High Performance Computational Geosciences at BRGM

Fabrice Dupros

Géosciences pour une Terre durable

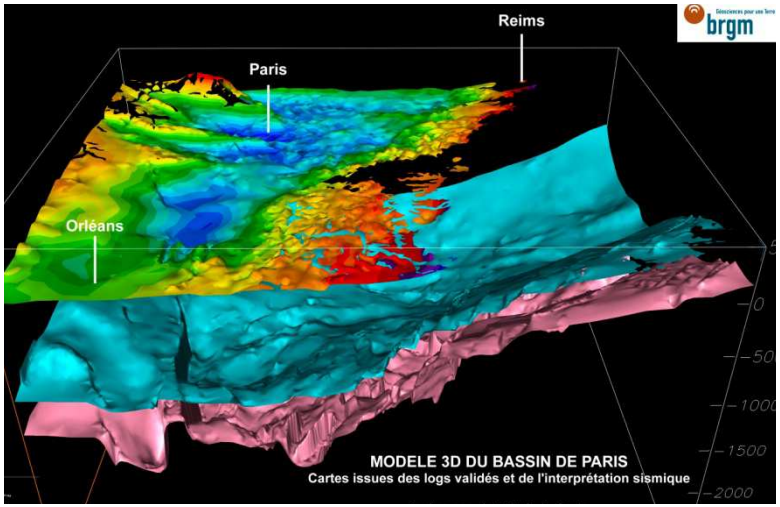
brgm

BRGM at a glance

> **The France's leading public institution in Earth Science applications for the management of resources and surface and subsurface risks**

> **Key objectives**

- **Understanding** geological processes, developing new methodologies and techniques, producing and disseminating relevant high-quality data.
- **Providing** necessary tools:
  - For surface, subsurface and resource management
  - For risk and pollution prevention
  - To support climate change policies



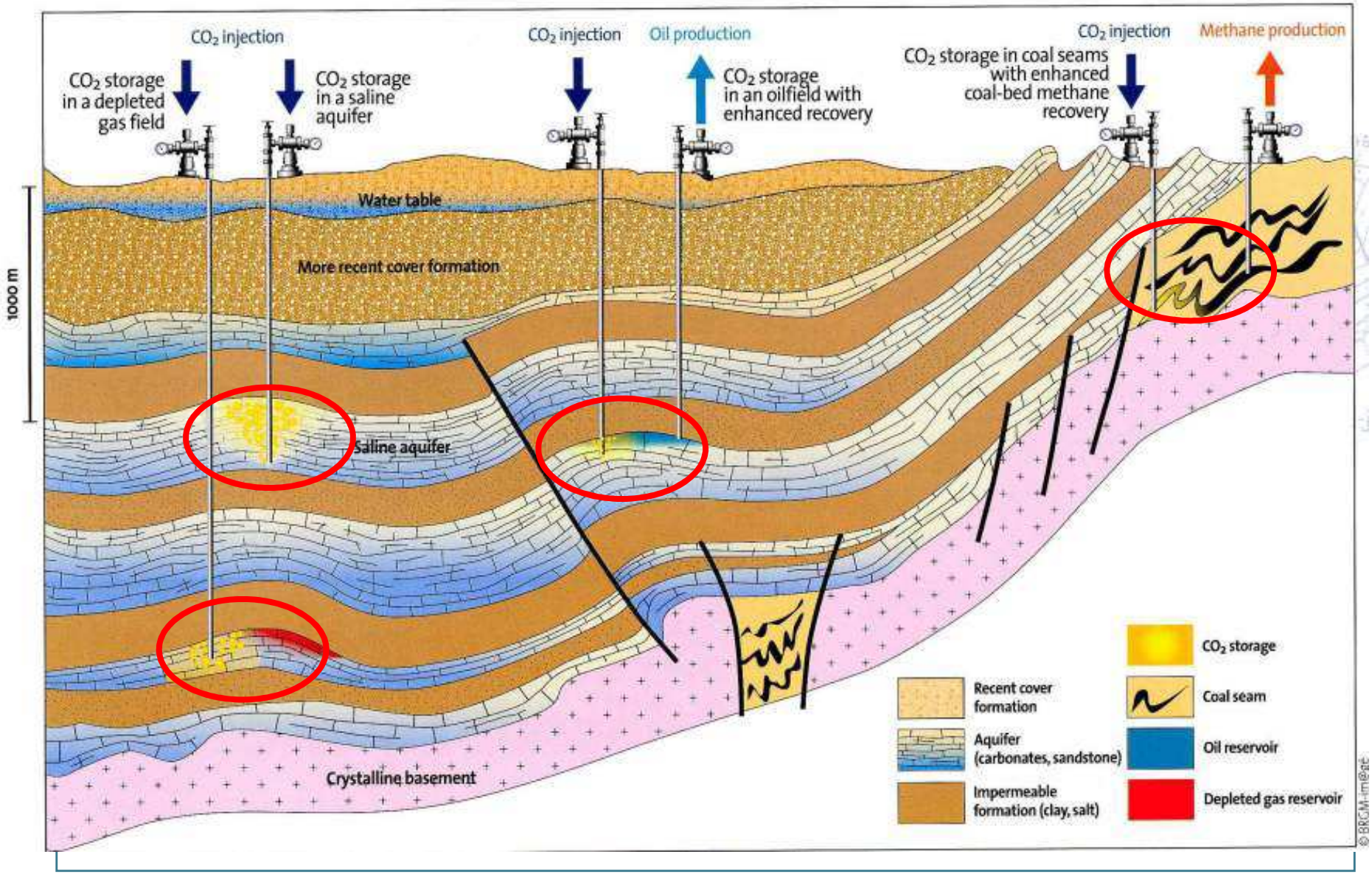
Research activities

> 11 main lines and 42 research programs.

> Main domains

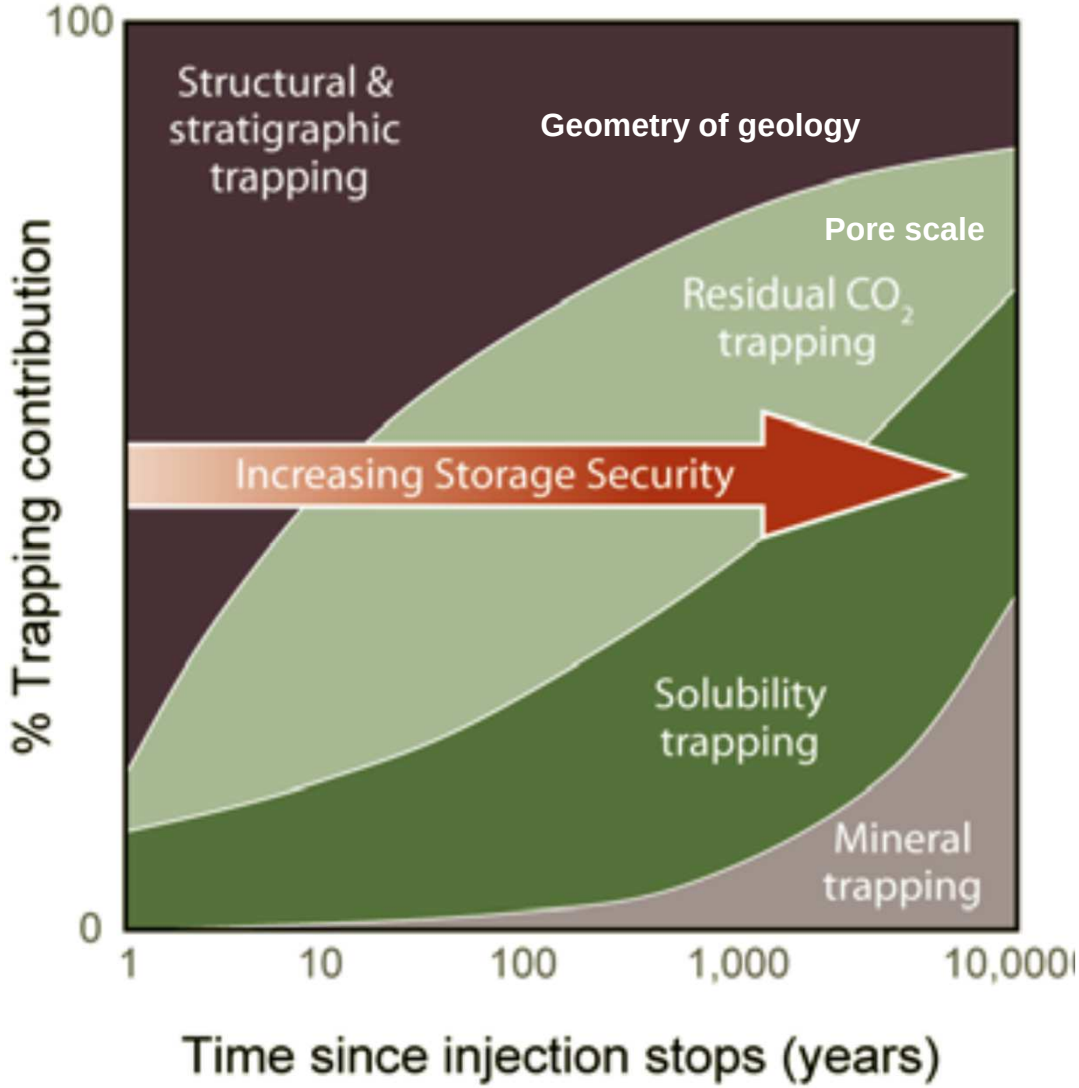
- Geology
- Geothermal Energy
- Post-Mining
- Water
- Lab and experiments
- Mineral resources
- Environement and Ecotechnology
- **Geological storage**
  - *CO2 storage and energy vectors (CH4, compressed air, hydrogen, etc.)*
- **Risks and impacts**
  - *Seismic, volcanic and tsunami risks*
- **Information Systems**
  - *Scientific Computing, software engineering and 3D.*

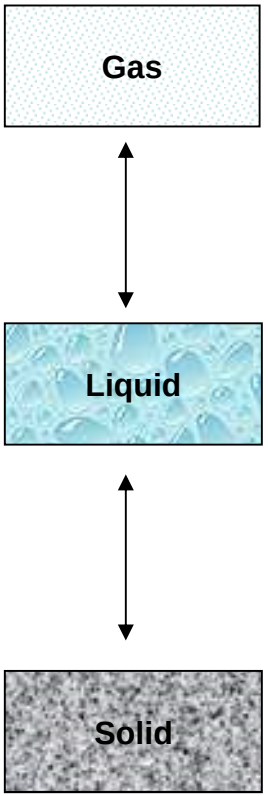
CCS


CO_2 injection
CO_2 storage in a depleted gas field
CO_2 storage in a saline aquifer
CO_2 injection
Oil production
CO_2 storage in an oilfield with enhanced recovery
CO_2 injection
Methane production
CO_2 storage in coal seams with enhanced coal-bed methane recovery
Water table
More recent cover formation
1000 m
Saline aquifer
Crystalline basement
CO_2 storage
Coal seam
Oil reservoir
Depleted gas reservoir
Recent cover formation
Aquifer (carbonates, sandstone)
Impermeable formation (clay, salt)
© BRGM-image

What we need to simulate

(IPCC, 2005)





Migration of supercritical CO_2?

Dissolution of CO_2 in brines/oil?

Geochemical fluid/rock interactions?

The H-Cube project

Hydrodynamics, Heterogeneity and Homogeneization in CO2 storage modeling

http://anr-h-cube.brgm.fr

Hydrodynamics
Heterogeneity
Homogenization

- **Financed by the French Agency of Research (ANR)**
- **Call SEED 2012: Systèmes Energétiques Efficaces et Décarbonnées**

- **Budget 636k€,**
- **Duration: 48 months (Started in January 2013- 2016)**
- **Partners:**

LSCE: Laboratoire des Sciences du Climat et de l'Environnement (**CEA-CNRS-UVQS**)

CEREGE: Centre Européen de Recherches Et d'Enseignement (**Université Aix Marseille**)

Terra3E
Energy Environment Expertise

Géosciences pour une Terre durable
brgm

ULTimateCO2 consortium

CSLF members

Dissemination stakeholders

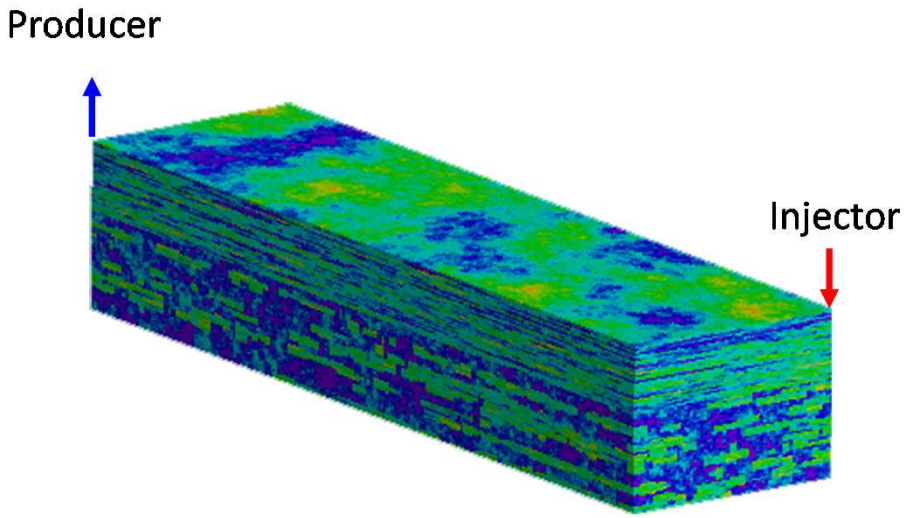
National Regulatory authorities

Private Consultant

T. Torp (retired STATOIL)

SPE10 model

- SPE-10 model, a highly heterogeneous reservoir model (Christie et.al 2001)
- irregular nature of sand/shale distribution
- The model size is 6.4km x 8.8 km × 170m (1.2 millions of cells).



CO2 in supercritical state is injected

- ***with the rate of 390 k tons / year,***
- ***with producing brine groundwater with the rate of 580 k tons / year***

TOUGH2
Transport **O**f **U**nsaturated **G**roundwater and **H**eat

- Developed by K. Pruess, C. Oldenburg, G. Moridis from the Earth Science Division (ESD) of Lawrence Berkeley National Laboratory (LBNL)

- http://www-esd.lbl.gov/TOUGH2/

  - Geothermal reservoir engineering
  - Nuclear waste disposal
  - Oil and gas
  - Carbon storage (sequestration)

Computing ressources

Earth Simulator 2

Upgrade of Earth Simulator system (2009) – Jamstec

- NEC SX-9/E system
- 160 nodes with 8 vector processors and 128 GB of memory
- Total of 1280 processors and 131 Tflops

Challenges on the Earth Simulator

> Opportunity based on a collaboration with research teams from Japan (Univ Tokyo, Taisei)

> *Challenges*
  - Code initially assumed for scalar processor (Linear solver → 70% of elapsed time)

  ⇒Increasing of the vector operation ratio (VOR)

- **Local operation and no global dependency**
- **Continuous memory access**
- **Long innermost loop for vectorization**

⇒Move from the DVBR matrix storage format to the DJDS *(Saad 89, Nakajima 05)*

Tough2-MP

AZTEC (Sandia national Lab.)

- DVBR storage
- Highly memory efficient
- Short innermost loop

HPC-MW

- DJDS storage
- Efficient length of innermost loop
- Contiguous memory access

- **Increasing vector operation ratio**
- **More than 30 times faster**

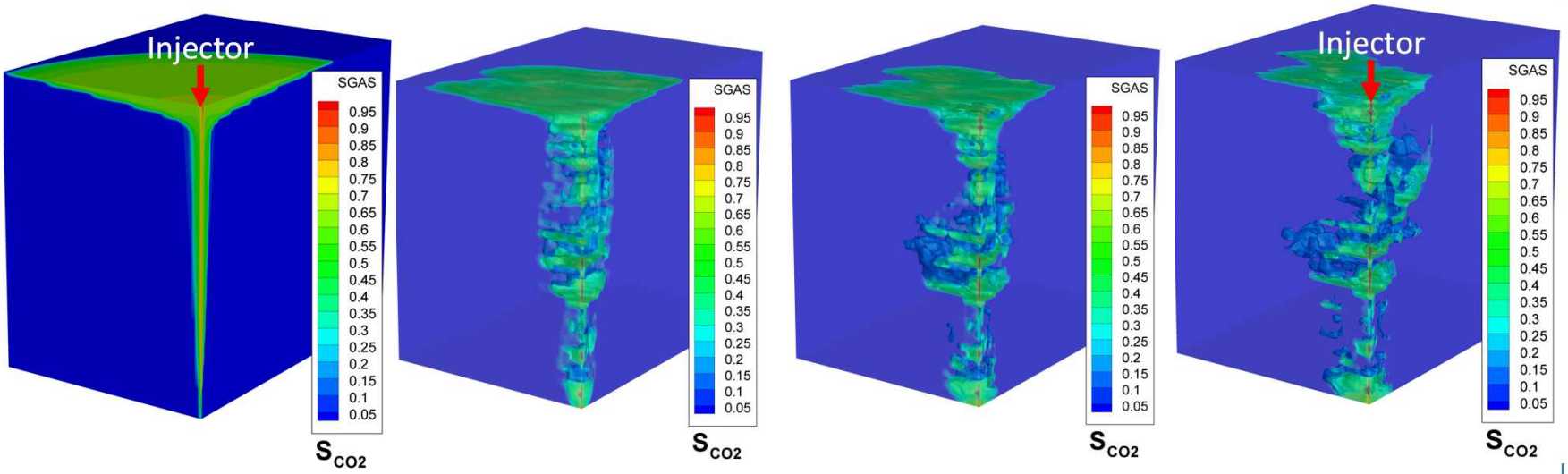
```
*---------------------*
  FTRACE ANALYSIS LIST
*---------------------*

Total CPU Time : 3:01'40"713 (10900.713 sec.)


FREQUENCY  EXCLUSIVE       AVER.TIME     MOPS   MFLOPS V.OP  AVER.    VECTOR I-CACHE O-CACHE   BANK CONFLICT   PROC.NAME
           TIME[sec](  % )    [msec]                   RATIO V.LEN      TIME    MISS    MISS CPU PORT  NETWORK

     1404  9866.592( 90.5)  7027.487  32801.3  10125.3 99.74 254.4  8650.944 308.097 335.221 2243.675 5600.790
solver_vbicgstab33
     1815   450.180(  4.1)   248.033  15795.4   2239.7 99.24 240.9   313.245   0.027 109.621   42.371  165.832 multi
        1   173.400(  1.6)173399.760    268.3      1.1  4.75 134.9     0.279  27.704  21.770    0.284    0.002 readmesh
     1601   154.540(  1.4)    96.527   9063.7   4299.5 95.33 255.4    37.474   0.223   9.599    2.184    9.855 eos
        2    95.918(  0.9) 47958.886    203.0      5.6  0.66  12.9     0.317  20.177  10.810    0.073    0.303 wrifi
        1    62.822(  0.6) 62822.241    240.8      2.0  2.14  47.7     0.411   9.985  10.375    0.048    0.314 rfile
        1    56.596(  0.5) 56596.402    373.9     10.2 67.68  67.5    22.890   0.161   2.993    0.057   19.280 cycit
        2     8.345(  0.1)  4172.604    190.5      3.1  0.42  79.7     0.007   1.263   1.571    0.007    0.006 finalout
      216     6.822(  0.1)    31.584    141.2     31.1 35.60   3.0     5.547   0.000   0.600    0.018    4.627 conver
```

Tough2-MP



3 node x hours (650 time steps)

95 node x hours (5,400 time steps)

250 node x hours (11,000 time steps)

900 node x hours (40,000 time steps)

Homogeneous model

Averaged permeability and porosity

Original SPE10

CO2 distribution after 20 years of injection

Dumux

- Developed by University of Stuttgart
- Multiphase, multicomponent flow and transport in porous media

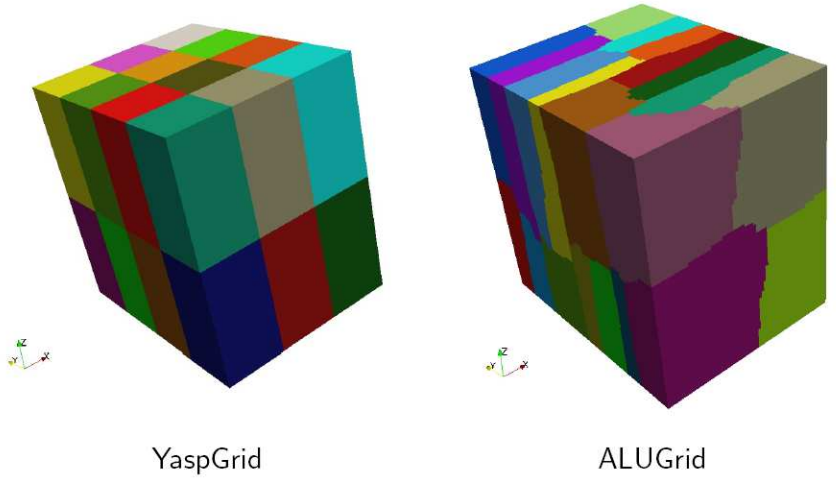
- http://www.dumux.org/

- Incompressible, immiscible flow
  - Isothermal condition → TwoP model (2p)
- Compressible, Miscible flow
  - Isothermal condition → TwoPTwoC model (2p2c)

Dumux

- Speedup = 4.01 (16)
- Speedup = 4.96 (16)

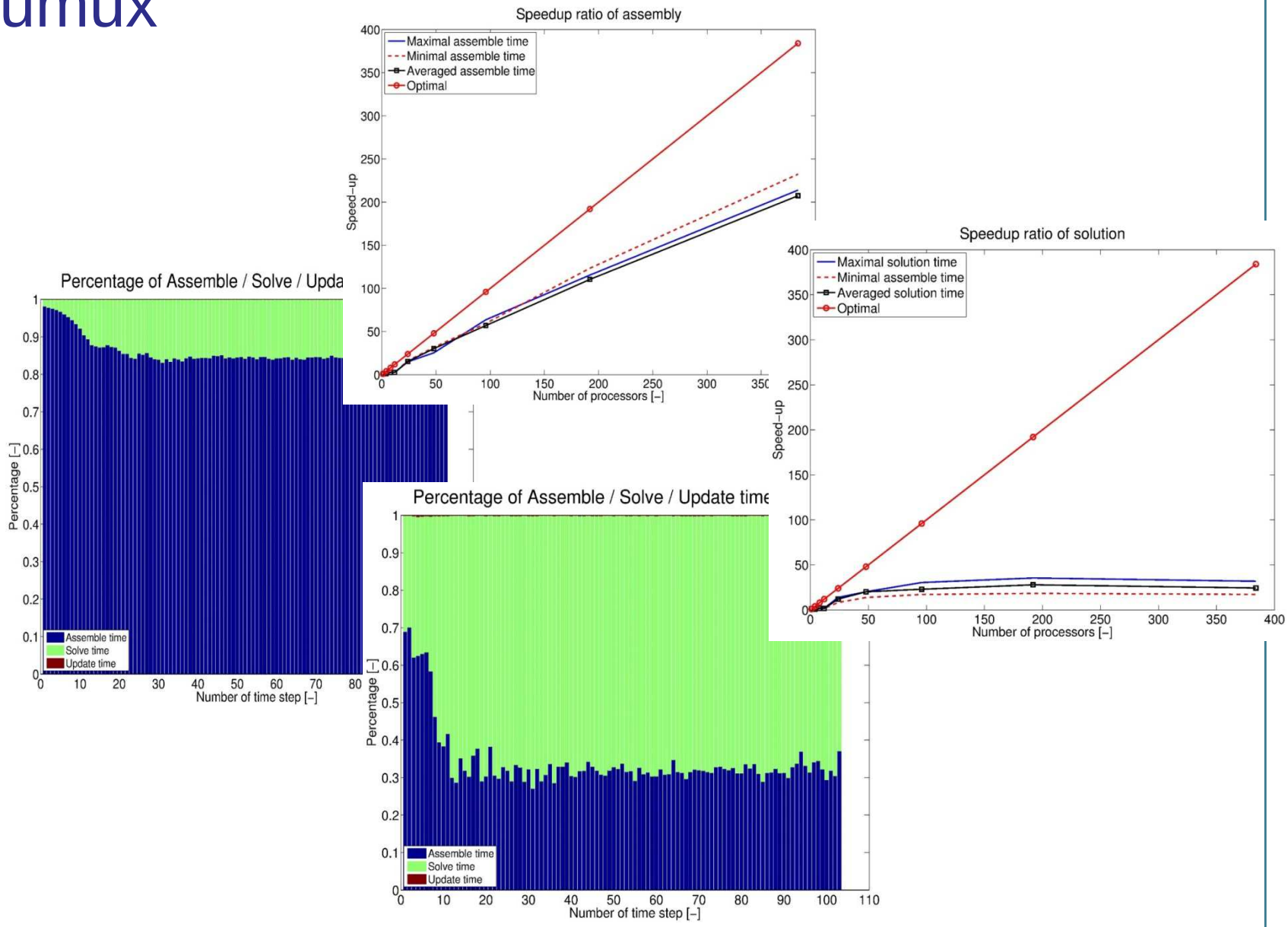
- Impact of linear solver



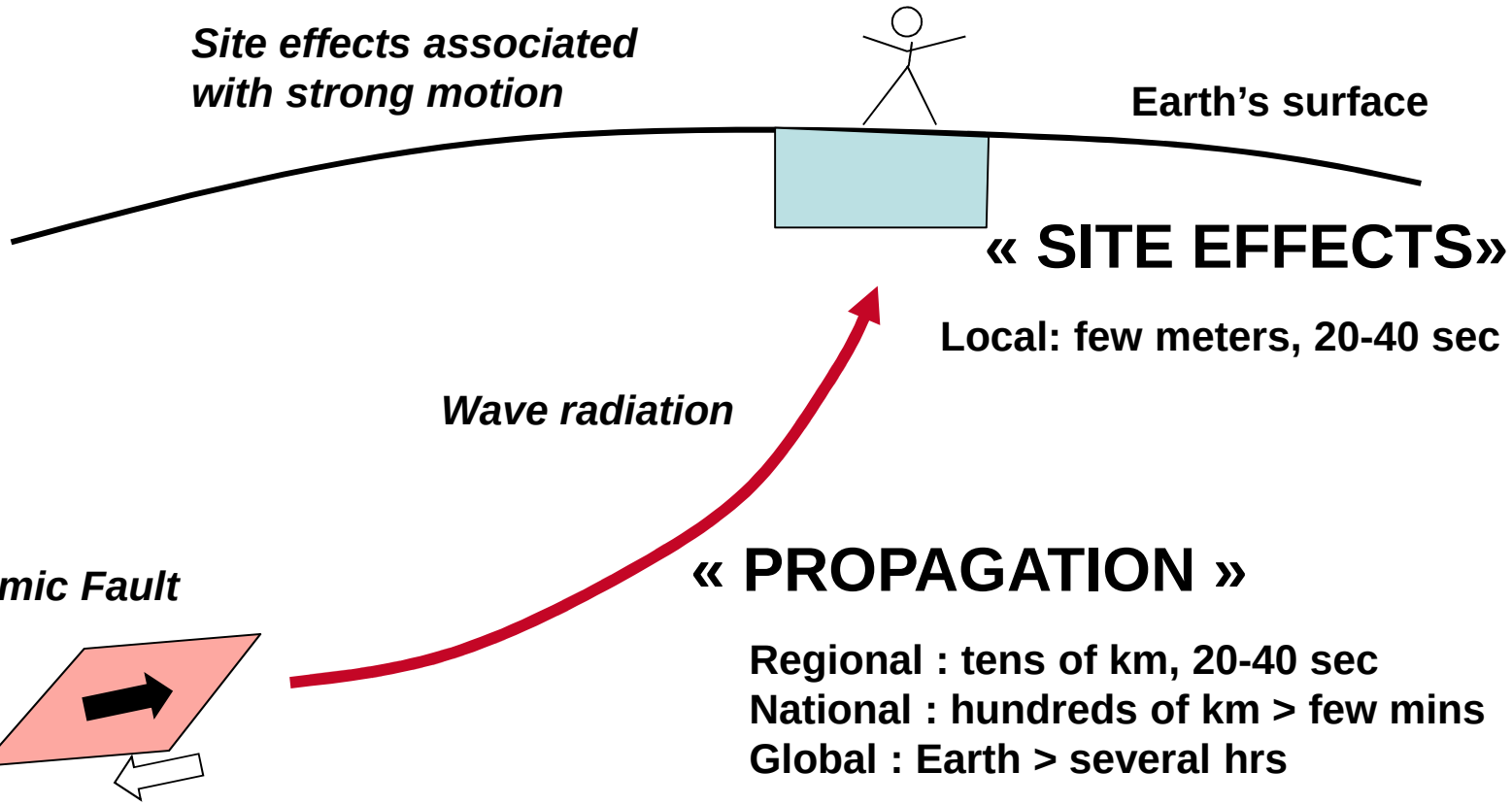
YaspGrid ALUGrid

	Implicit-YaspGrid		Implicit-ALUGrid	
	CPU Time [s]	Time step	CPU Time [s]	Time step
24 cores	3934.94	76	4587.27	73
48 cores	2175.51	73	2503.85	73
96 cores	1668.16	76	1296.53	73
192 cores	1169.24	73	1116.46	79
384 cores	981.37	73	923.626	75

Dumux

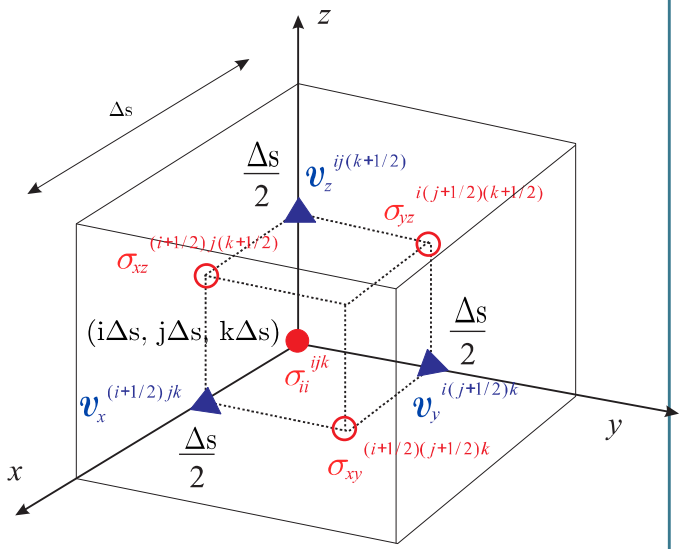

Speedup ratio of assembly
Maximal assemble time
Minimal assemble time
Averaged assemble time
Optimal
Speed−up
Number of processors [−]
Speedup ratio of solution
Maximal solution time
Minimal assemble time
Averaged solution time
Optimal
Speed−up
Number of processors [−]
Percentage of Assemble / Solve / Upda
Assemble time
Solve time
Update time
Percentage [−]
Number of time step [−]
Percentage of Assemble / Solve / Update time
Assemble time
Solve time
Update time
Percentage [−]
Number of time step [−]

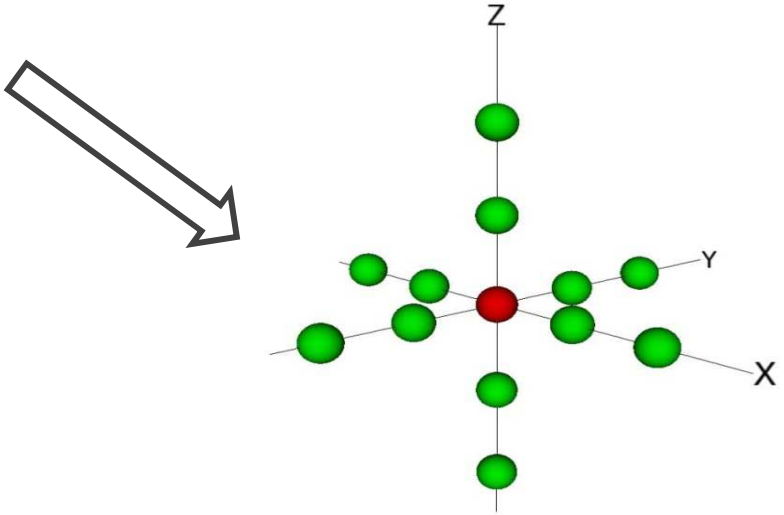
Earthquake risk assessment



Finite-differences → ONDES3D code

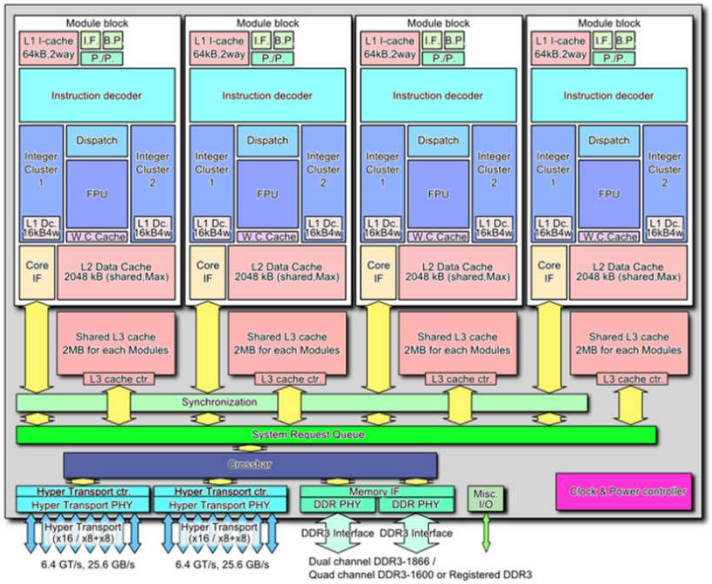
```
for x ← 1 to x_dimension
  do { for y ← 1 to y_dimension
       do { for z ← 1 to z_dimension
            do { COMPUTE_VELOCITY() }
for x ← 1 to x_dimension
  do { for y ← 1 to y_dimension
       do { for z ← 1 to z_dimension
            do { COMPUTE_STRESS() }
```

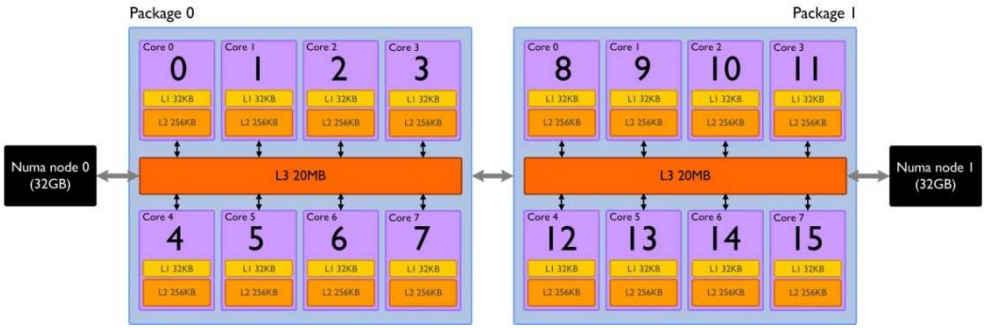


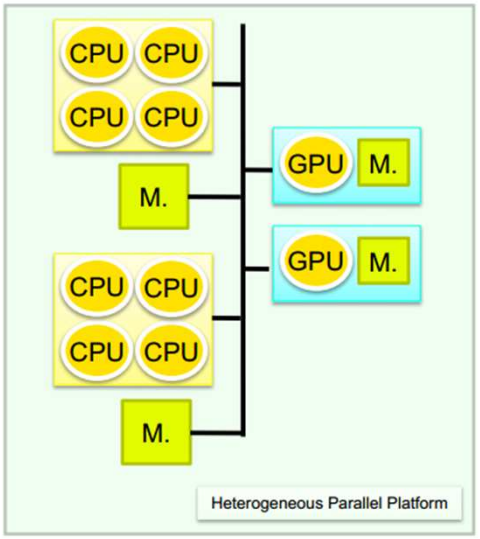


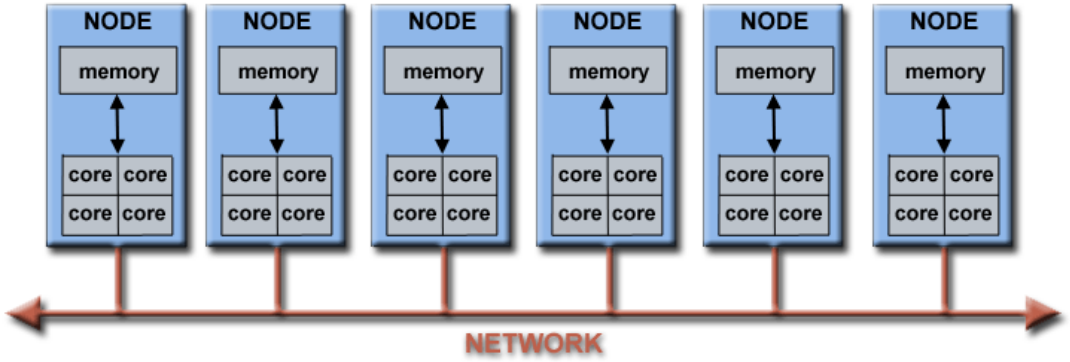
4th-order stencil

data transfer and load imbalance





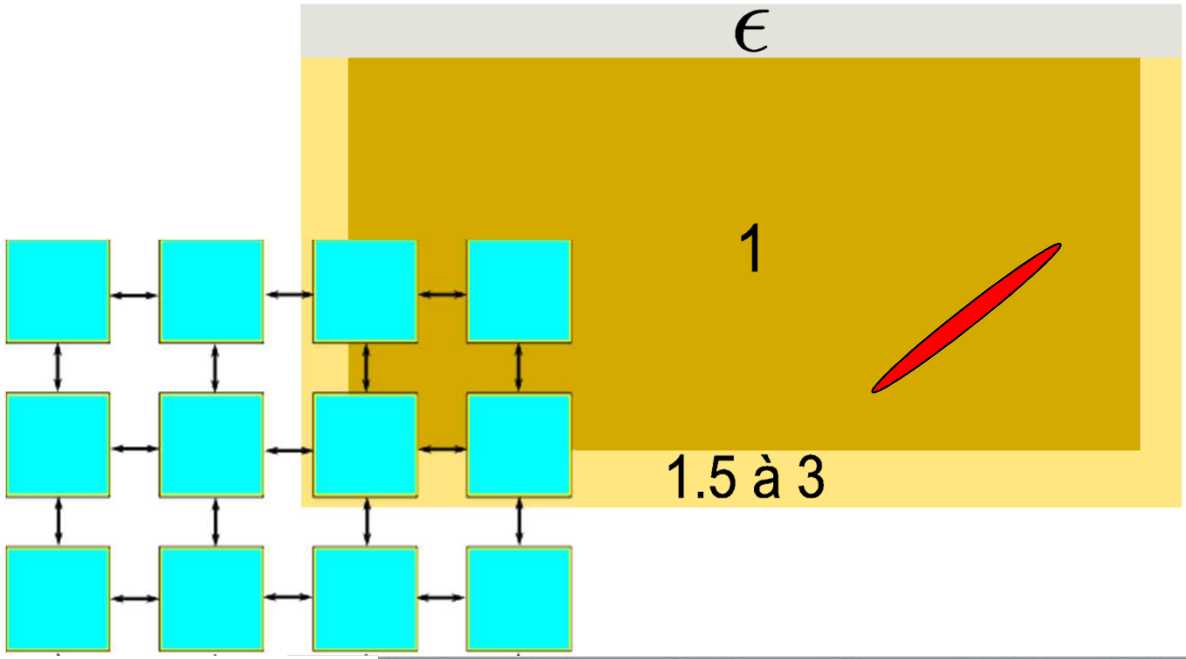


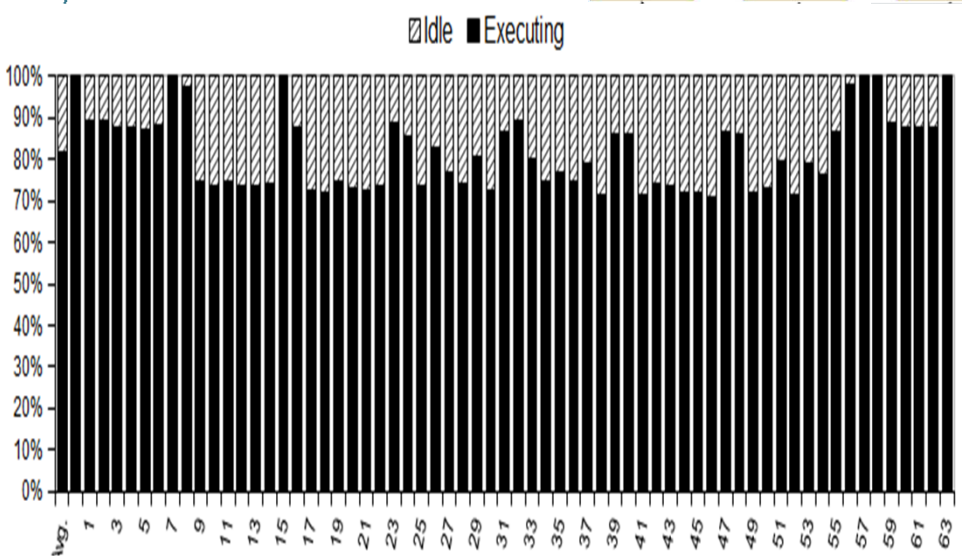


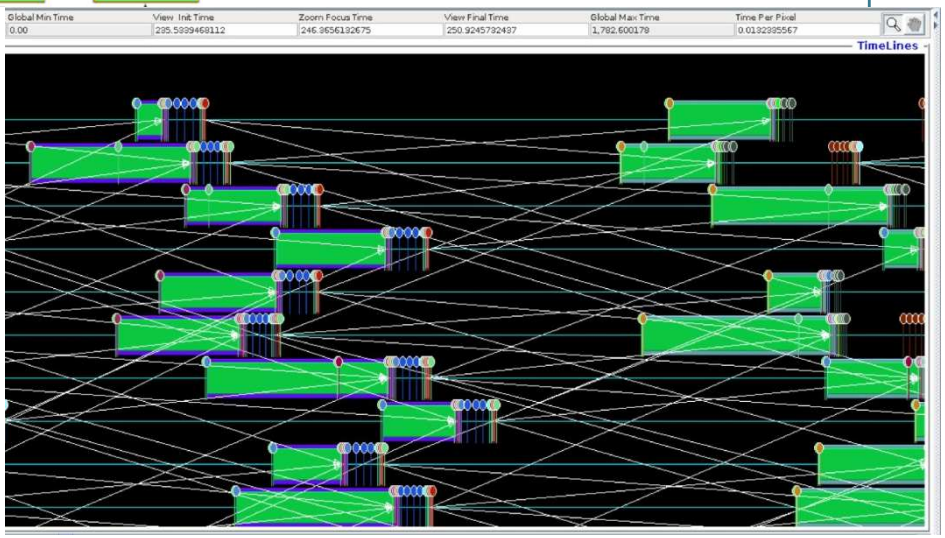
Motivation

Load imbalance

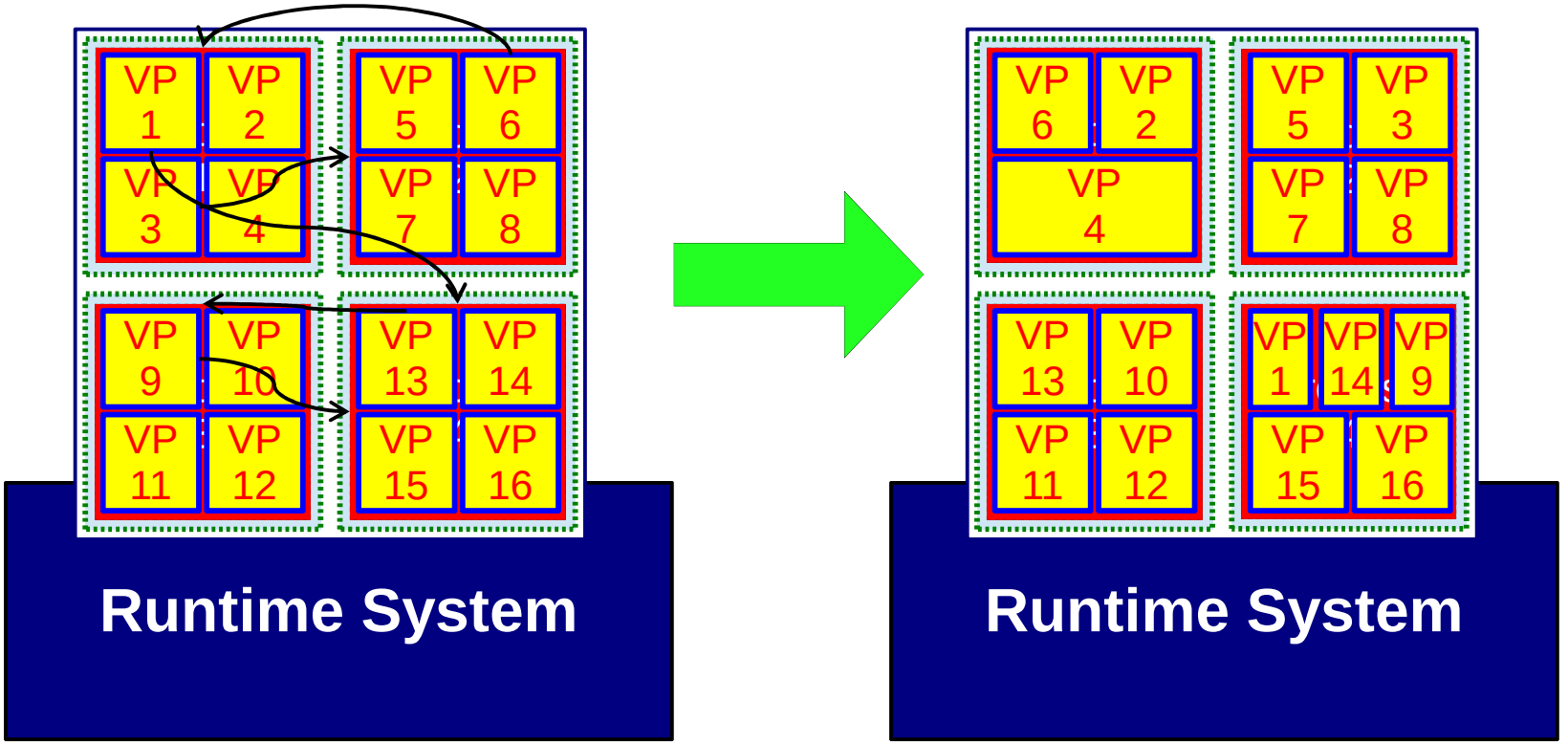
- From ABC
- From I/O
- From seismic sources







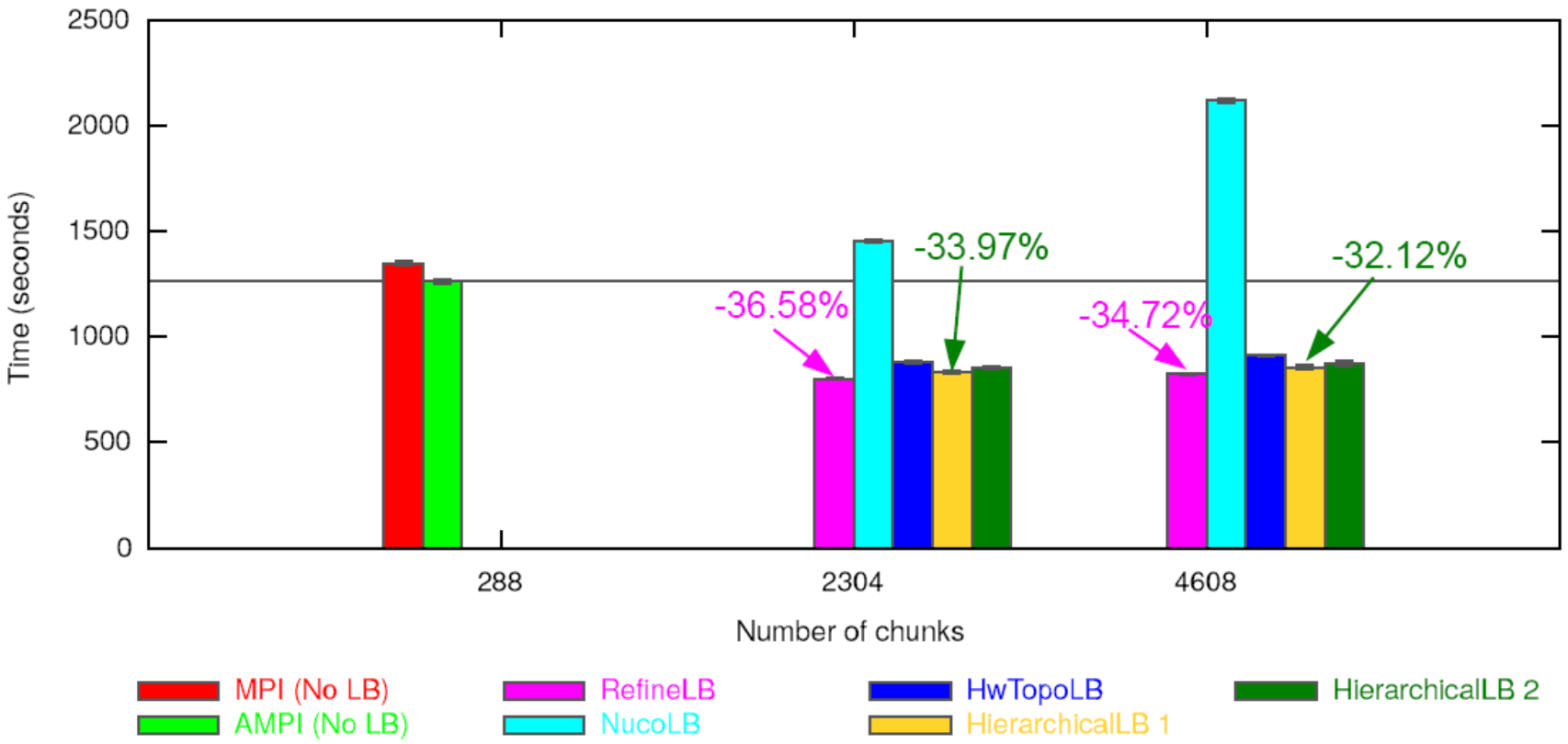
Dynamic Load Balancing



Charm++/AMPI overdecomposes the domain into virtual processors (VPs)

- **Multiple VPs per core / Each VP** implements **one MPI task**

Results

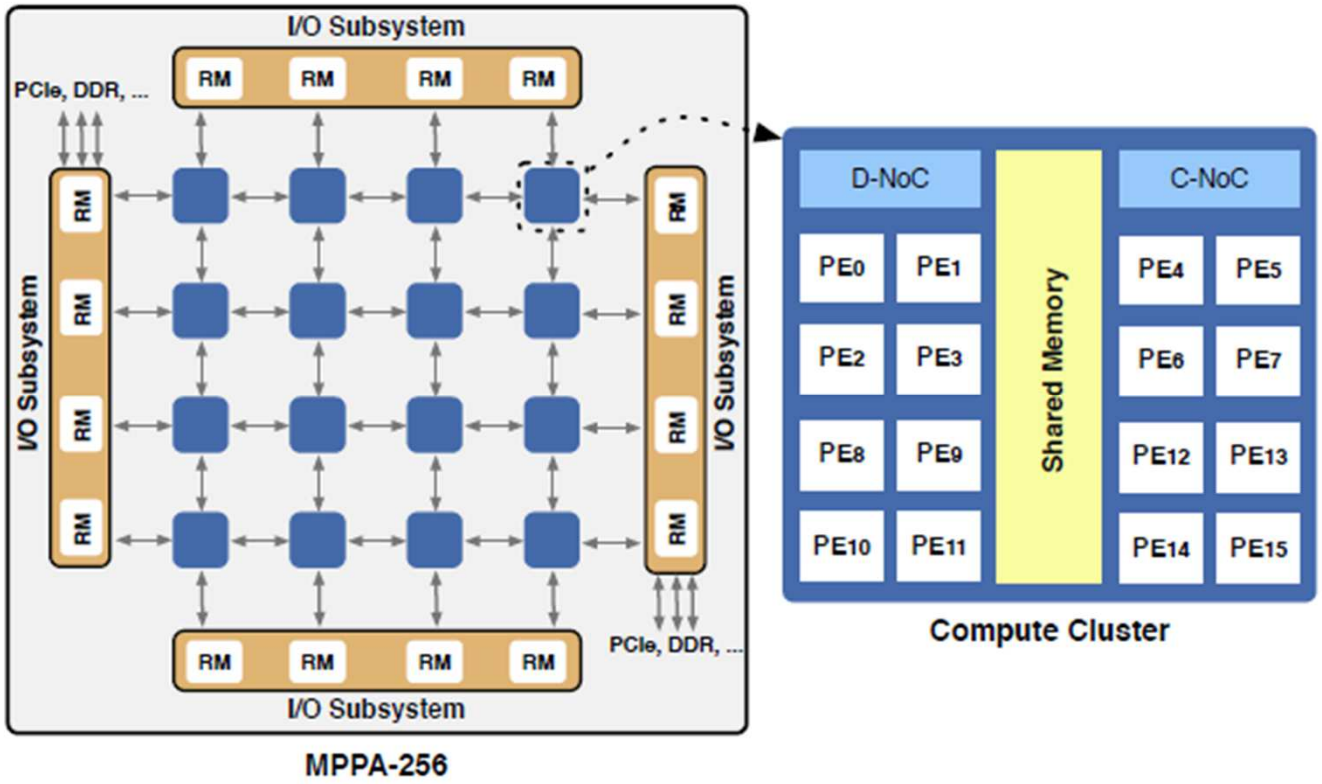


Up to 36.58% performance improvement on a 288 core system.

Motivation → manycores architecture

MPPA-256 from Kalray

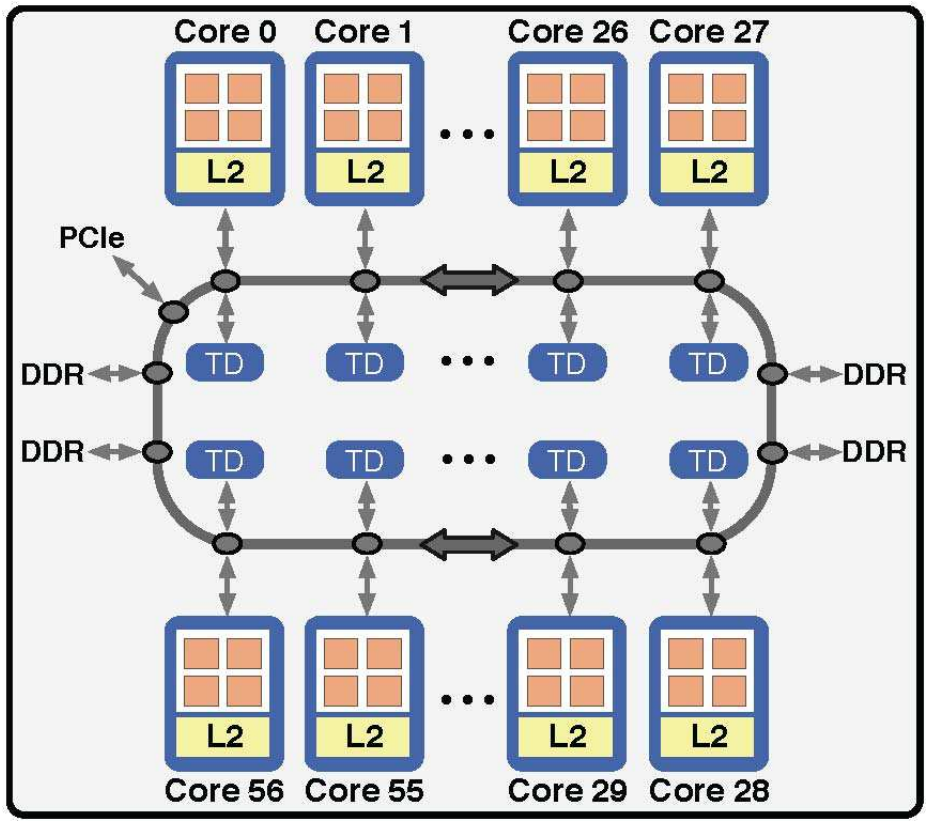
- 2 MB on compute cluster - 4 GB on I/O nodes
- Theoretical 230 Gflops – 5 W



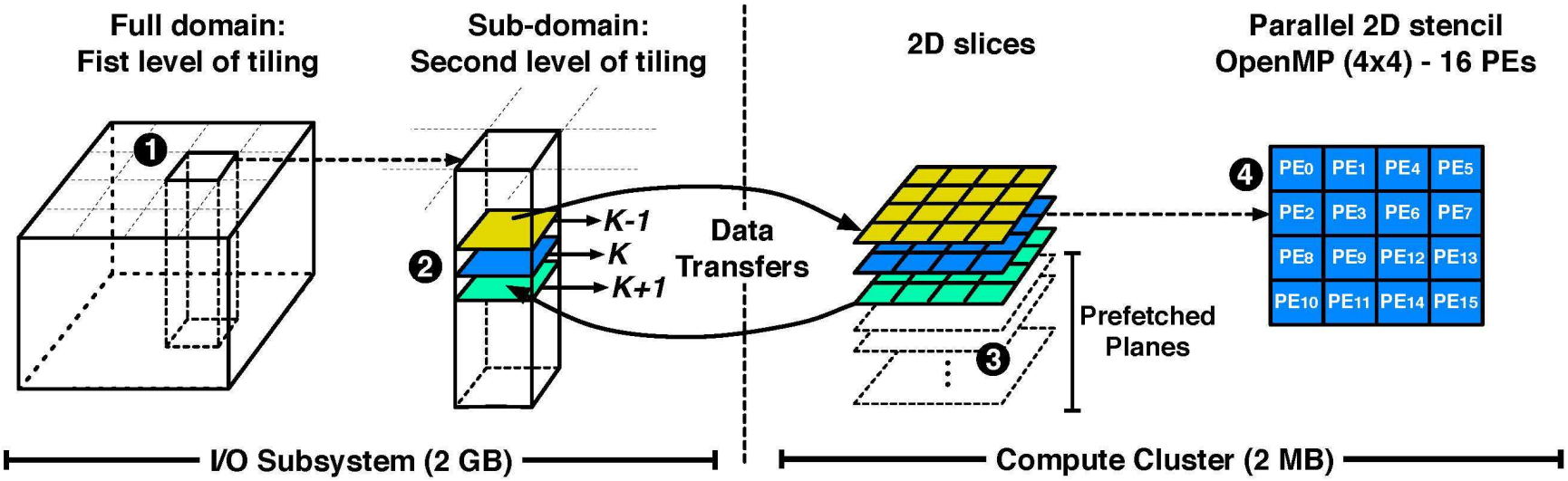
Motivation → manycores architecture

Intel Xeon Phi

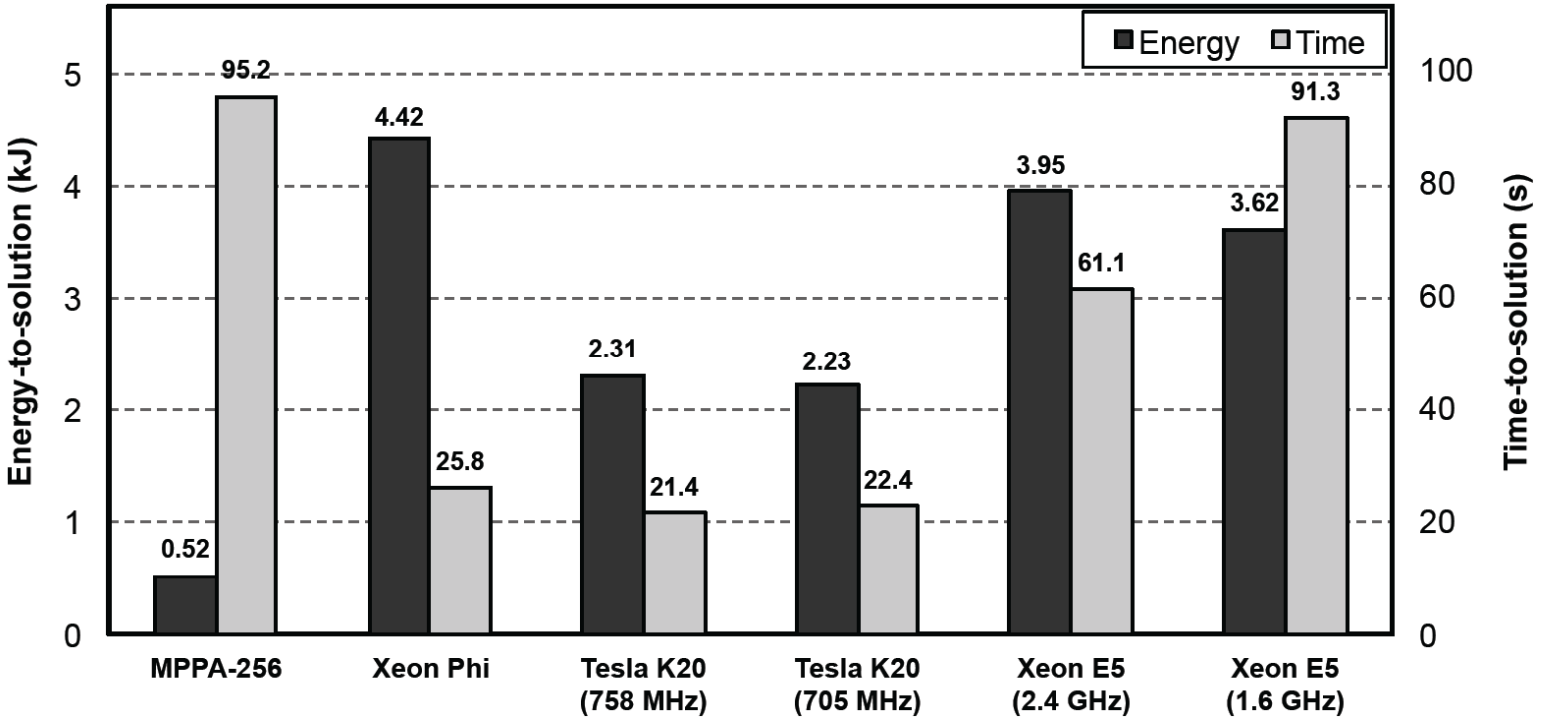
- 57 cores – 512 KB of L2 Cache
- 6 GB of shared DDR memory / 2.0 TFLOPS in SP



Seismic stencil - algorithm



Results



Energy consumption

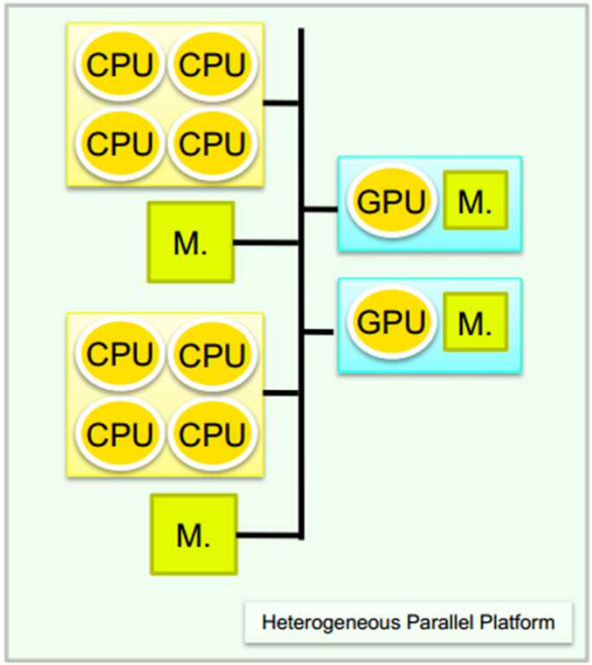
- MPPA : average reduction of energy consumption ~ 80%

Time-to-solution

- Ratio GPU/Phi → Peak = 1.75 – Measured = 1.20
- Ratio MPPA/Phi → Peak = 14 – **Measured = 3.68**

Motivation

- Reuse existing CPU and GPU numerical kernels
- Evaluate task programming model.
- Tackle Heterogeneous architectures (CPU + GPU / CPU + Xeon Phi)



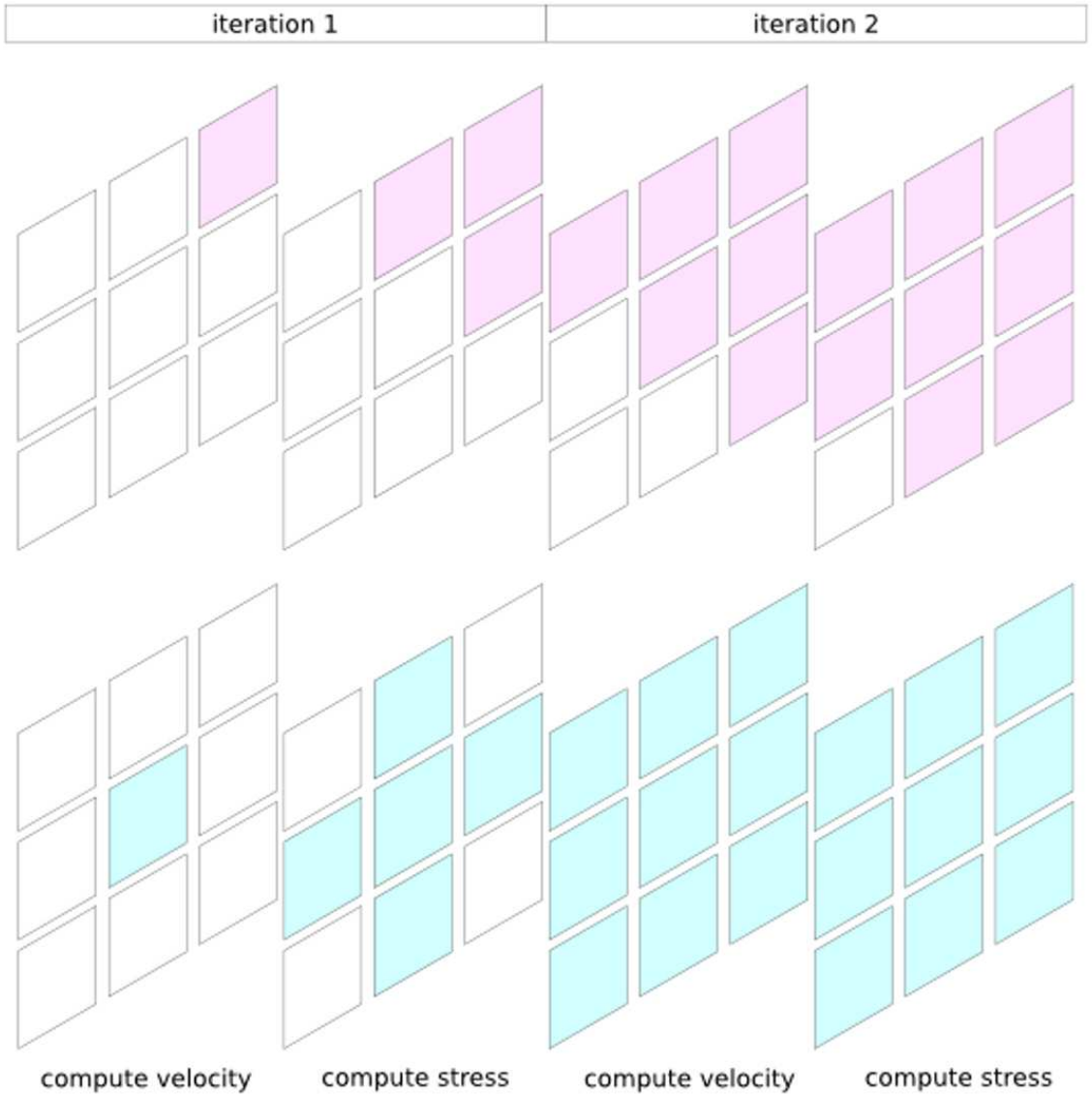
Ondes3D on top of StarPU

StarPU programming model

- Task programming model
- Directed Acyclic Graph according to data dependencies
- StarPU manages memory transfers .

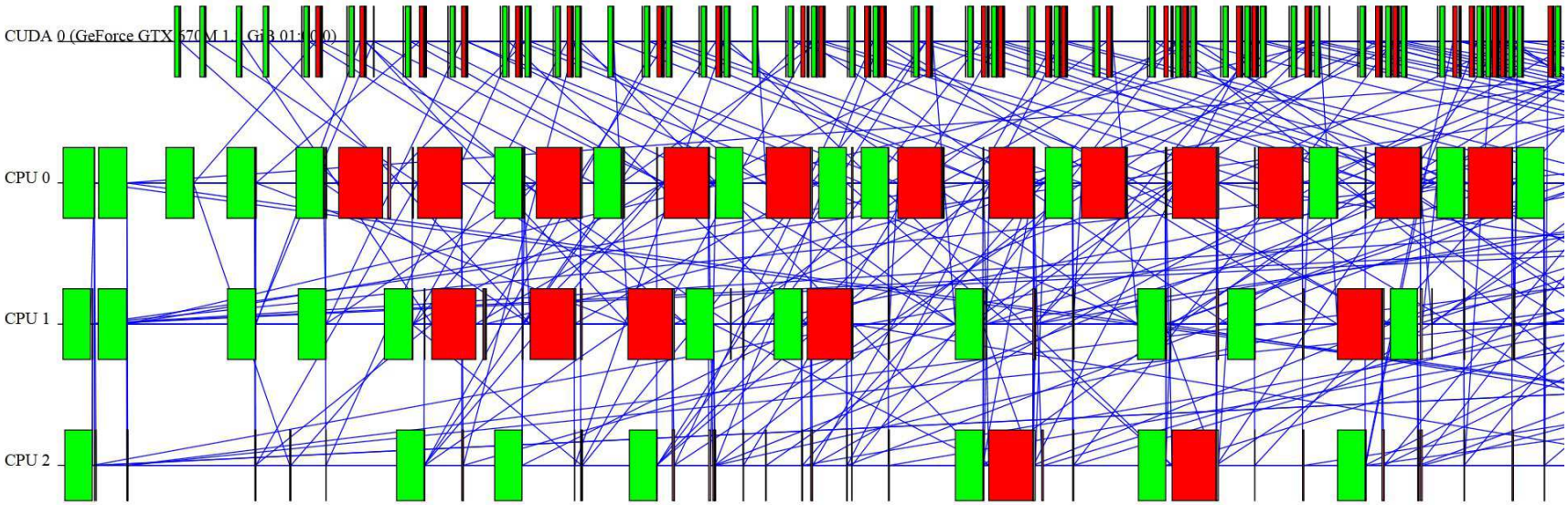
- **For each iteration and each block :**
  - Update source
  - Load stress boundaries & Compute velocity from stress
  - Save & send velocity boundaries for each direction
  - Load velocity boundaries & Compute stress from velocity
  - Save & send stress boundaries for each direction
  - Record seismograms

Scheduling strategies



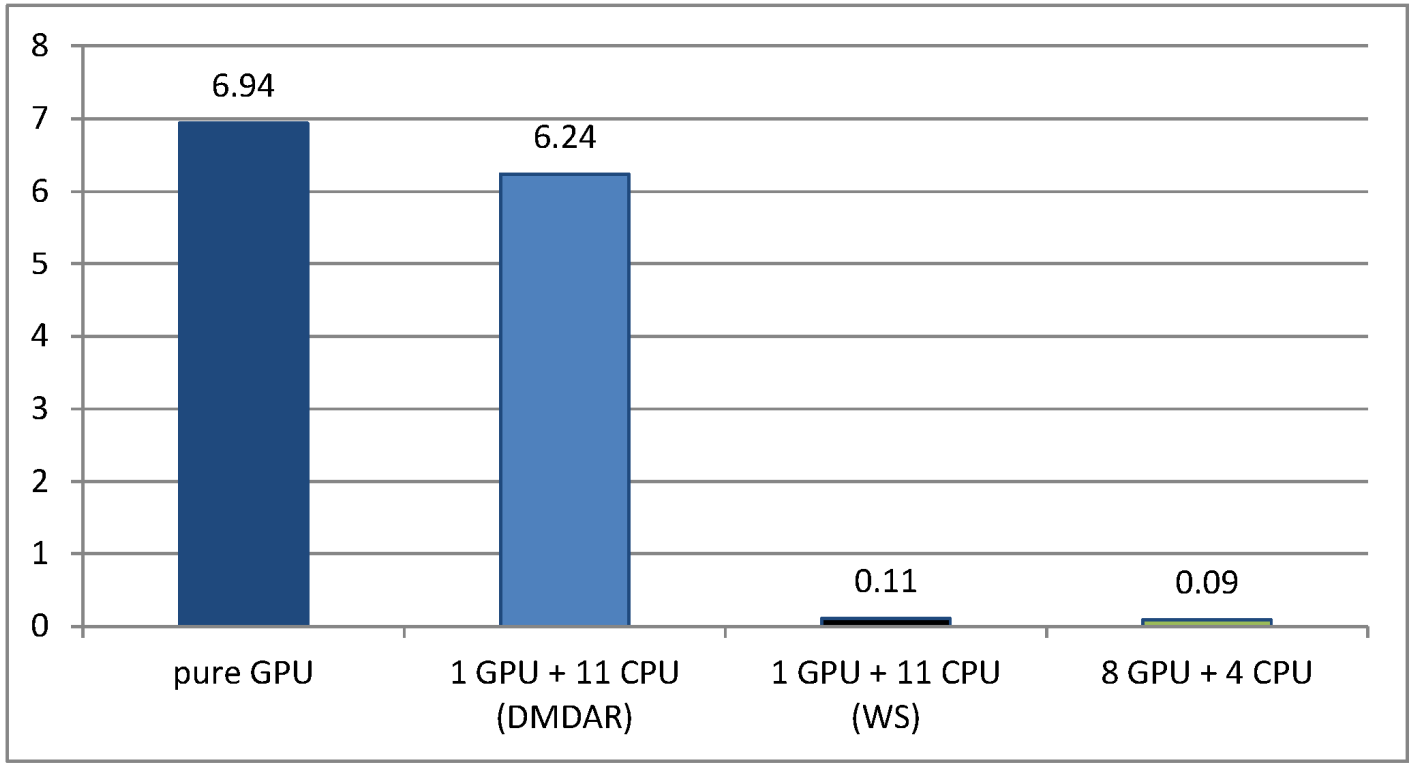
- Dependencies in the DAG for a kernel on a given block (here in a grid of 3x3 blocks) Colored tasks can not be scheduled before the initial one is over.
- If a task take longer to compute on a slow PU than the rest of the grid on a fast PU, there will be waiting for the fast PU on the next step
- Time taken to compute a task (here a given kernel) on a slow PU should not exceed the time taken to compute the rest of the grid on faster PU for the same kernel

Scheduling strategies



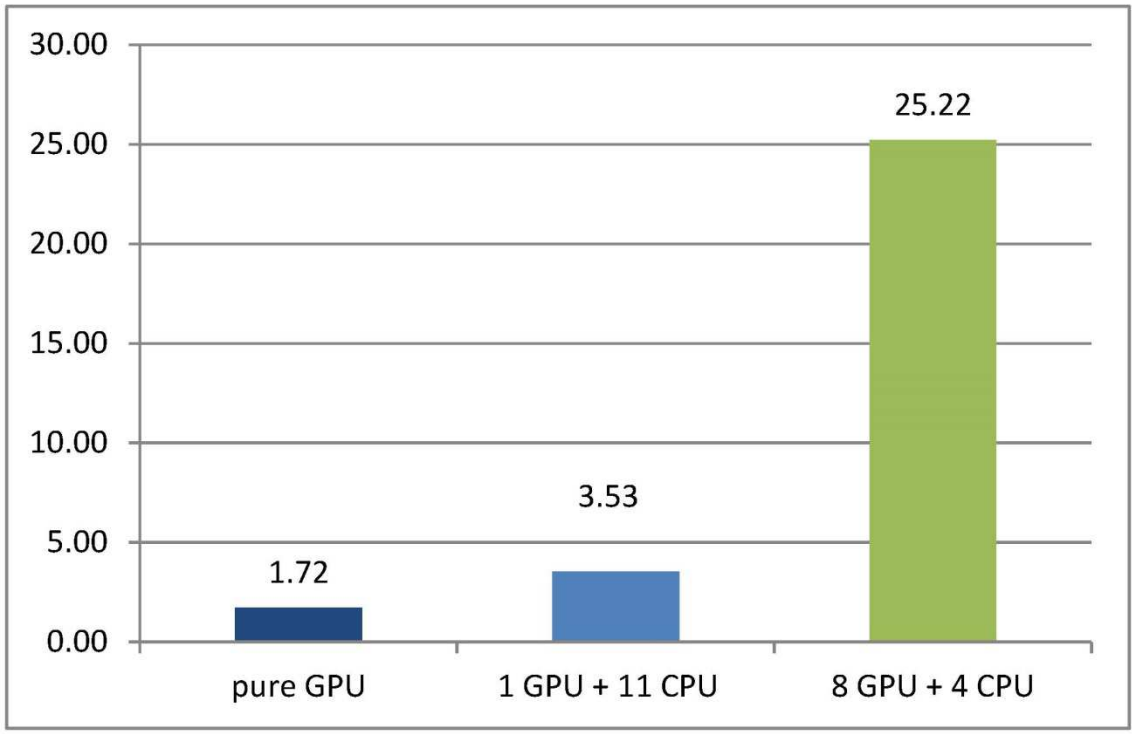
- Complexity of heterogeneous scheduling
- Red and green velocity and stress kernels
  - Speedup ratio between architectures
  - Significant part of idle time due to data dependency

Incore results



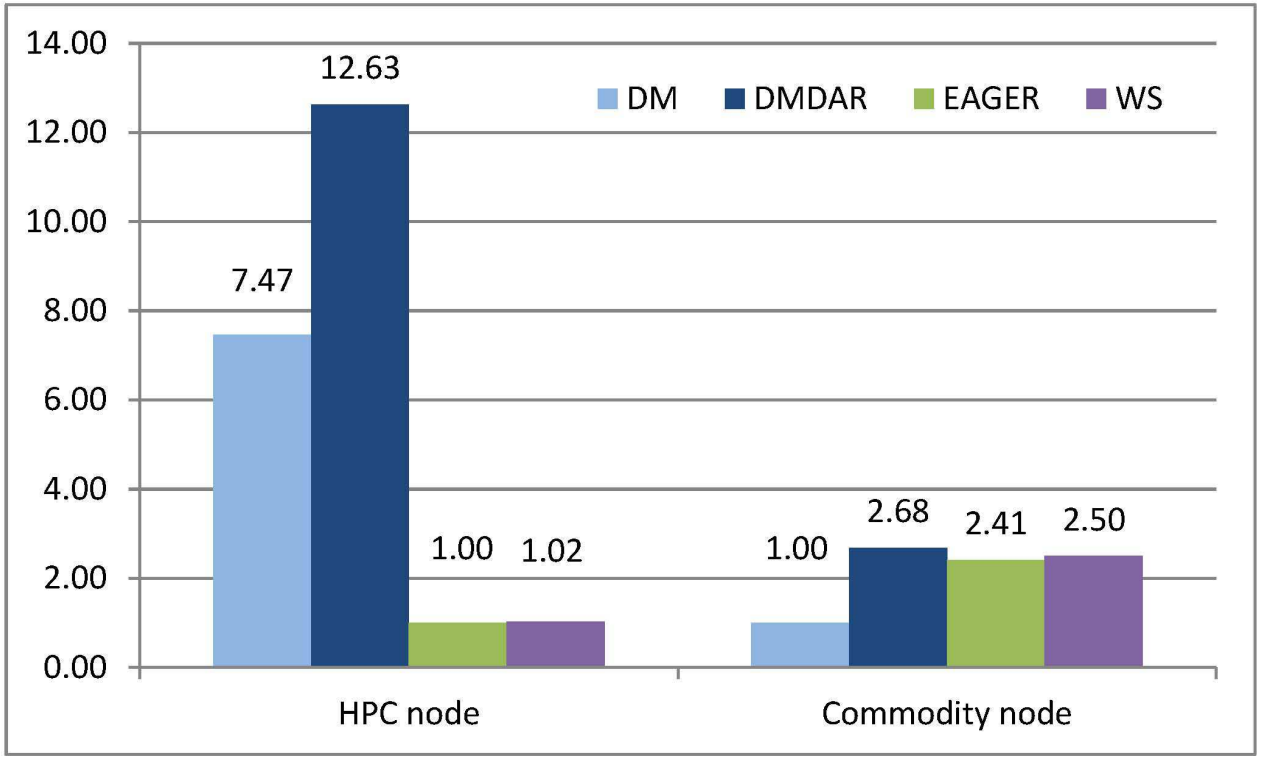
- Cost of data transfer using the PCI bus
- Hybrid (DMDAR) vs Hybrid (WS)

Out-of-core results



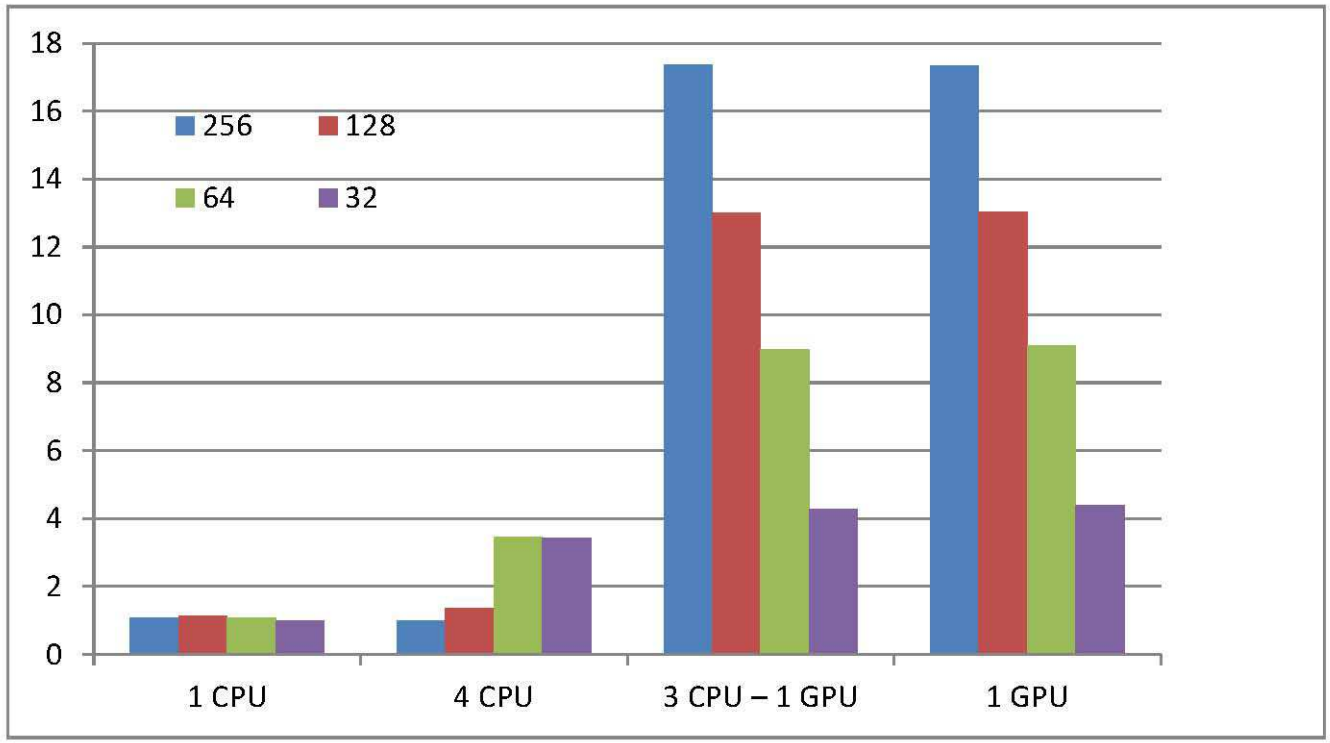
- Impact of data transfer for incore and ooc experiments (pure GPU)
- We are far from peak performance level (i.e. speedup with incore GPU)

Impact → Scheduling strategies



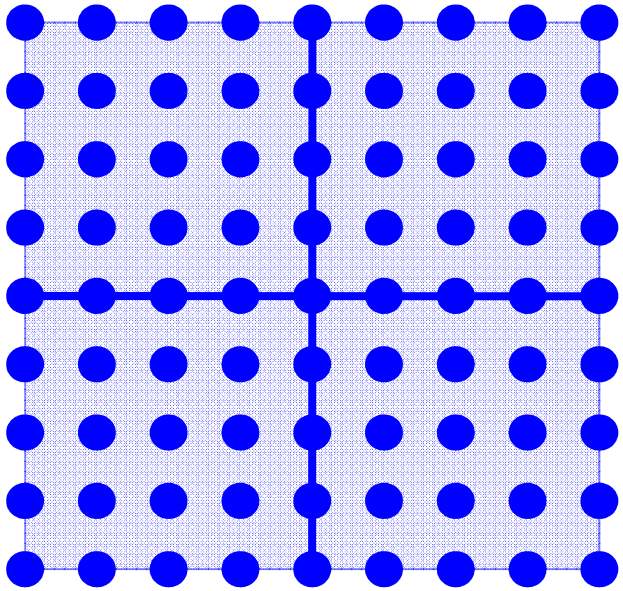
- Contention on the memory bus.
  - Strong impact on the HPC node → 8 GPU + 4 CPU
  - Reduced on commodity-node → 1 GPU + 3 CPU

Impact → Size of the block



- Competition between two behaviors with a fixed problem size
  - Create enough (but not too much) tasks for parallelism at StarPU level
  - Suitable block size with respect to the underlying architecture.

Stencil computation – space blocking



> **Standard methodology**

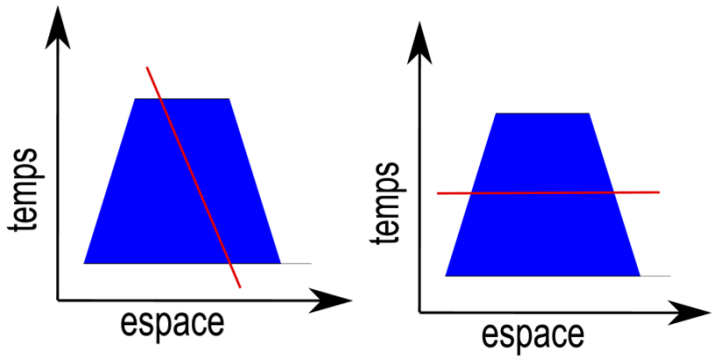
> 2D blocking and reuse in the third direction *(Rivera 2000)*

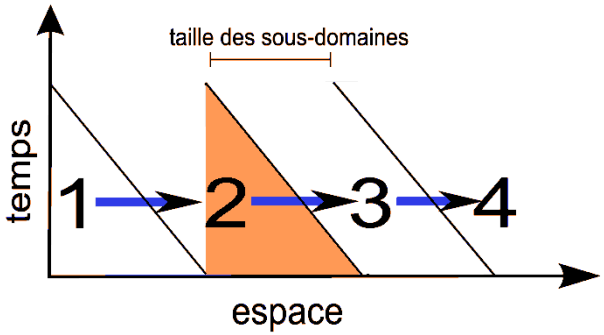
> Poor reuse opportunity → Blas 1 formulation **O(1)**

Stencil computation – spacetime blocking

$$\rho \frac{\partial^2 u_i}{\partial t^2} = \frac{\partial \sigma_{ij}}{\partial x_j} + f_i$$

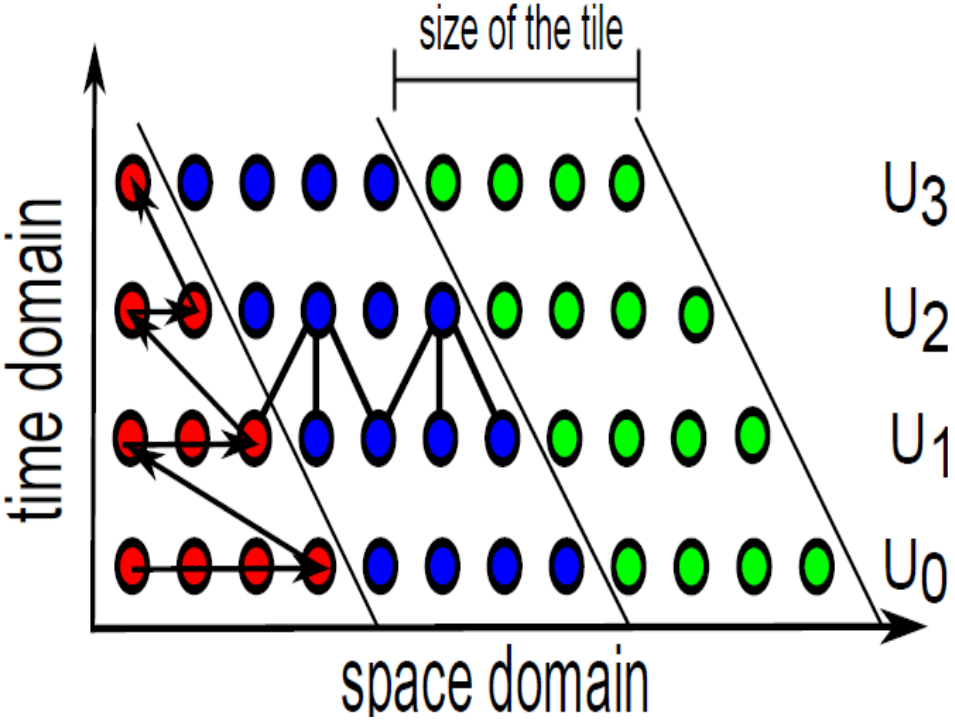
> **Computer science approaches** → Frigo 2005, Datta 2009

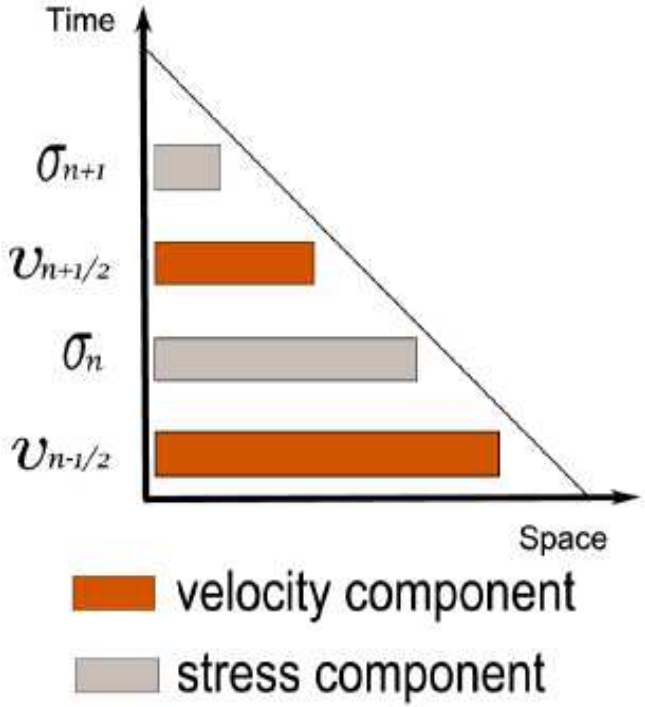




> **Numerical analysis** → *Parareal, PITA (Lions 2001)*

Stencil computation – spacetime blocking





- Reuse data accros several time steps
- Wavefront algorithm
- Blas 3 formulation → **O(n)**

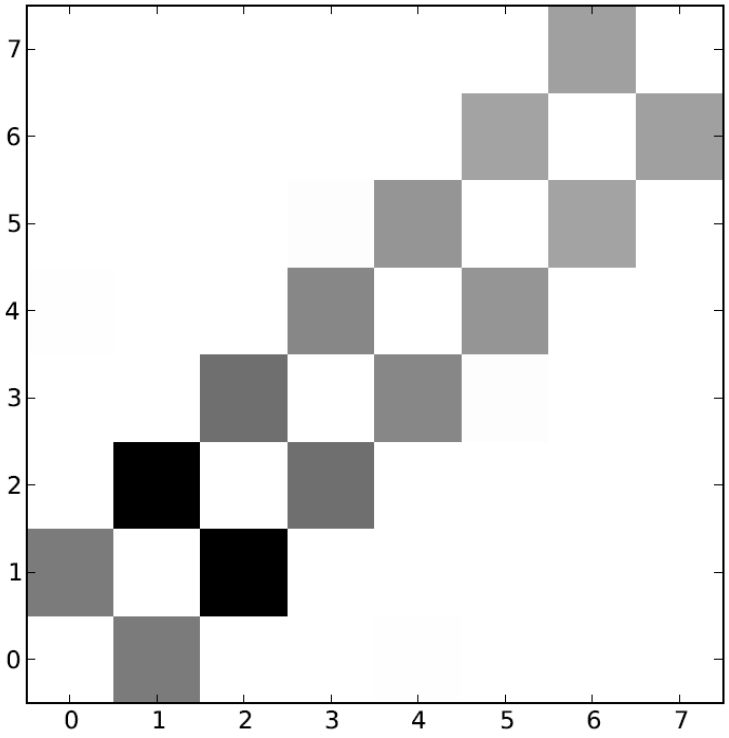
Characterization

How do threads communicate.

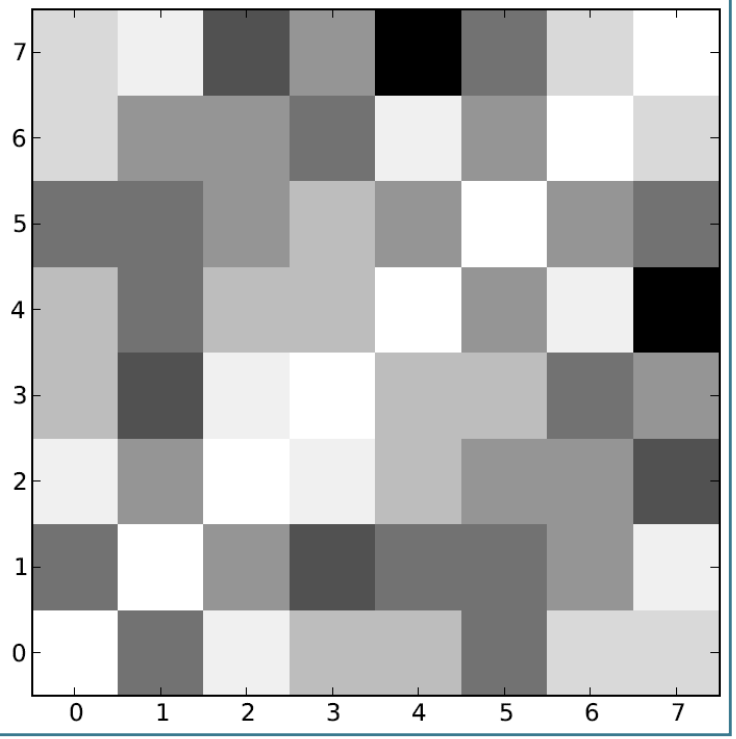
- **Amount** of communication
- **Heterogeneity** (Communication Pattern)

SPCD: shared pages and communication detection

Naive version



Skew version



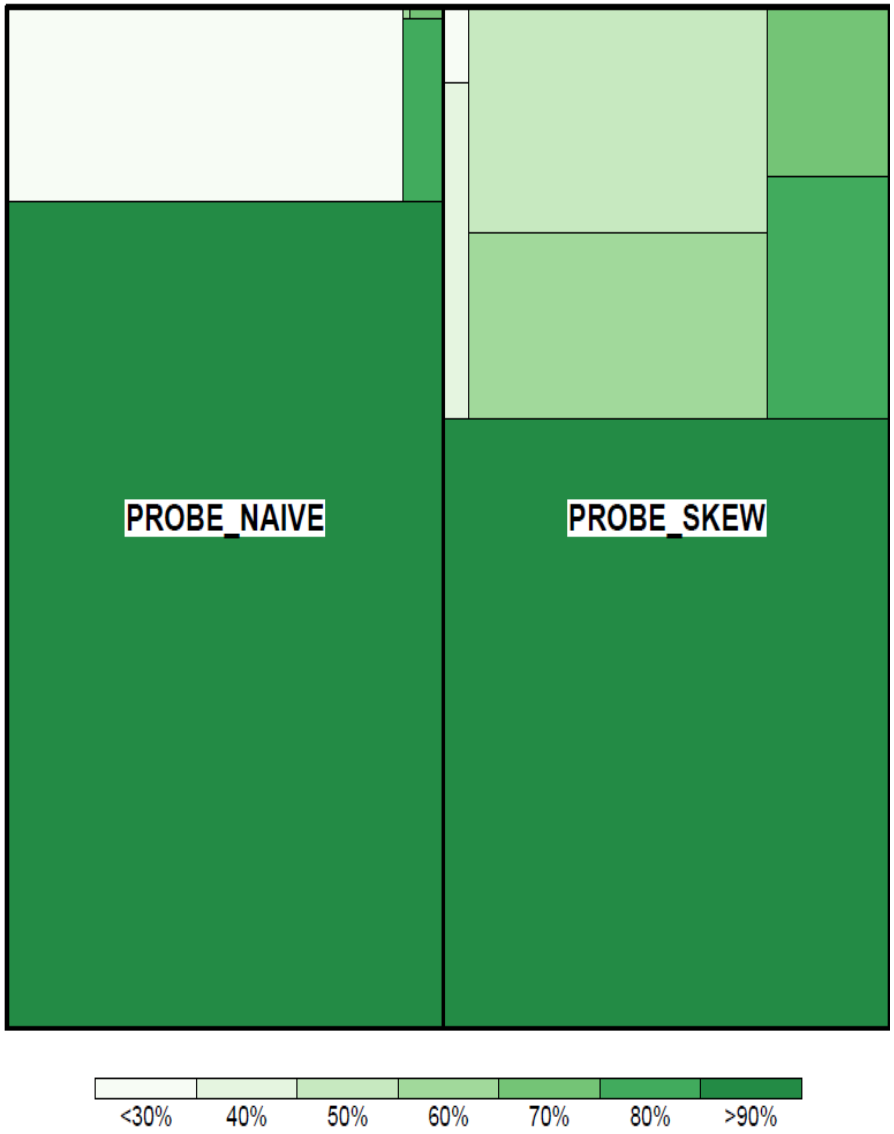
Characterization

Exclusivity:

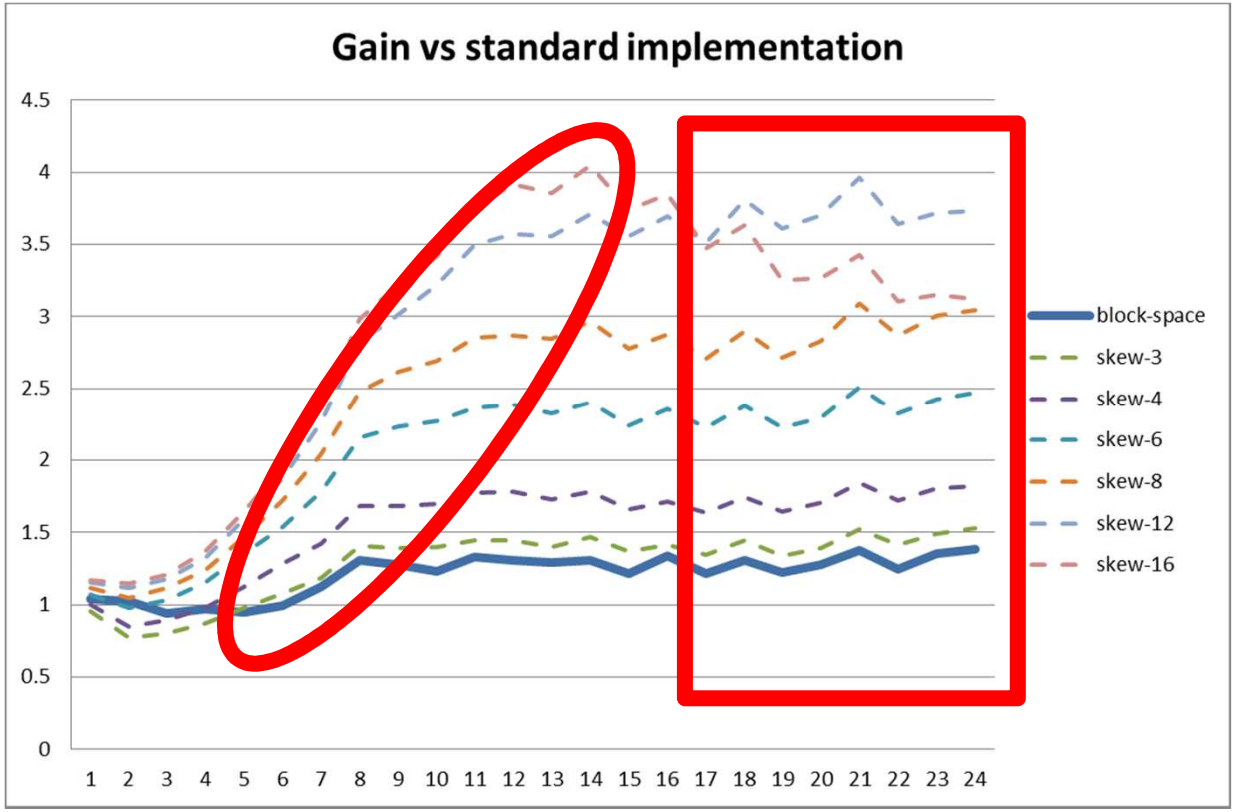
$$\frac{\#\ highest\ memory\ accesses\ from\ a\ node}{\#\ memory\ accesses\ from\ all\ nodes}$$

- Max: 100% (page fully exclusive)
- Min: $1/\#nodes$ (page fully shared)
  → with 4 nodes, min: 25%

- ***Treemap***



7-points → *Impact of the « reuse » parameter*



- Tradeoff between optimal size of the tiles and « reuse » opportunity
- Shift in the upper-bound limit with respect to the number of cores

⇒Better usage of memory BW available

Seismic wave equation

Platform	Spaceblocking	Spacetime blocking
Intel Xeon E5-2697v2	*x1.19*	*x3.59*
Intel Xeon E5-2697v3	*x1.27*	*x2.68*

- Limited speedup from space only blocking
- Stronger impact of vectorization on Haswell vs complexity of the kernel.

Variability of earthquakes ground motions at sedimentary basin scale (1D)

⇒**Uncertain sources: wave velocities, thickness of layers, damping ratio**

⇒**Quantities of interest : PGA, acceleration, spectral ratio**

> Meta-modeling:

- Polynomial regression
- Neural network
- Kriging (or Gaussian process)
- Chaos Polynomial: convergence in case of smooth process

> Chaos Polynomial (spectral expansion)

- Need to solve some direct models (Design of experiments)
- Low number of evaluations in comparison with Monte-Carlo
- Evaluation of the polynomial coefficents by projection

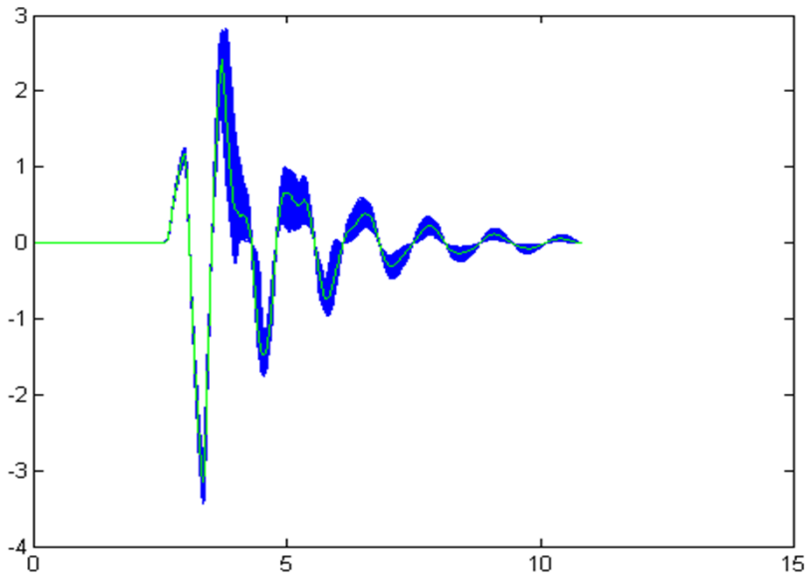
> Difficulties: - Stochastic stiffness of the quantities of interest

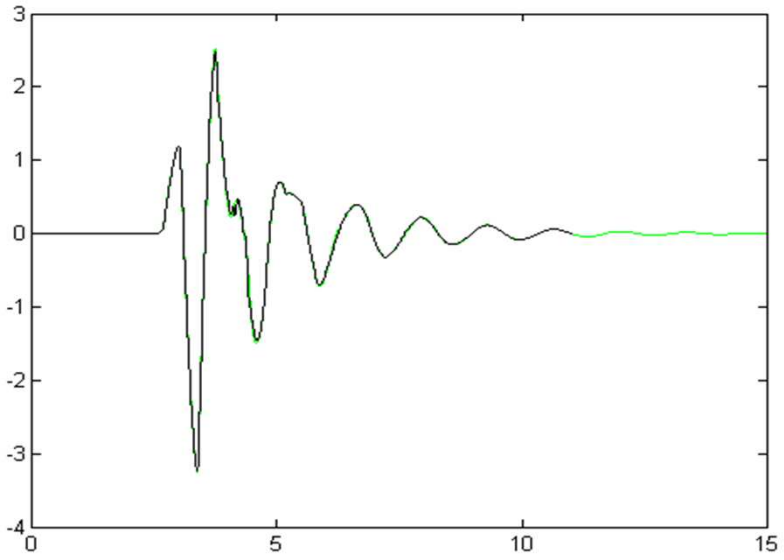
- Limited number of simulation (in 3D)

Variability of earthquakes ground motions at sedimentary basin scale (1D)

> Design of experiments
- 799 runs for 7 parameters
- synchronization of the zeros

> Estimation of one realization stochastic: MC and **PPC**





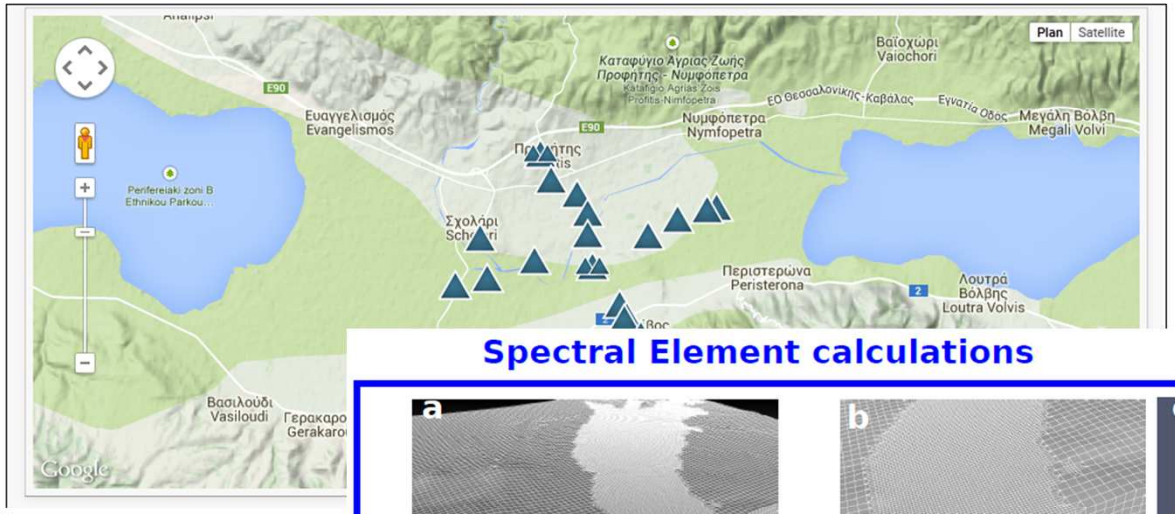
Variability of earthquakes ground motions at sedimentary basin scale (3D)

EUROSEISTEST DATABASE

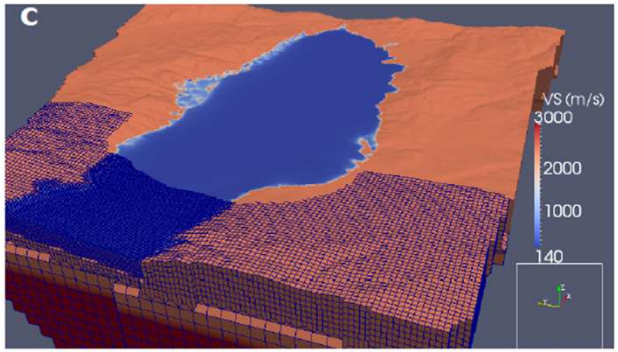
Bassin Mygdonien, Grèce

SIGN IN

HOME | THE EUROSEISTEST | DATABASE SEARCH | SFSI | ANNOUNCEMENTS



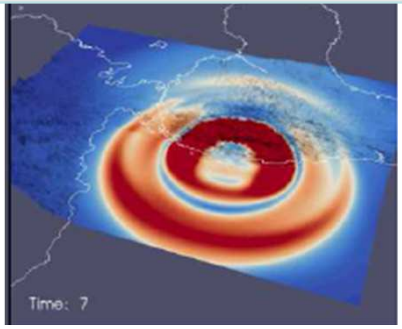
Spectral Element calculations

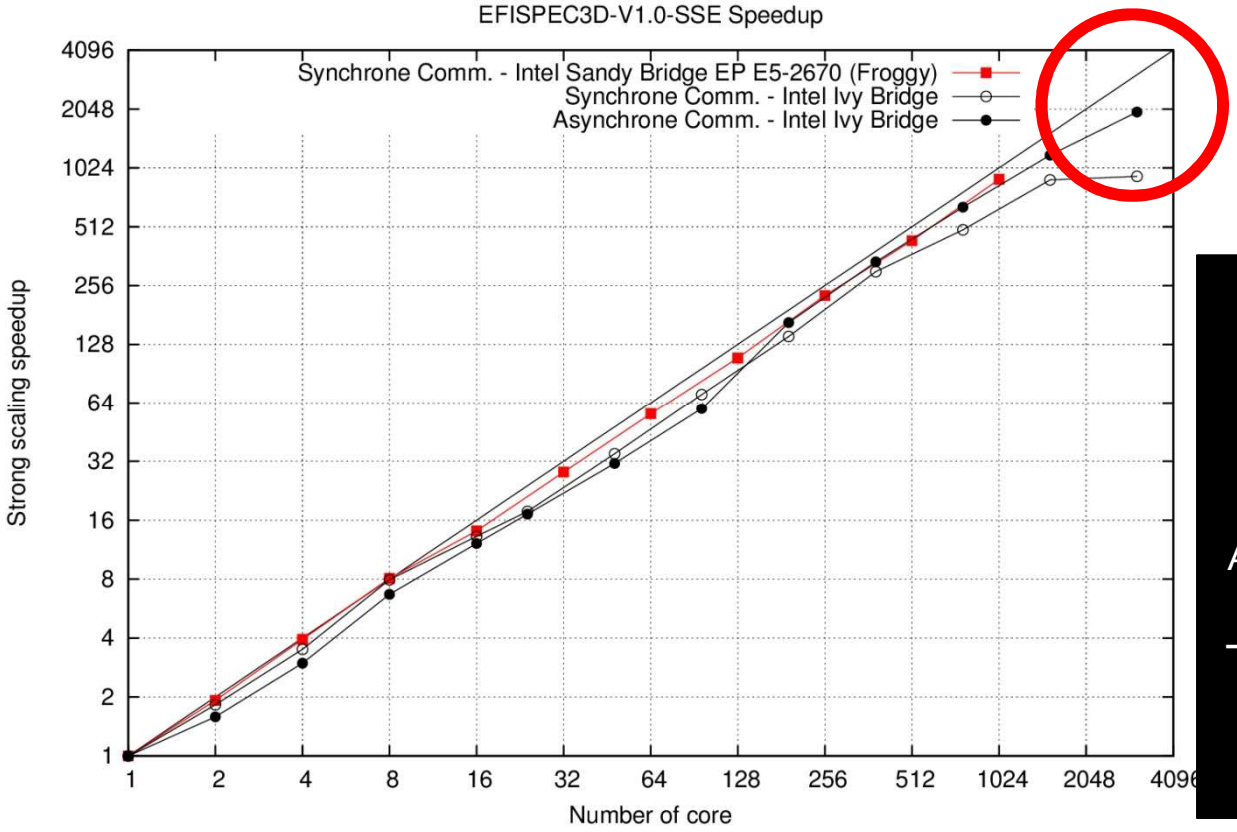


We designed a spectral element mesh of the Mygdonian basin which is refined both horizontally and vertically (see Figs. **a** and **b**). The mesh includes surface topography and the size of the elements is adapted to the *S* wavelengths in the basin (see Fig. **c**).
The size of the computational domain is 47 km x 64 km in EW and NS directions, and 30 km in the vertical direction. The mesh is made of 4.7 millions elements and contains 314 millions gridpoints.
The velocity model in the basin is defined by two linear gradients, one between the surface and the premygdonian-mygdonian limit and the second one down to the sediment-bedrock interface. The surface S wave velocity is 137 m/s and the grid allows accurate calculations for frequencies up to 4 Hz. .

Variability of earthquakes ground motions at sedimentary basin scale (3D)

Spectral elements





Break down of synchronous comm. above 1536 cores

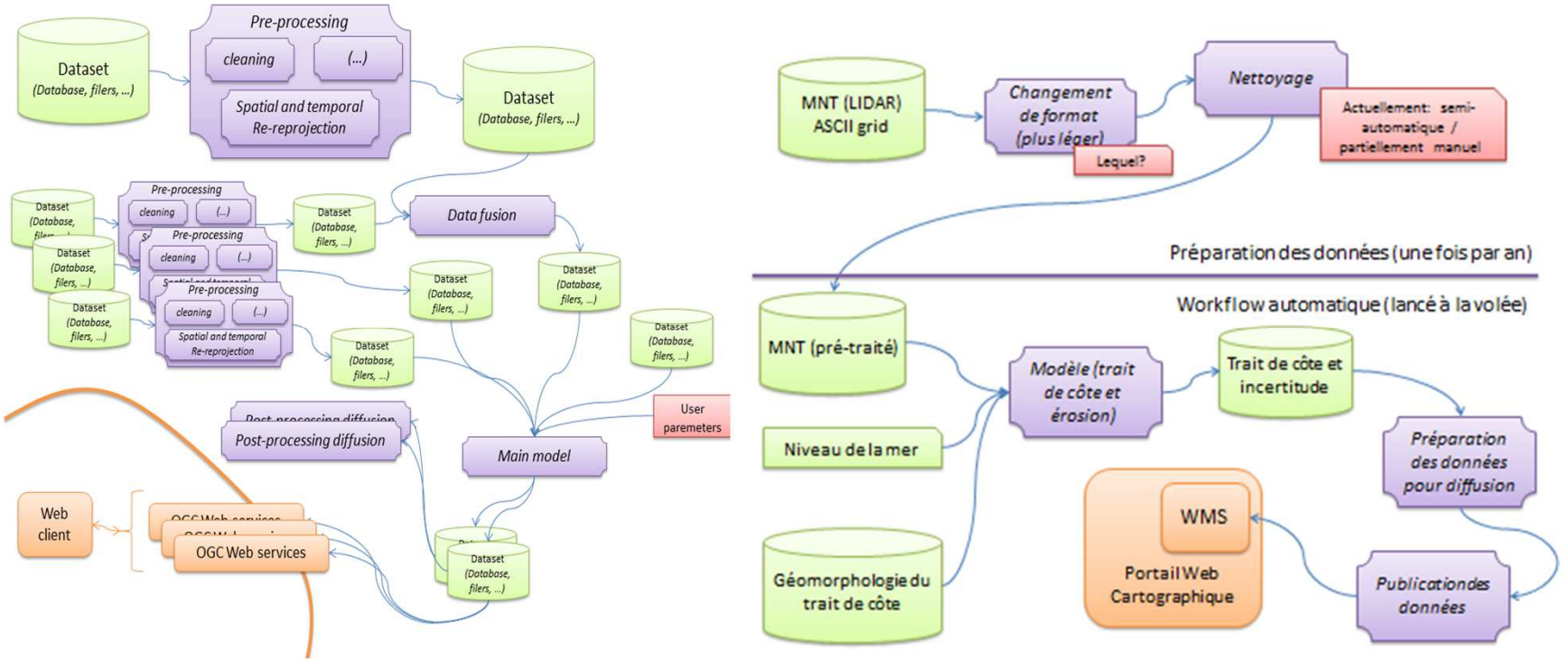
But

Async. comm. comes to rescue
→ More computations needed to reach async. comm. break down

Scientific workflow

Large Data sets

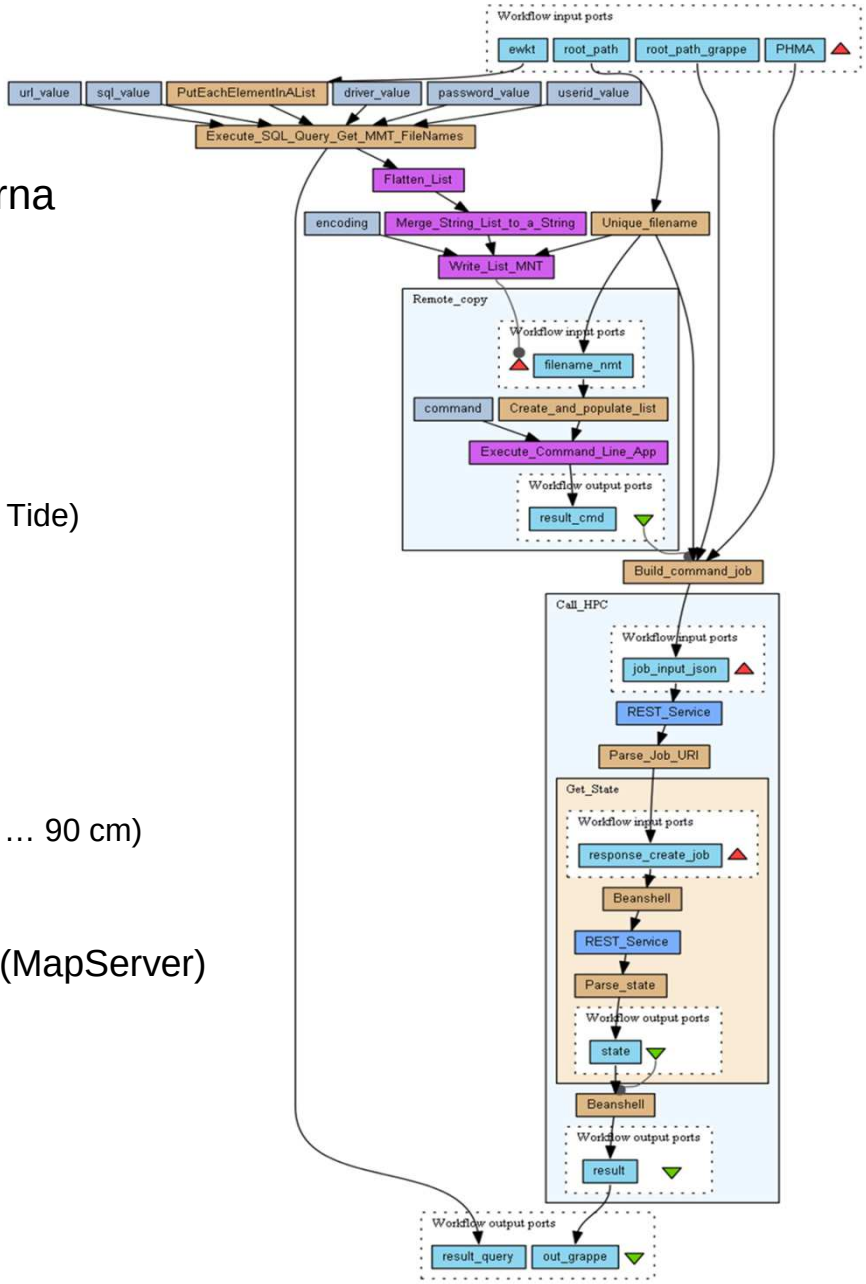
- LiDAR: Digital Elevation Model
  - → French coastal area: 160Go, 25000 files
  - → Several Terabytes for global map
- Data on tides and water levels (SHOM National Hydrographic Service)
- Population and land use (Medde, INSEE, EEA)
- Extrem events and coastal erosion (BRGM)



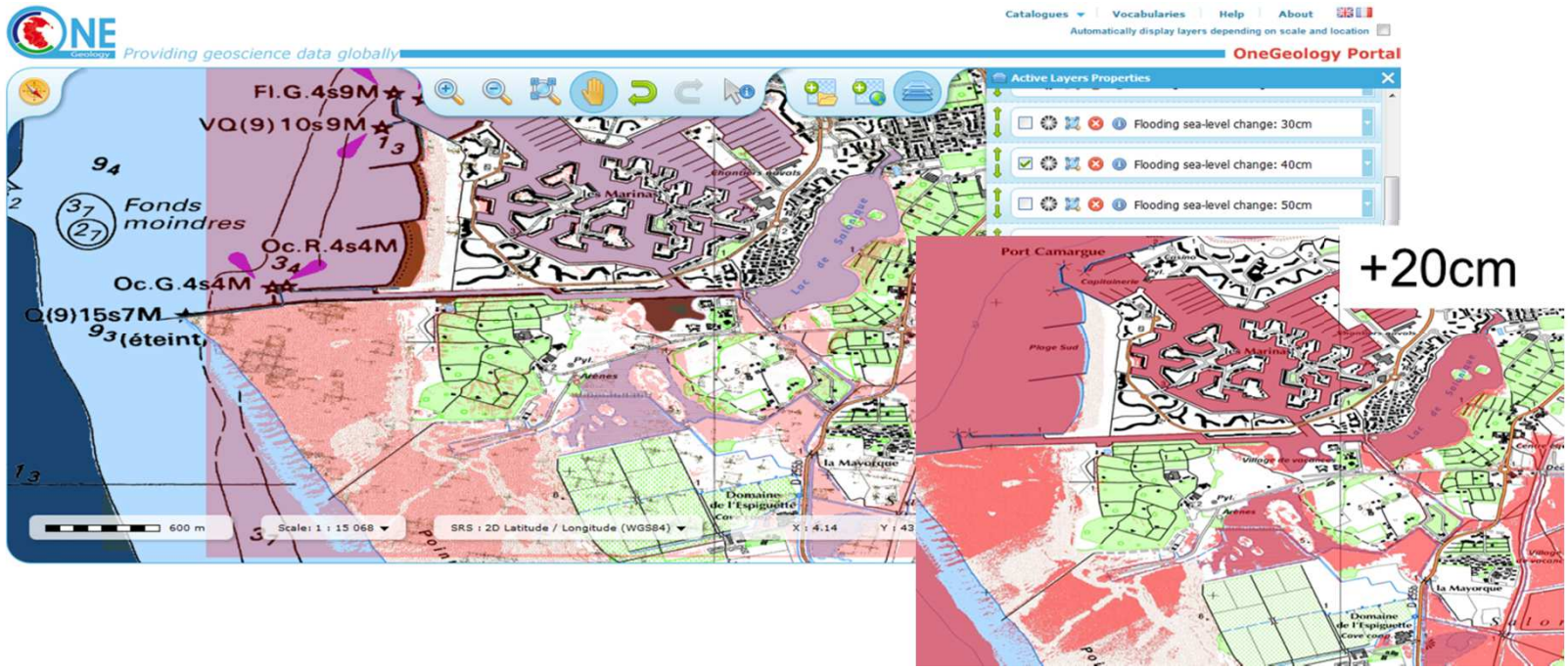
Scientific workflow

- Workflow management system → Taverna
  - Synchronous for general public
  - Asynchronous for experts
- On-the-fly processing for a small area
  - Selection of the Area of interest
  - Computation of the HAT (Highest Astronomical Tide)
  - Computation of Marine Extreme Level
  - Selection of grids to be used.
- For each
  - Preprocessing steps
  - Sea flooding modelling for 10 sea levels (10cm … 90 cm)
- Publication of WMS with Tile index (MapServer)
  - Generation of 10 shapefiles of the grids

20 seconds for a small area 15 km²



Scientific workflow



- Take full advantage of large scale processing facilities in order to speedup grid-based data analysis including multi-resolution strategies (**Rasdaman**)
- Increase the complexity of overflow model including PDE-based simulation tools (**Swash, Swan**).
- Extend this framework capabilities to tackle large environmental data collections (**EPOS, Copernicus ...**)
- Application to **Earthquake early warning → <u>exploiting GPU implementation</u>**

Conclusion

- **Harnessing high performance computing is critical for future challenges in Geosciences**
  - Large scale, 3D, coupled simulations associated with uncertainties analysis

- **Effort at several levels at BRGM**
  - Balance between internal efforts and external collaborations at the software level.
  - Maintain high level scientific collaborations

- **New challenges → *mostly related to data !***
  - Geoprocessing → High Performance Data
  - Uncertainties (non intrusive) → all aspects including visualization.
  - Moving code paradigm and complex high-performance workflows.

- **Happy with the collaboration with Brazil**
  - Ongoing work on I/O (Efispec3D software)
  - Co-advising the Ph.D of Victor Martinez

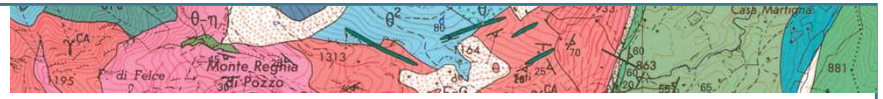
References

> Journal

- R.Keller Tesser, L.Pilla, F.Dupros, P.O.Navaux, J-F Mehaut, C.Mendes *: Dynamic load balancing for seismic wave propagation models IJHPCA (revision)*
- M.Castro, F.Dupros, E.Francesquini, J-F Méhaut, P.O.Navaux : Seismic Wave Propagation Simulations on Low-power and Performance-centric Manycores. *Parallel Computing (revision)*
- M.Diener, E.Cruz, L.Pilla, F.Dupros, P.O.Navaux : *Characterizing Communication and Page Usage of Parallel Applications for Thread and Data Mapping .* Performance Evaluation (2015)
- M.Castro, E.Francesquini, P-H Penna, F.Dupros, H.Freitas, J-F Méhaut, P.O.Navaux : *On the Energy Efficiency and Performance of Irregular Applications on Multicore, NUMA and Manycore* Processors : *JPDC (2014)*

> Conferences

- V.Martinez, F.Dupros, D.Michea, O.Aumage, S.Thibaut, P.O.A.Navaux : Towards seismic wave modeling on heterogeneous many-core architectures using task-based runtime system. SBAC-PAD (2015)
- F. Dupros, F.Boulahya, H.Aochi, P.Thierry : Communication-avoiding seismic numerical kernels on multicore processors: HPCC (2015)
- M.Castro, F.Dupros, E.Francesquini, J-F Méhaut, P.Navaux : Energy Efficient Seismic Wave Propagation Simulation on a Low-Power Manycore Processor. Proceedings of SBAC-PAD (2014) – Best Paper
- R.Keller Tesser, L.Pilla, F.Dupros, P.O.Navaux, J-F Mehaut, C.Mendes *: Improving the performance of seismic wave simulations with dynamic load balancing :* Proceedings of PDP (2013)
- Hajime Yamamoto et.al. : Numerical Simulation of Long-Term Fate of CO_2 Stored in Deep Reservoir Rocks on Massively Parallel Vector Supercomputer. Proceedings of VECPAR (2012)



Thanks for your attention

f.dupros@brgm.fr

Géosciences pour une Terre durable

brgm