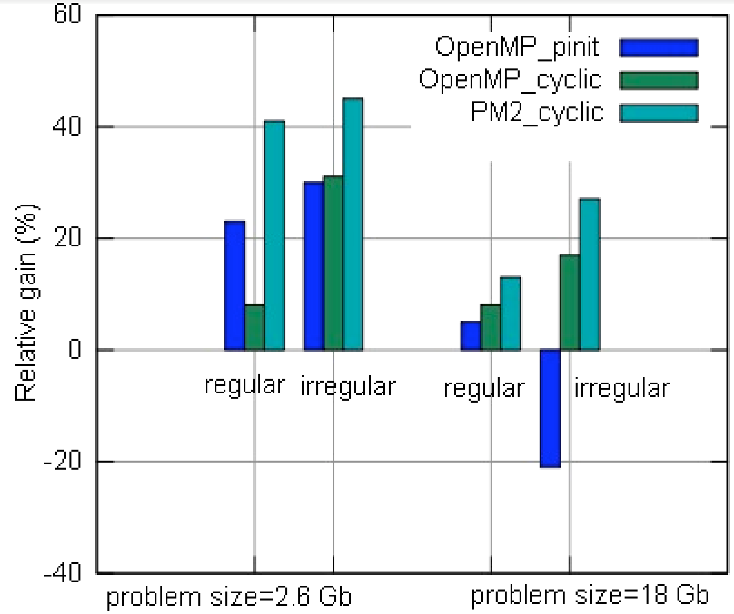
<u>Application requirement</u>

Need for dynamic thread scheduling strategy to consider load-balancing at the shared memory level

ONDES3D: Results with Marcel/PM2 and MAI



Balanced situation - physical domain



Unbalanced situation – Complete simulation

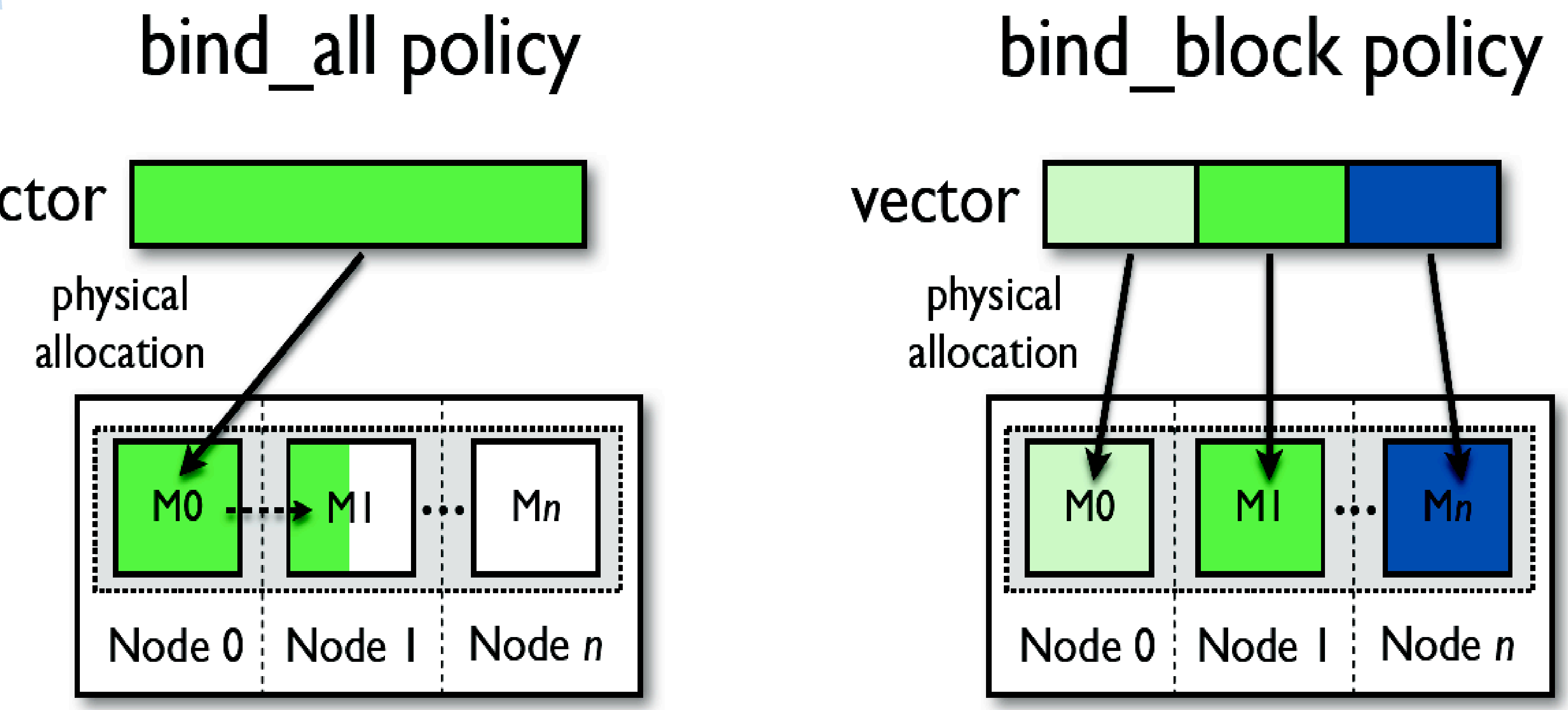
- *Gain relative to OpenMP regular and irregular memory pattern and FT memory policy*
- More 1000 threads on 16 cores for the irregular PM2 version
- *Importance of the size of the problem and the thread scheduling to choose the best memory policy on NUMA architecture.*

Memory Affinity Interface (MAi)

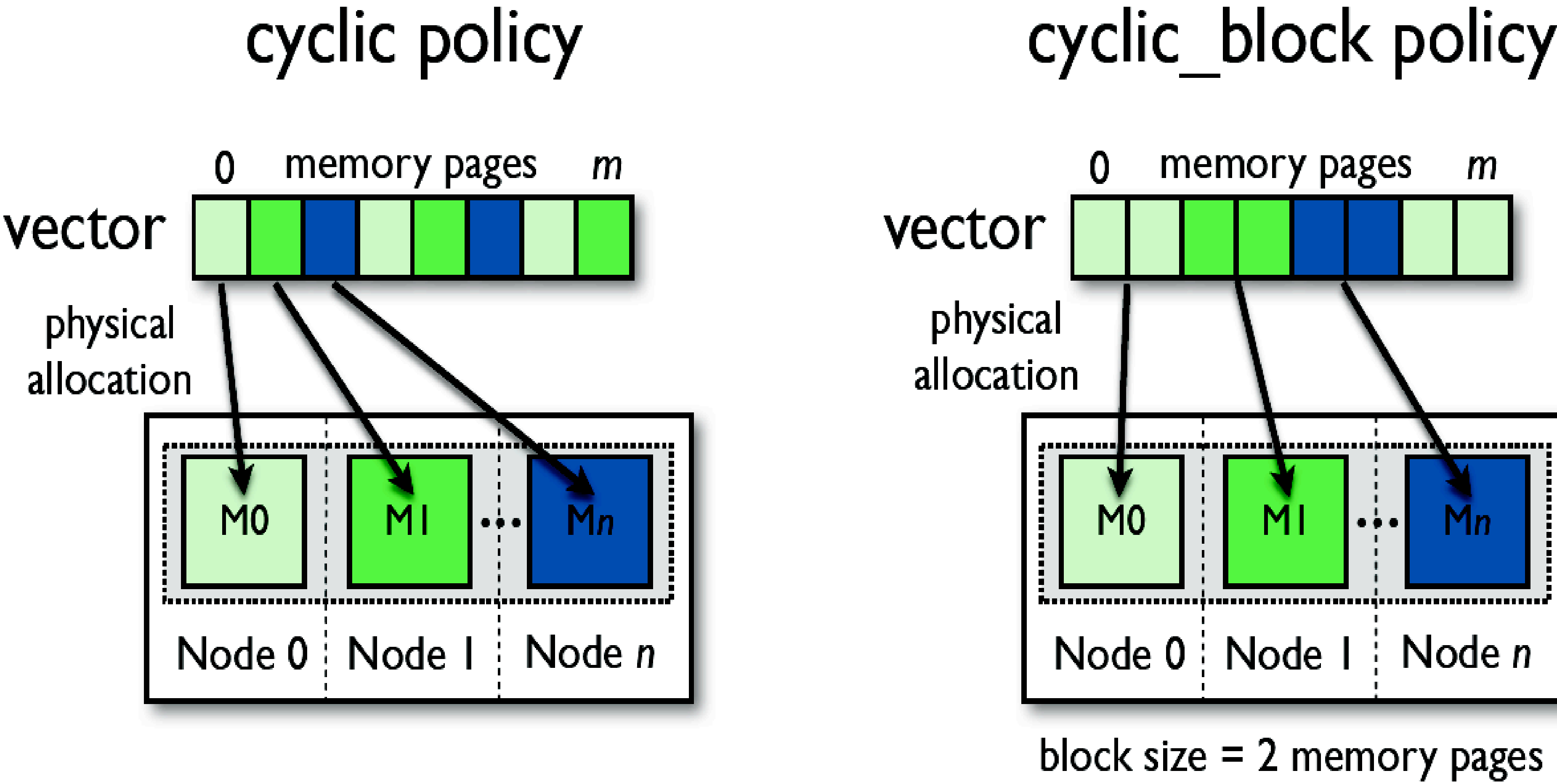
- A user-level interface to control memory affinity on NUMAs

- Memory Policies:
  - bind_all
  - bind_block
  - cyclic
  - cyclic_block
  - ...

MAi (Memory Affinity Interface)



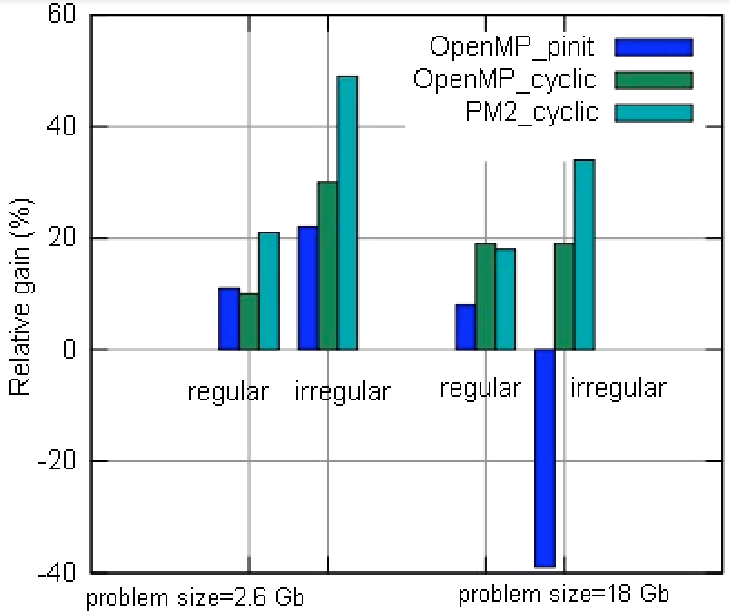
MAi (Memory Affinity Interface)



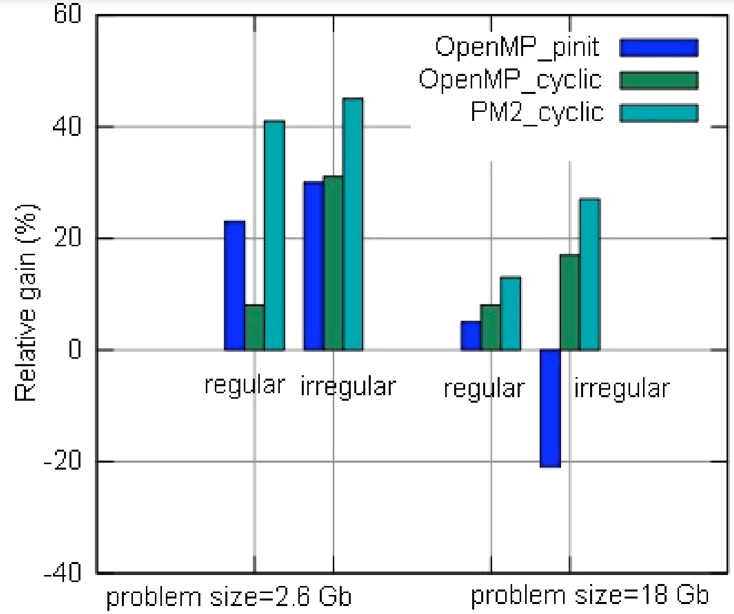
MAi (Memory Affinity Interface)

```
absolute_matrix = mai_alloc(Nx,Ny,sizeof(double));
resltave_matrix = mai_alloc(Nx,Ny,sizeof(double));

mail_bind_all(absolute_matrix);
mai_bind_all(relative_matrix);


........

#pragma omp parallel for
for(....)
  for(...)
      compute(relative_matrix);
```

ONDES3D: Results with Marcel/PM2 and MAi



Balanced situation - physical domain



Unbalanced situation – Complete simulation

- *Gain relative to OpenMP regular and irregular memory pattern and FT memory policy*
- More 1000 threads on 16 cores for the irregular PM2 version
- *Importance of the size of the problem and the thread scheduling to choose the best memory policy on NUMA architecture.*