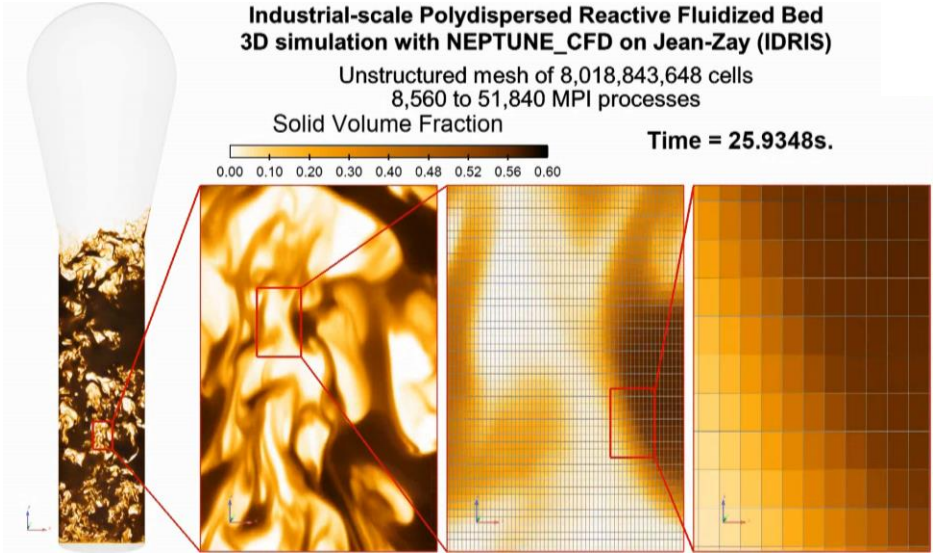
code_saturne

neptune_cfd

Mesh refinement and checkpoint interpolation to reach Exascale with neptune_cfd

HPC and computer codes:
a constantly evolving ecosystem transition



Hervé NEAU[1,2,7] - Maxime PIGOU[1,2,7]

Nicolas RENON[5,7] - Cyril BAUDRY[6] - Yvan FOURNIER[6] - Nicolas MERIGOUX[6]

Renaud ANSART[3,4,7] - Olivier SIMONIN[1,4,7]

❶ IMFT Institut de Mécanique des Fluides de Toulouse

❷ cnrs dépasser les frontières

❸ Laboratoire de Génie Chimique Toulouse • UMR 5503

❹ Toulouse INP

❺ calmip

❻ EDF R&D

❼ Université de Toulouse

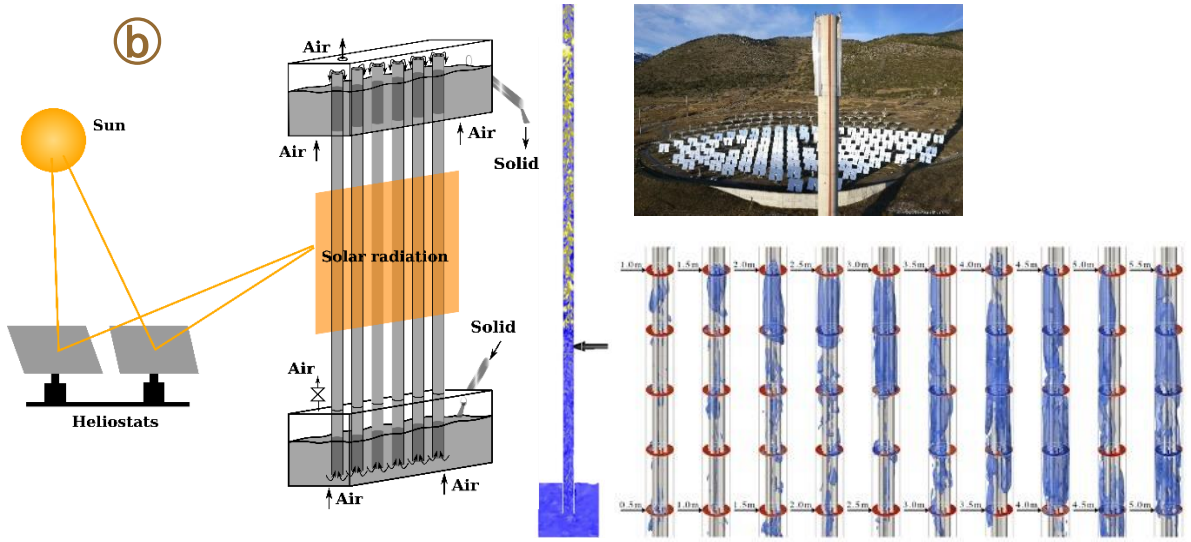
Research topic: modeling of fluid-particle reactive flows: IMFT / LGC

⇒ **Develop innovative processes and decarbonated applications for energy production**:

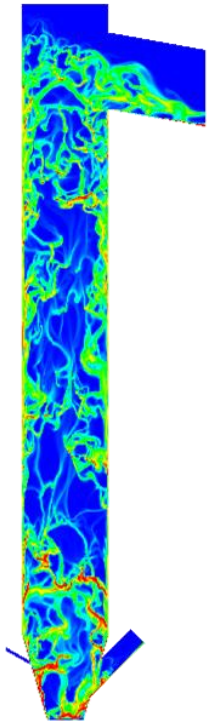
- ⓐ hydrogen combustion
- ⓑ fluidized bed solar receivers and heat storage on particles
- ⓒ biomass gasification plants
- ⓓ chemical looping combustion (coal/gas)
- ⓔ CO_2 mineralization process

ⓐ

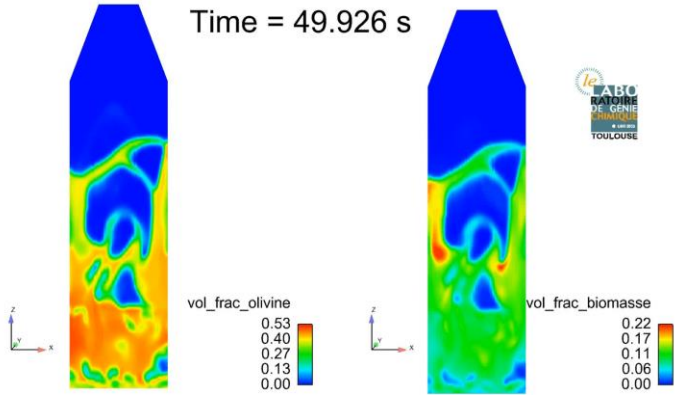
ⓑ



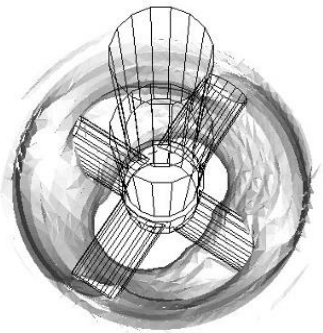
ⓒ



ⓓ



ⓔ



Breakthrough innovative technology for solar thermal power plants neptune_cfd HPC simulation on 6m height solar receiver

Experimental mock-up vs simulation of CO_2 mineralization process

Research on gas-particles reactive flows at Toulouse

A multi-scale numerical approach with experimental comparison

Microscopic scale
(~1*mm*) DNS

Lab/Industrial experimental measurements

Mesoscopic scale
(~10*cm*) DNS/LES+DPS

Macroscopic scale
~1m ⇨ ~30m
URANS/LES Euler-Euler

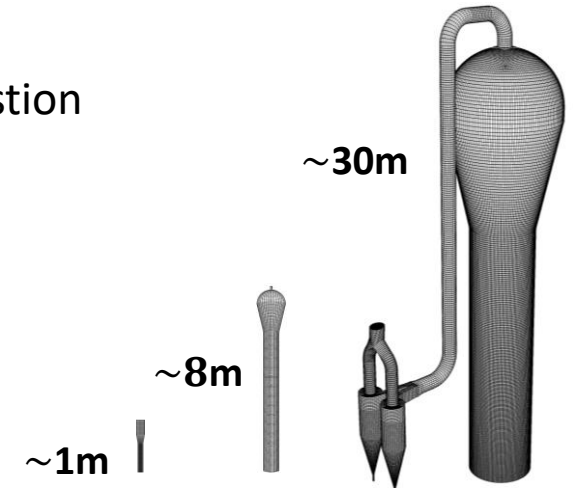
Scale-up

Industrial scale

Modeling challenges

- 3D unsteady flows, poly-dispersed solid mixture, gaseous or heterogeneous combustion
- Granular flow regime transition, non-spherical particles
- Radiative transfers, electrostatic effects

Simulation from laboratory scale up to industrial scale



neptune_cfd: Finite-volume Eulerian multiphase solver for 3D reactive turbulent flows

neptune_cfd*: proprietary solver build upon the open-source code_saturne HPC framework

- **Massively Parallel** code: C/C++, MPI, multigrid solver, parallel I/O
- Unsteady Multi-Fluid Modeling approach (N-Euler)
- **Numerical methods:** semi-implicit solver
  - Numerical schemes
    - Spatial discretization: 2nd-order unstructured finite-volume, centered scheme
    - Time integration: 1st-order fractional step method
  - **Unstructured meshing**
    - Non-matching meshes
    - Rotating meshes
    - **Any type of cell** (tetrahedral, hexahedral, prismatic, pyramidal, polyhedral...)
    - Co-located cell-centered finite volumes

code_saturne ⇨ http://code-saturne.org

neptune_cfd

* neptune_cfd is developed jointly by EDF and CEA with financial support of IRSN and FRAMATOME in the framework of the NEPTUNE project for nuclear applications

Why High Performance Computing?

Large, complex geometries of research and industrial projects ⇨ refined meshes

- **To design and improve industrial processes**
- **To better understand coupling of scales**

Optimal accuracy ⇔ highly refined meshes ⇔ HPC software/hardware

- Better prediction of mesoscale structures and improved description of bed hydrodynamics and transfers
- Computations limited by available resources and performances

Euler/Euler modeling approach for industrial-scale geometries
⇨ strong sensitivity to mesh size

Why High Performance Computing?

Large, complex geometries of research and industrial projects ⇨ refined meshes

- **To design and improve industrial processes**
- **To better understand coupling of scales**

Optimal accuracy ⇔ highly refined meshes ⇔ HPC

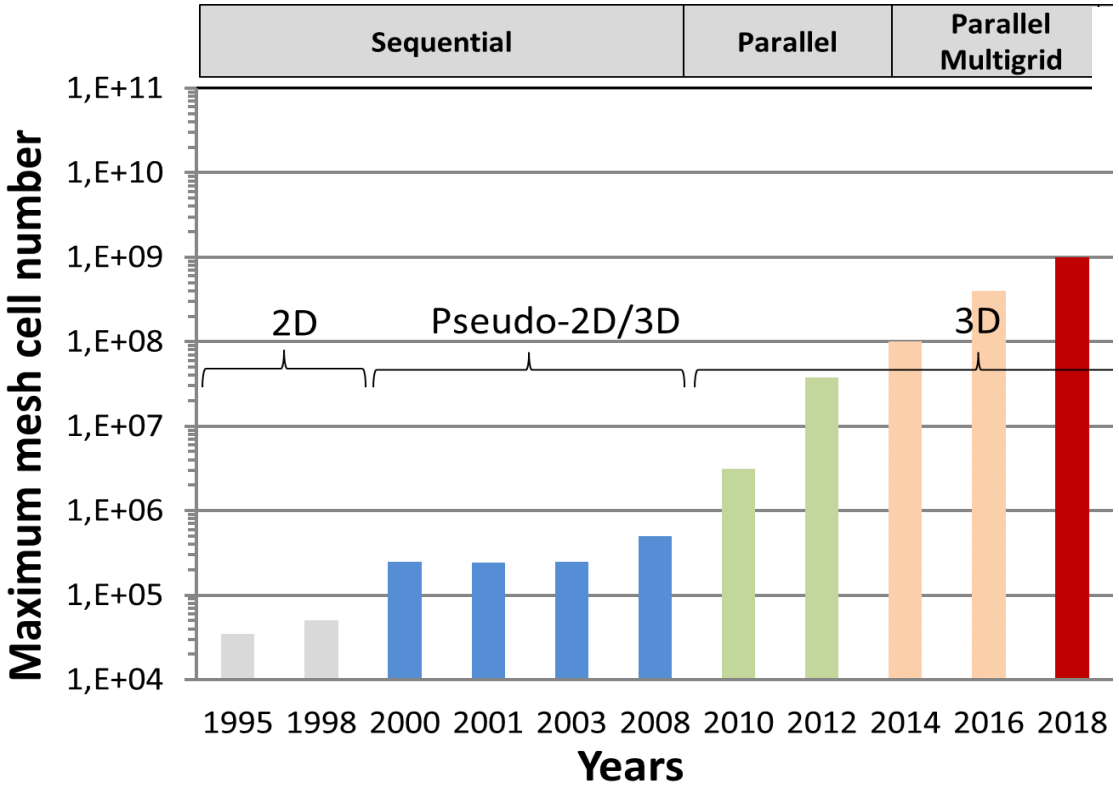
- Better prediction of mesoscale structures and improved description of bed hydrodynamics and transfers
- Computations limited by available resources and performances

Euler/Euler modeling approach for industrial-scale geometries ⇨ strong sensitivity to mesh size

Since 2008, handled mesh size increased by **x1,000**

⇒ same increase in **cost for simulating transient hydrodynamics:** **~50% of total computation time**

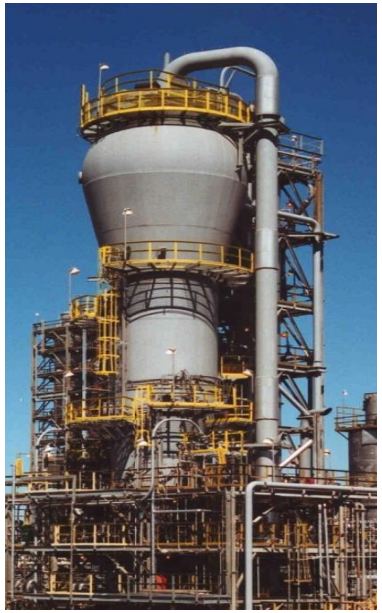
⇒ generally, only **"steady" results interesting:** temporal mean computation, temporal fluctuations, …

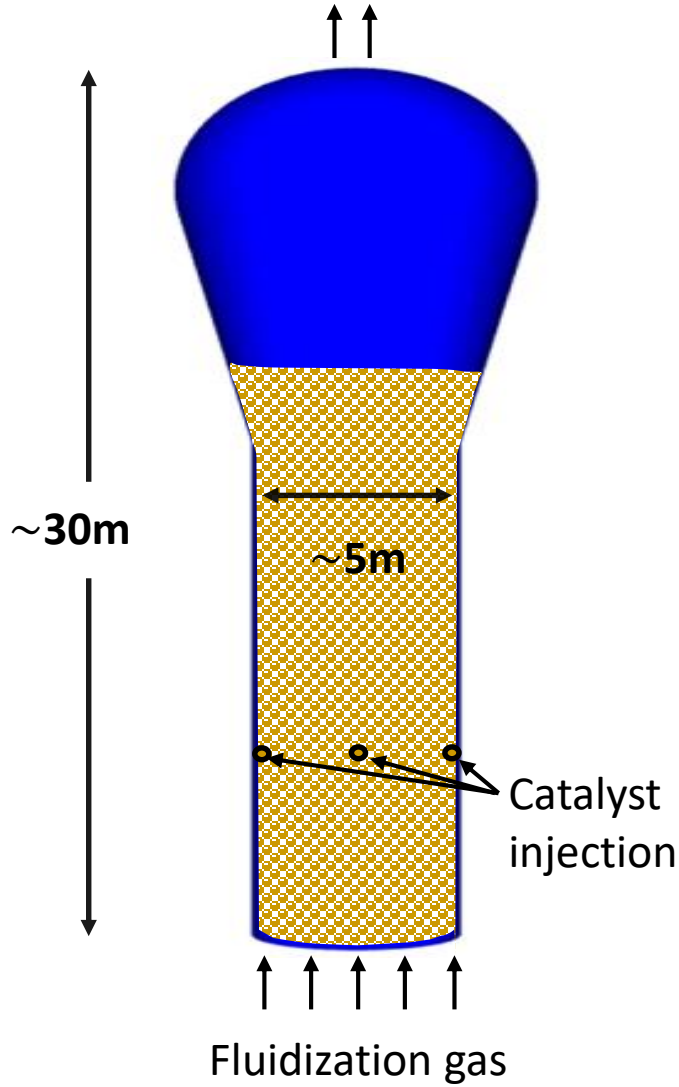


Representative industrial case: Reactive gas-solid fluidized bed polymerization reactor

- 2 kinds of particles: polyethylene (large) and catalyst (fine)
- Reactive: Polymerization exothermic reaction
  Gas/particle heat transfers
- Multi-scale unsteady turbulent flow
- T∼100°C - P∼24bars – 100 tons of particles – V∼$600m^3$

⇨ **21 coupled PDE to solve**

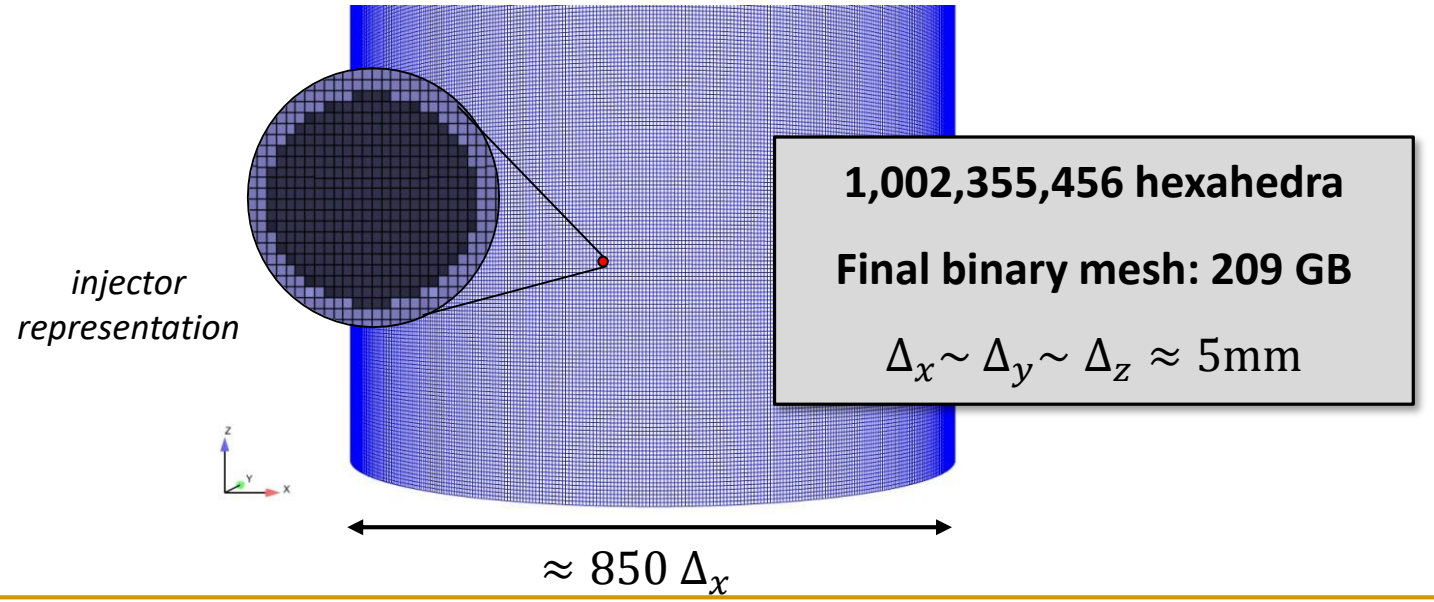


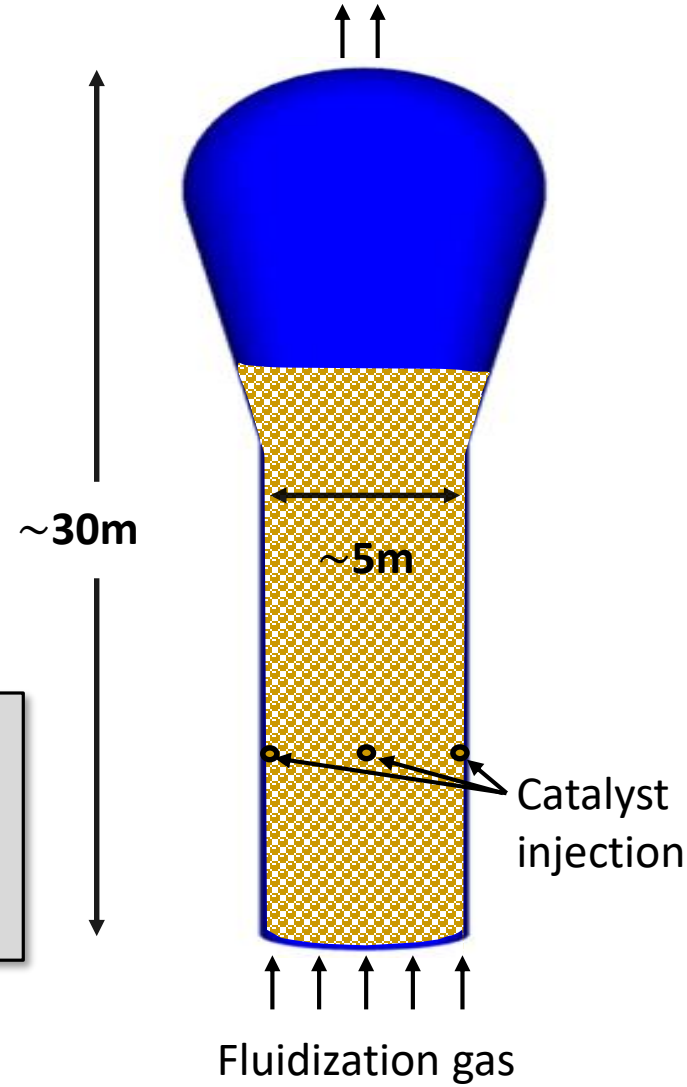


Representative industrial case: Reactive gas-solid fluidized bed polymerization reactor

- 2 kinds of particles: polyethylene (large) and catalyst (fine)
- Reactive: Polymerization exothermic reaction
  Gas/particle heat transfers
- Multi-scale unsteady turbulent flow
- T∼100°C - P∼24bars – 100 tons of particles – V∼$600m^3$

1 billion cells mesh construction (Simail, code_saturne)





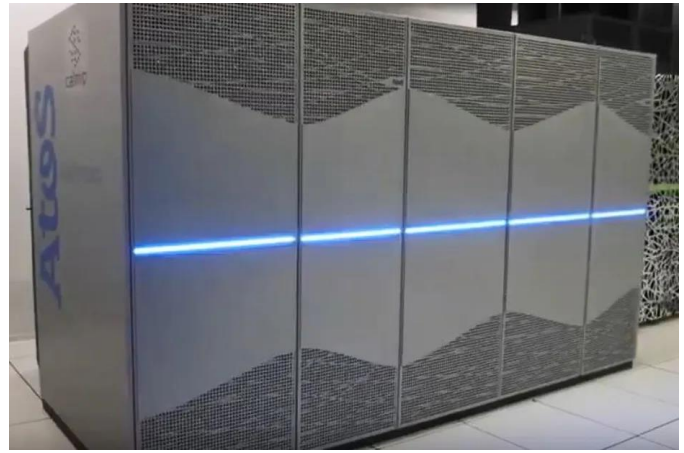
2018: Meso- / Grands-Challenges at CALMIP and EDF computing centers

⇨ **an opportunity to use whole supercomputers during a "short" dedicated period**

CALMIP supercomputer: Olympe

Atos Bull SEQUANA X1000 cluster - Perf. Peak: 1.37 Pflop/s
13,392 cores (2.3GHz) - Infiniband EDR (100 Gb/s) - Lustre

CPU nodes	Cores / node – Processor	RAM / node
360 bi socket	2x18 - Intel® Xeon® Gold Skylake 6140	192 GB



EDF R&D supercomputer: Gaia

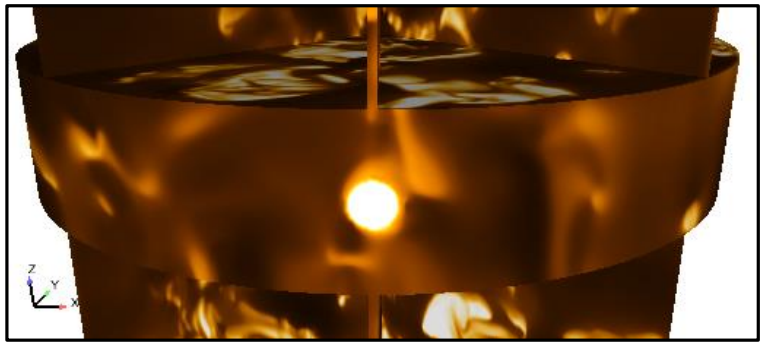
Atos Bull Cluster - Perf. Peak: 3.05 Pflop/s
42,912 cores (2.3GHz) - Intel OPA v1 - GPFS

CPU nodes	Cores / node – Processor	RAM / node
1,192 bi socket	2x18 - Intel® Xeon® Gold Skylake 6140	192 GB

Principal metrics of our initial computation

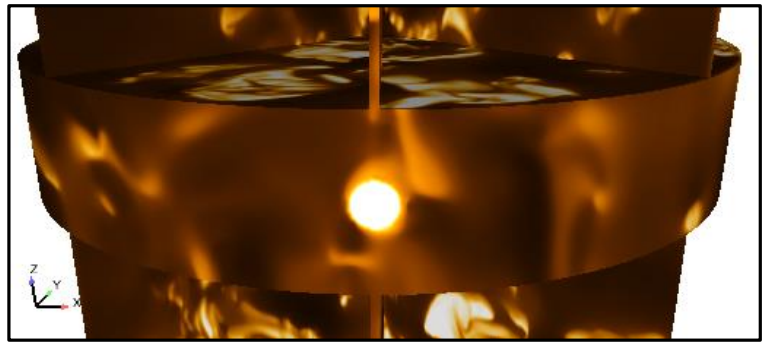
- Multiple runs from 35 nodes (1,260 cores) up to **1,000 nodes** (36,000 cores)
- **15 million CPU hours ⇔ 25 s of physical time** ⇔ elapsed time: 30 days

- Checkpoint Restart file: **1.3 TB** – Overall data volume: ~**200 TB**
- Mesh, partitioning and checkpoint **reading time: 13 minutes / run**



Principal metrics of our initial computation

- Multiple runs from 35 nodes (1,260 cores) up to **1,000 nodes** (36,000 cores)
- **15 million CPU hours ⇔ 25 s of physical time** ⇔ elapsed time: 30 days

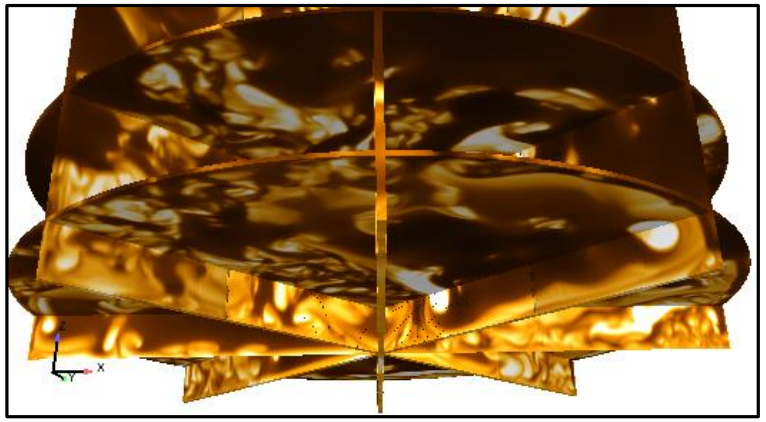


- Checkpoint Restart file: **1.3 TB** – Overall data volume: ~**200 TB**
- Mesh, partitioning and checkpoint **reading time: 13 minutes / run**

Results Data base: ~10 TB by saving 16 variables

⇨ 1 full volume data file: 0.5 TB

⇨ EnSight Gold binary data files for 660 time steps
on 12 selected thick planes, cylinders and external surfaces



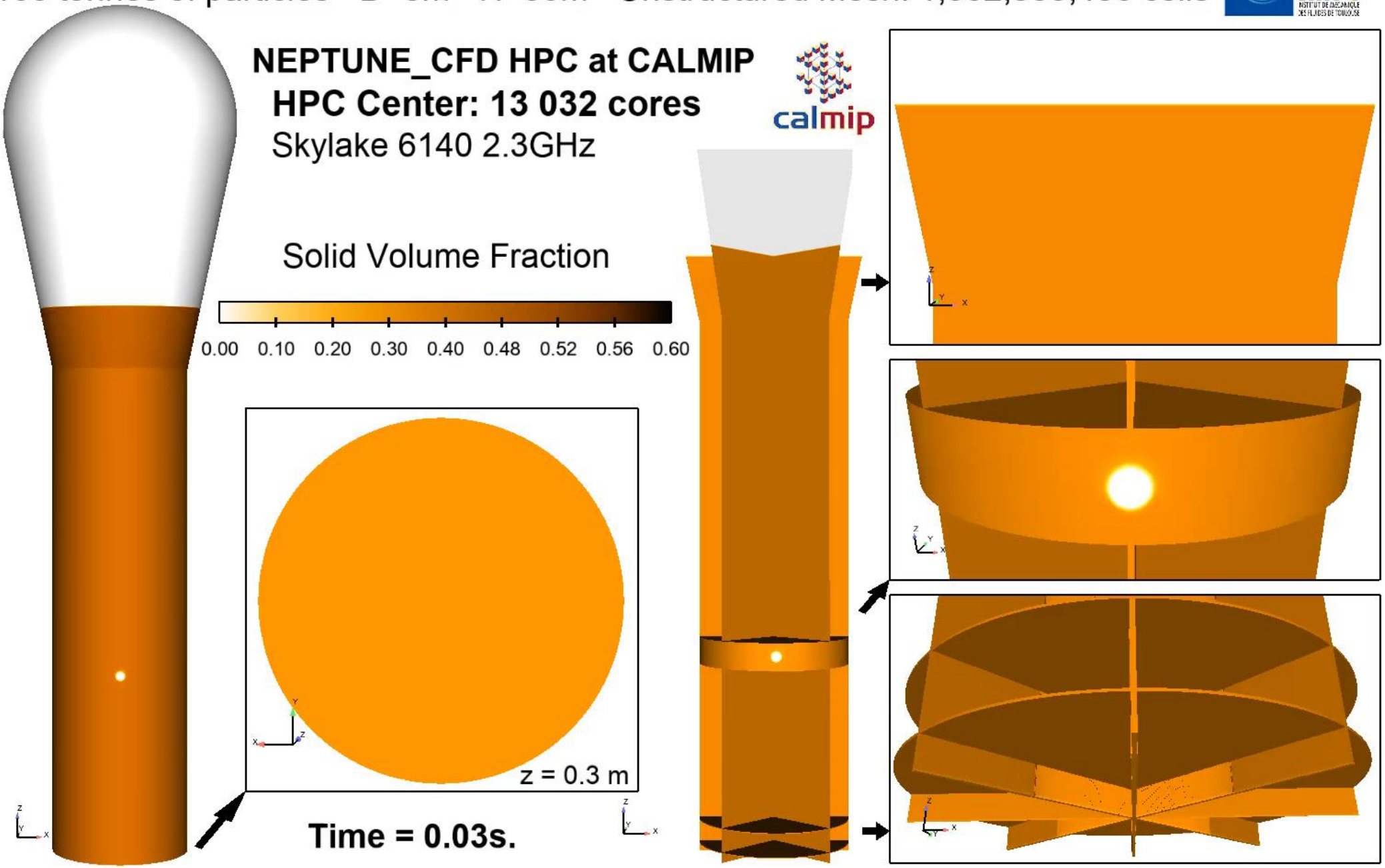
Industrial Scale Bidispersed Reactive Fluidized Bed Reactor

100 tonnes of particles - D~5m - H~30m - Unstructured Mesh: 1,002,355,456 cells

NEPTUNE_CFD HPC at CALMIP

HPC Center: 13 032 cores

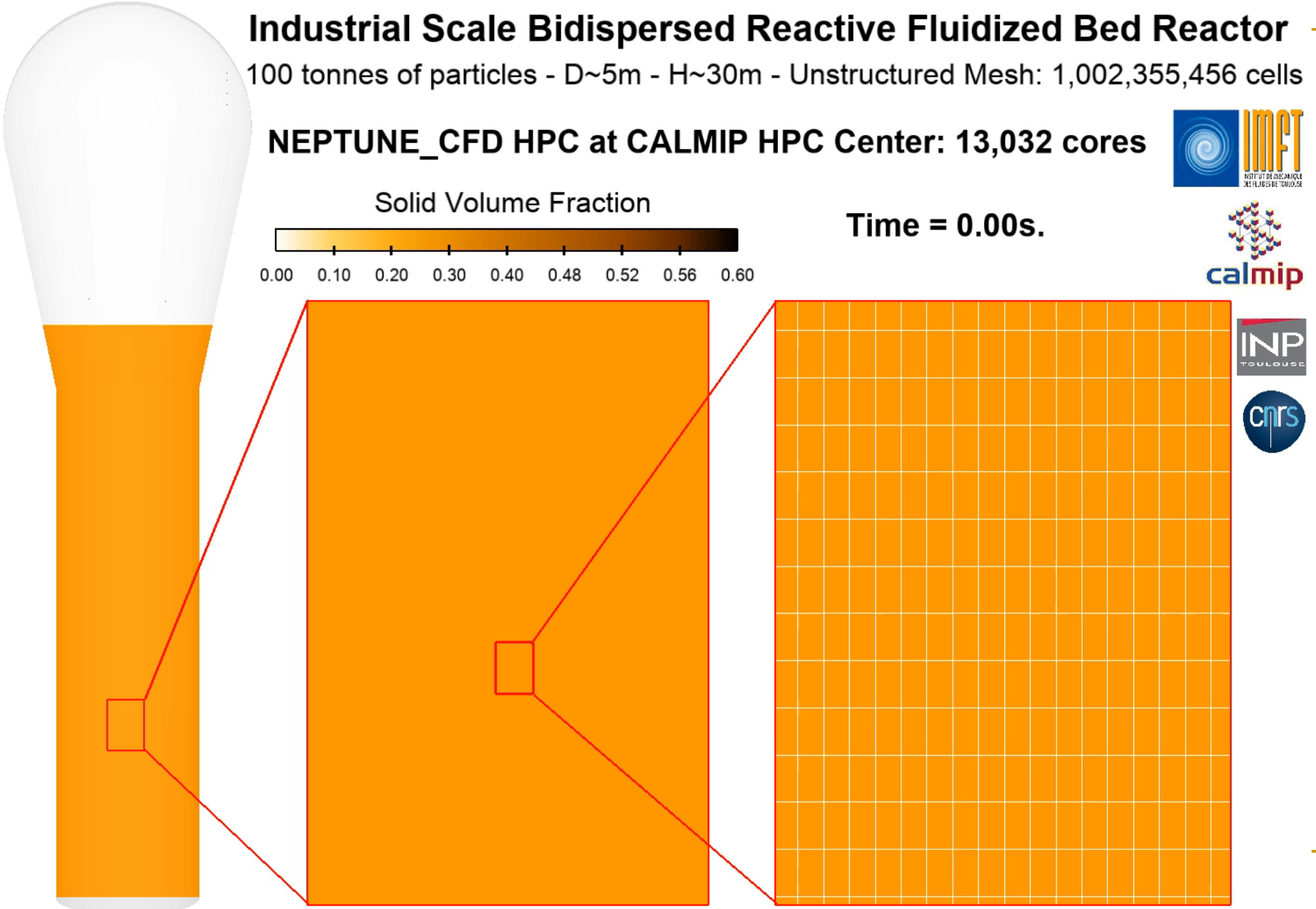
Skylake 6140 2.3GHz



Industrial Scale Bidispersed Reactive Fluidized Bed Reactor

100 tonnes of particles - D~5m - H~30m - Unstructured Mesh: 1,002,355,456 cells

NEPTUNE_CFD HPC at CALMIP HPC Center: 13,032 cores

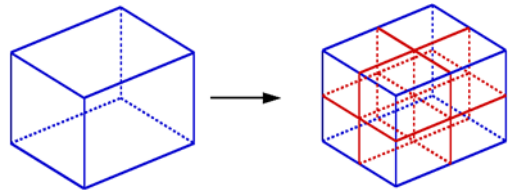


Next step: refining from 1 to 8 billion cells

- **Need to refine further:**
  - cell size still large compare to particle clusters
  - evaluate HPC capabilities at larger scales (Tiers 1 – IDRIS)

- **Creation of a new mesh:**
  - from scratch (simail + code_saturne)

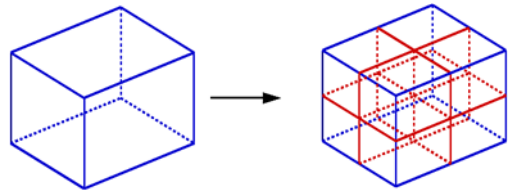
1 hexa ⇨ 8 hexa



Next step: refining from 1 to 8 billion cells

- **Need to refine further:**
  - cell size still large compare to particle clusters
  - evaluate HPC capabilities at larger scales (Tiers 1 – IDRIS)

1 hexa ⇨ 8 hexa

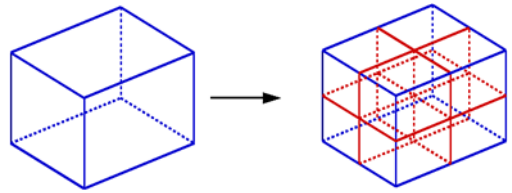


- **Creation of a new mesh:**
  - ~~from scratch (simail + code_saturne)~~
  - **automatically split by 2 in each direction** the 1 billion cell mesh using **code_saturne features**

Next step: refining from 1 to 8 billion cells

- **Need to refine further:**
  - cell size still large compare to particle clusters
  - evaluate HPC capabilities at larger scales (Tiers 1 – IDRIS)

1 hexa ⇨ 8 hexa



- **Creation of a new mesh:**
  - ~~from scratch (simail + code_saturne)~~
  - **automatically split by 2 in each direction** the 1 billion cell mesh using **code_saturne features**

Problem: number of cells => x8, adaptive time step => twice/thrice smaller with gas acceleration

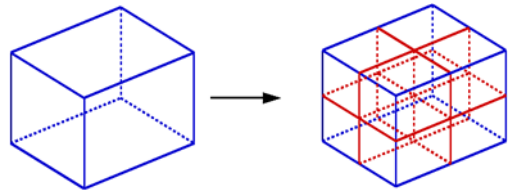
⇒ Simulation time: x20

⇒ **Significant increase in computation cost**

Next step: refining from 1 to 8 billion cells

- **Need to refine further:**
  - cell size still large compare to particle clusters
  - evaluate HPC capabilities at larger scales (Tiers 1 – IDRIS)

1 hexa ⇨ 8 hexa



- **Creation of a new mesh:**
  - ~~from scratch (simail + code_saturne)~~
  - **automatically split by 2 in each direction** the 1 billion cell mesh using **code_saturne features**

Problem: number of cells => x8, adaptive time step => twice/thrice smaller with gas acceleration

⇒ Simulation time: x20

⇒ Significant increase in computation cost

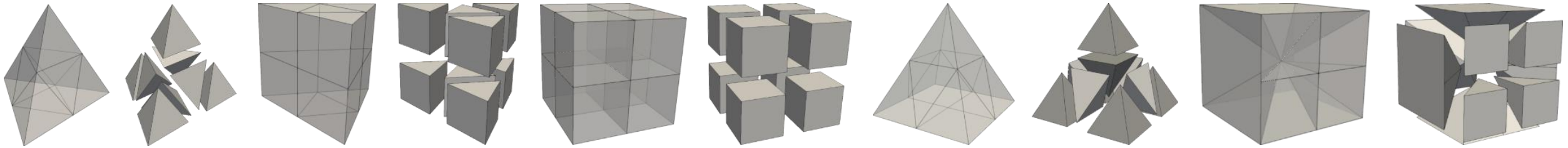
⇒ Savings **reusing results from the 1 billion-cells** mesh as initial conditions **for the 8 billion cells simulation**

⇒ **Mesh to mesh interpolation**

Automatic mesh refinement ⇨ 1 to 8 billion cells

Use code_saturne's built-in mesh refinement (pre/post-processing) feature

- **Works on all element types**, using reference patterns where possible, generic ones otherwise



- **Based on user-selected cells** ⇨ Our case: all cells
- **Fully distributed algorithm**
  - Mostly local
  - Some many-to-many data movement (domain boundaries synchronization, possible mesh repartitioning)
  - Limited only by memory usage and the few MPI data movement
- Works with code_saturne's internal mesh structure
  - May be **saved to internal checkpoint format**

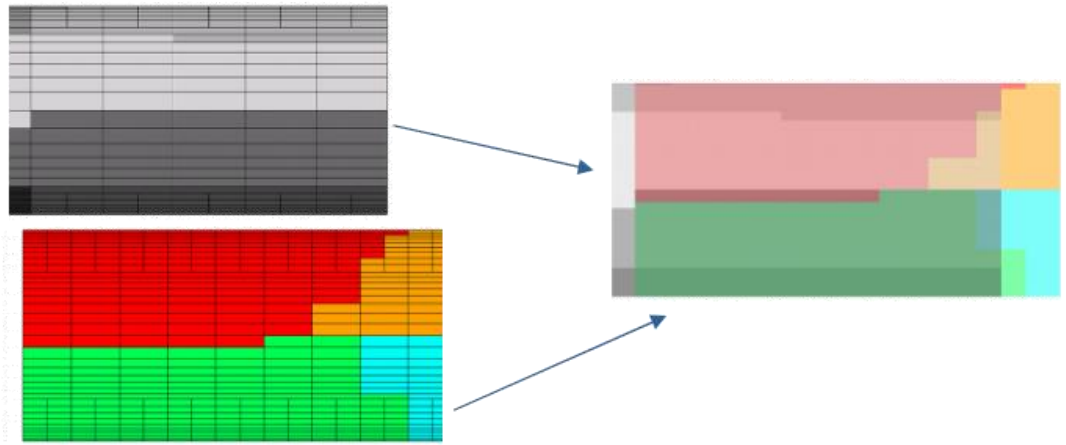
8,002,355,456 hexahedra
3,008,918,880 faces
1,004,210,878 nodes
Final binary mesh: 1.6 TB

code_saturne feature: restart simulation with different mesh

- Parallel Location and Exchange (PLE) library

  ⇨ Handles multi-programs and/or multi-domains MPI communications

  ⇨ Detects partitioned-domains overlap of both coarse and fine meshes



⇨ **Read “coarse mesh” checkpoint file, project it to fine mesh**

- **Cell-centered values of fine-cells based on value of matching coarse cell**
- **Simple Laplacian smoothing step** applied before actual time stepping start

⇒ Generated restart file: 9.8 TB

See https://doi.org/10.1080/10618562.2020.1810676 or https://hal.archives-ouvertes.fr/hal-02494687

Application during IDRIS Grand-Challenge: 1 to 8 billion cells

IDRIS supercomputer: Jean-Zay

HPE SGI 8600 - Perf. Peak: 16 Pflop/s

61,120 cores (2.5GHz) – Intel OPA (100 Gb/s) – IBM spectrum scale

CPU nodes	Cores / node – Processor - GPU	RAM / node
1,528 bi socket	2x20 - Intel® Cascade Lake 6248	192 GB/n

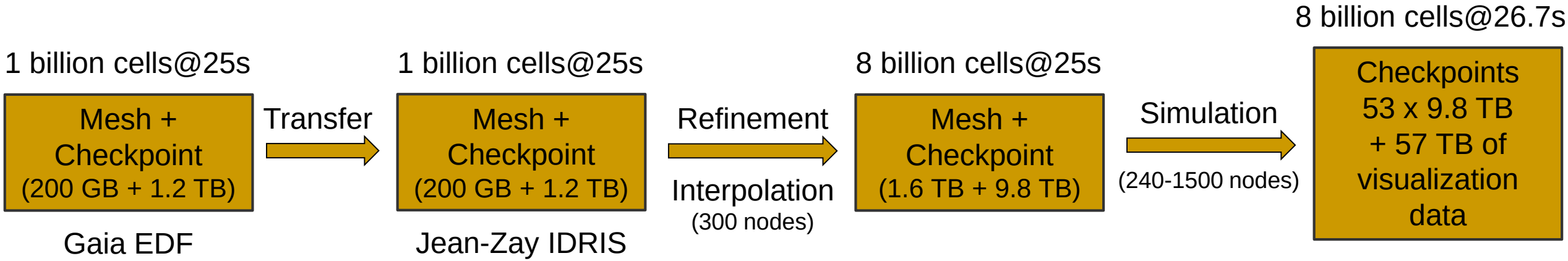
Application during IDRIS Grand-Challenge: 1 to 8 billion cells

IDRIS supercomputer: Jean-Zay

HPE SGI 8600 - Perf. Peak: 16 Pflop/s

61,120 cores (2.5 GHz) – Intel OPA (100 Gb/s) – IBM spectrum scale

CPU nodes	Cores / node – Processor - GPU	RAM / node
1,528 bi socket	2x20 - Intel® Cascade Lake 6248	192 GB/n



- **31 M CPU h ⇔ 1.7 s of physical time**
- Significantly long computation (1000 h) with “short” walltime (20 h) ⇨ 80 h spent in checkpoint I/O
- **HPC assessments:** speed-up, efficiency, sensibility studies (IO, process CPU binding)

Industrial-scale Polydispersed Reactive Fluidized Bed
3D simulation with NEPTUNE_CFD on Jean-Zay (IDRIS)

Unstructured mesh of 8,018,843,648 cells
8,560 to 51,840 MPI processes

Time = 25.0006s.

Solid Volume Fraction

0.00 0.10 0.20 0.30 0.40 0.48 0.52 0.56 0.60

z=16m

z=11m

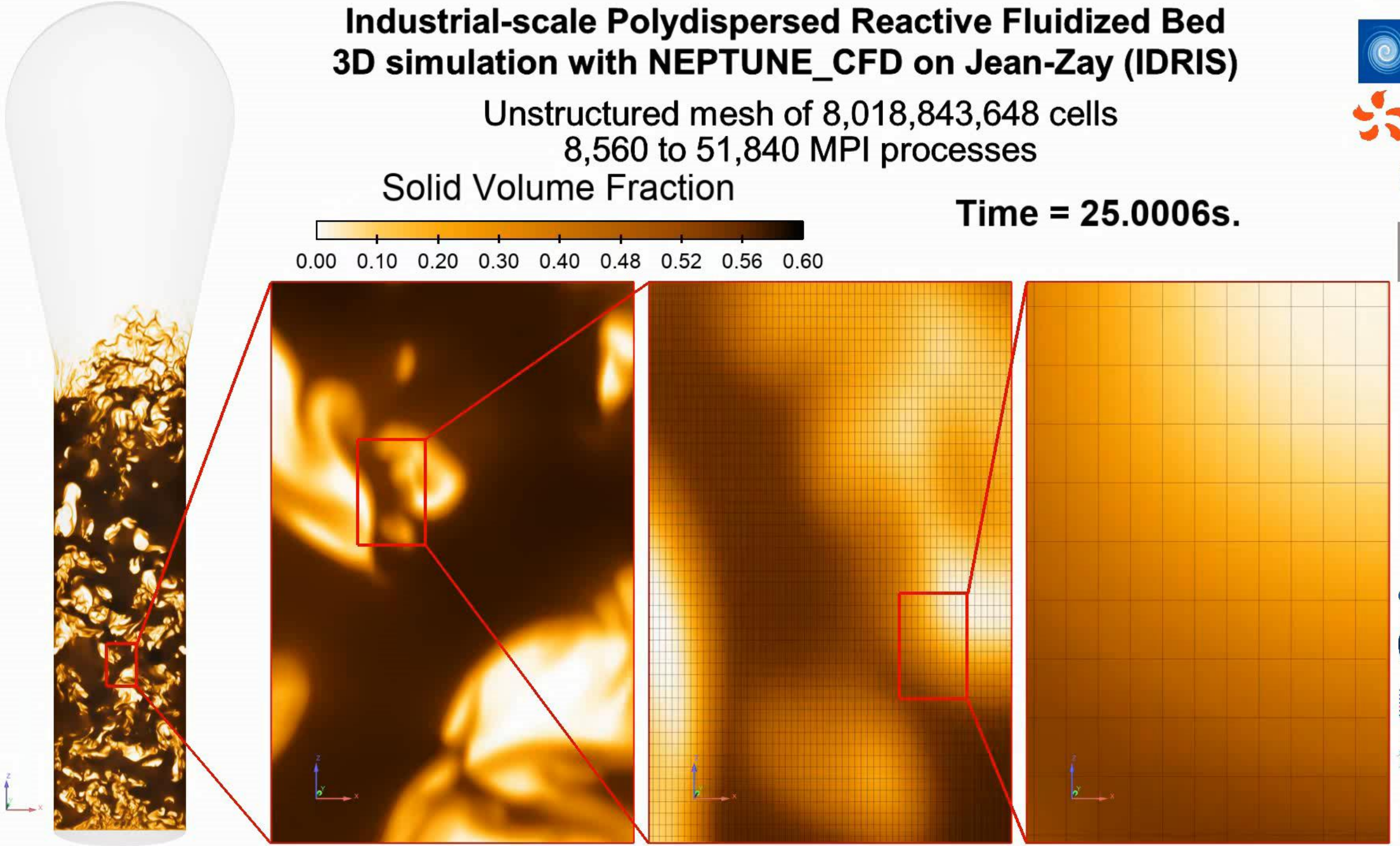
z=6.7m

z=0.3m

Herve Neau
Maxime Pigou

Industrial-scale Polydispersed Reactive Fluidized Bed
3D simulation with NEPTUNE_CFD on Jean-Zay (IDRIS)

Unstructured mesh of 8,018,843,648 cells
8,560 to 51,840 MPI processes



H. Neau
M. Pigou

Another opportunity: Grand-Challenge at TGCC

TGCC supercomputer: Joliot-Curie Irene-AMD

Bull Sequana XH2000 - Perf. Peak: 11.75 Pflop/s

293,376 cores (2.6 GHz) – Infiniband HDR100 (100 Gb/s) – Lustre

CPU nodes	Cores / node – Processor - GPU	RAM / node
2,292 CPU nodes	2x64 c/n – **AMD Rome** (Epyc) 7H12	256 GB/n

⇨ Assessing AMD Rome architecture

⇨ Going further up to 64 billion cells

Another opportunity: Grand-Challenge at TGCC

TGCC supercomputer: Joliot-Curie Irene-AMD

Bull Sequana XH2000 - Perf. Peak: 11.75 Pflop/s

293,376 cores (2.6 GHz) – Infiniband HDR100 (100 Gb/s) – Lustre

CPU nodes	Cores / node – Processor - GPU	RAM / node
2,292 CPU nodes	2x64 c/n – **AMD Rome** (Epyc) 7H12	256 GB/n

8 billion cells@26.7s

Mesh + Checkpoint (1.6 TB + 9.8 TB)

Jean-Zay IDRIS

Transfer

8 billion cells@26.7s

Mesh + Checkpoint (1.6 TB + 9.8 TB)

Irene-AMD TGCC

Refinement (400 nodes)

Interpolation (2000 nodes)

64 billion cells@26.7s

Mesh + Checkpoint (12 TB + 80 TB)

Simulation

(1800-2250 nodes)

Profiling and performance data only

Another opportunity: Grand-Challenge at TGCC

TGCC supercomputer: Joliot-Curie Irene-AMD

Bull Sequana XH2000 - Perf. Peak: 11.75 Pflop/s

293,376 cores (2.6 GHz) – Infiniband HDR100 (100 Gb/s) – Lustre

CPU nodes	Cores / node – Processor - GPU	RAM / node
2,292 CPU nodes	2x64 c/n – **AMD Rome** (Epyc) 7H12	256 GB/n

8 billion cells@26.7s

Mesh + Checkpoint (1.6 TB + 9.8 TB)

Jean-Zay IDRIS

Transfer

8 billion cells@26.7s

Mesh + Checkpoint (1.6 TB + 9.8 TB)

Irene-AMD TGCC

Refinement (400 nodes)

Interpolation (2000 nodes)

64 billion cells@26.7s

Mesh + Checkpoint (12 TB + 80 TB)

Simulation (1800-2250 nodes)

Profiling and performance data only

Standard interpolation algorithm too slow (PLE algorithm complexity dependent on partitioning)

Here, **only hexahedra** are needed, and **all cells** are refined

⇨ Simple static mapping from original to refined mesh is possible in this case

⇨ **Allows for specialized, much simpler and faster approach**

Another opportunity: Grand-Challenge at TGCC

8 billion cells@26.7s

Mesh + Checkpoint (1.6 TB + 9.8 TB)

Jean-Zay IDRIS

Transfer

8 billion cells@26.7s

Mesh + Checkpoint (1.6 TB + 9.8 TB)

Irene-AMD TGCC

Refinement (400 nodes)

Interpolation (2000 nodes)

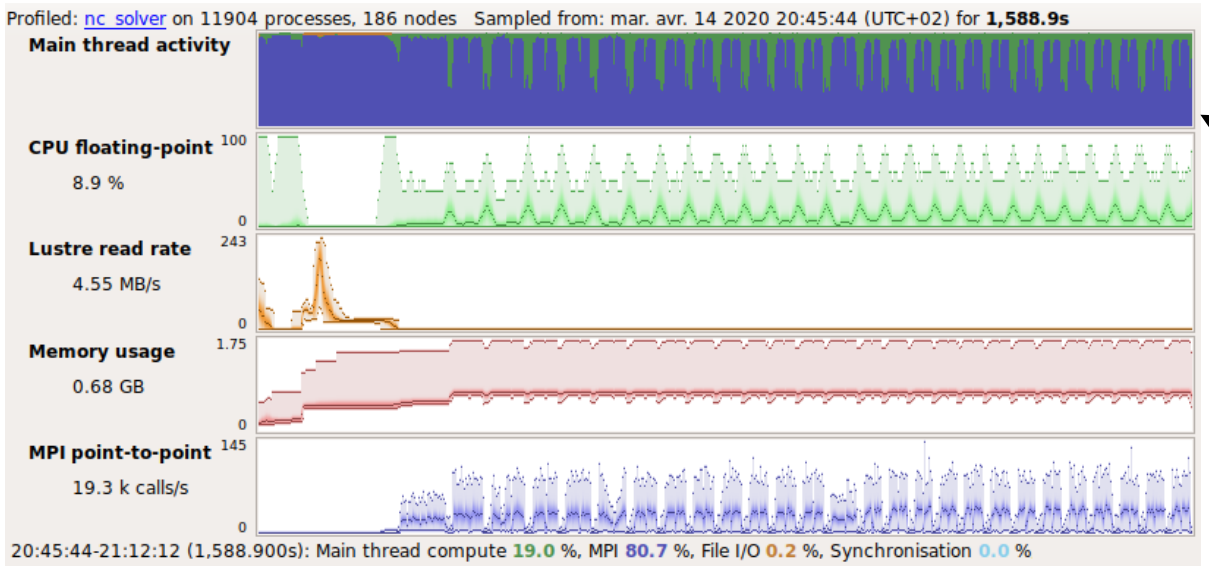
64 billion cells@26.7s

Mesh + Checkpoint (12 TB + 80 TB)

Simulation

(1800-2250 nodes)

Profiling and performance data only

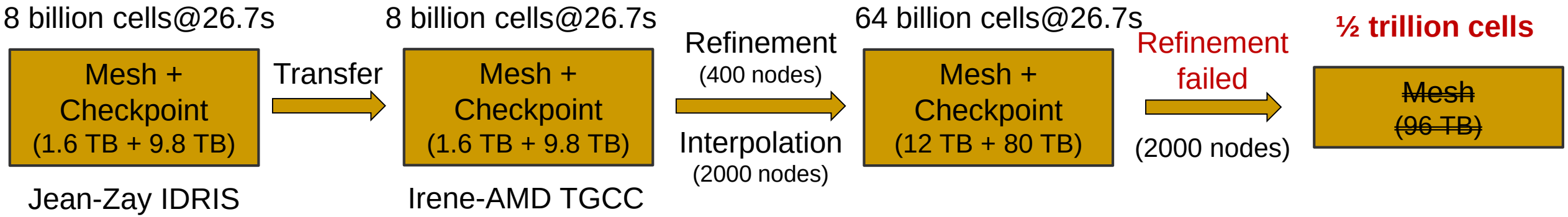


ARM MAP Profiling

Many bottlenecks at this scale:

- Whole supercomputer availability
- MPI_init and MPI_finalize slow and unstable (~1 h)
- Dominating MPI communications (>80% CPU time)
- Unmanageable file sizes
- …

Another opportunity: Grand-Challenge at TGCC



- **8 billion cells** ⇨ **limits reached for production**
- **64 billion cells** ⇨ **limits reached for simulations**
- **512 billion cells** ⇨ **limits reached for refinement procedure:**
  - Faced issue of 32 bits index overflow
  - Interpolation not feasible
  - Theoretical checkpoint size ~640 TB
- **Difficult data management even at 8 billion cells:** storage, transfers, analysis…

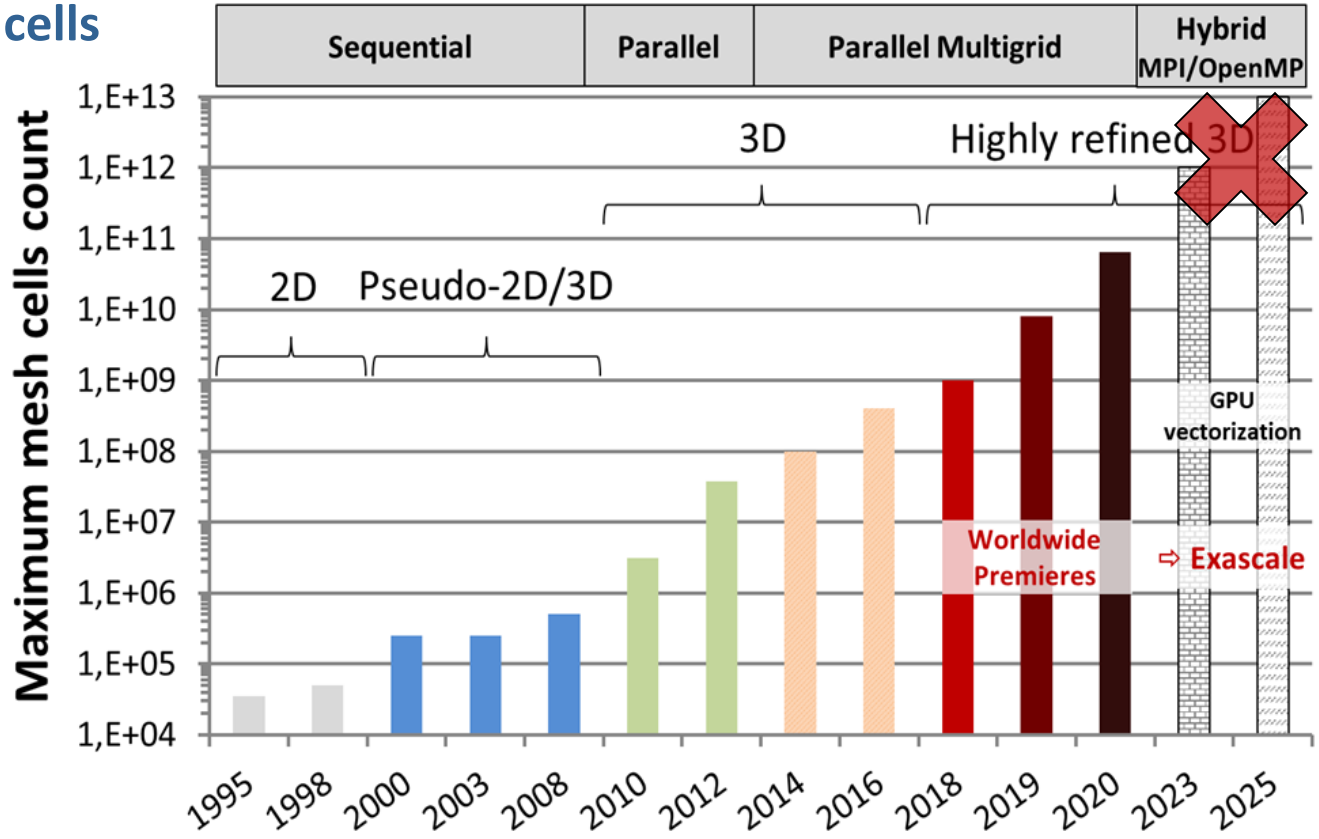
Refinement / interpolation to go beyond billion cells

- **Exponential growth of simulated case size**
- **Maximum mesh size for preprocessor ⇨ 2018**

⇨ Parallel mesh splitting approach to go beyond

⇨ Solver adaptation required at these scales:

- slow simulation (small Δt), huge CPU time
- saturated MPI communications
- availability of whole supercomputer



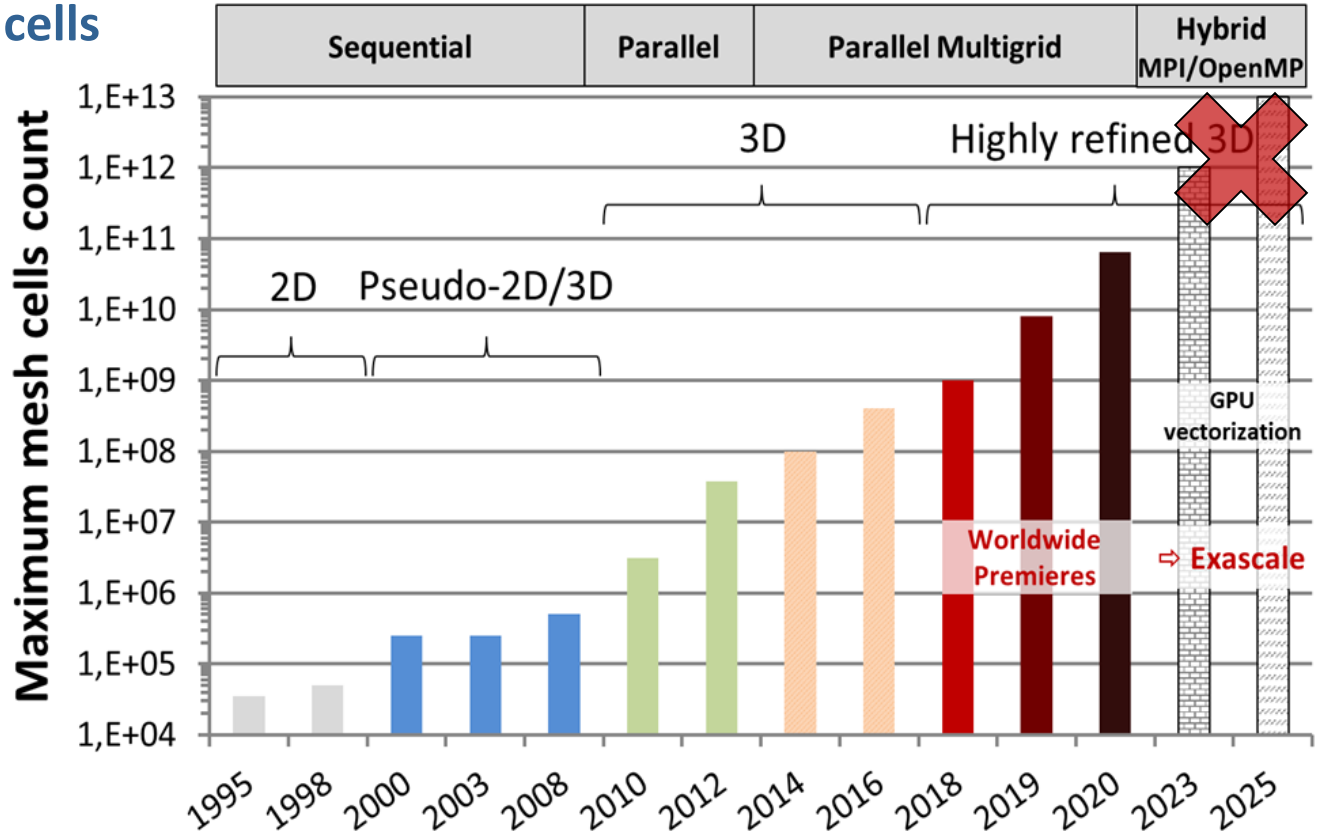
Refinement / interpolation to go beyond billion cells

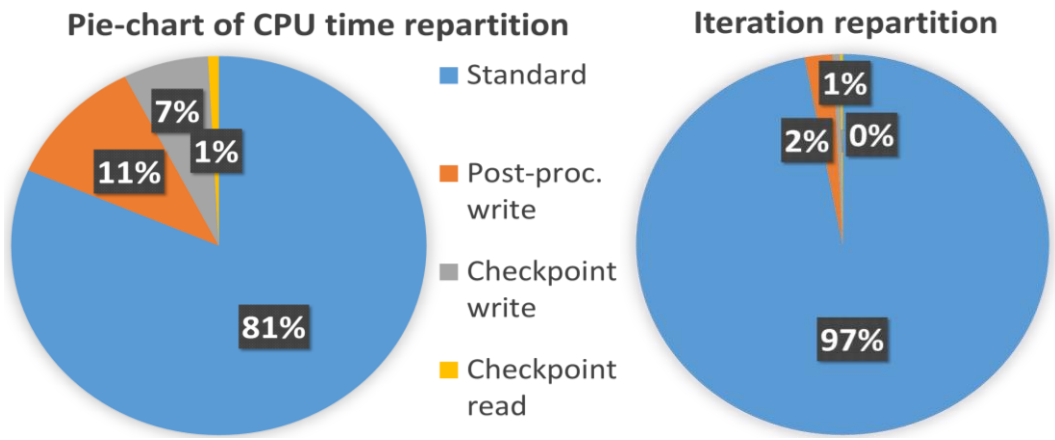
- **Exponential growth of simulated case size**
- **Maximum mesh size for preprocessor ⇨ 2018**

⇨ Parallel mesh splitting approach to go beyond

⇨ Solver adaptation required at these scales:

- slow simulation (small Δt), huge CPU time
- saturated MPI communications
- availability of whole supercomputer





- **Most significant limitations >1,000 nodes: I/O, MPI**
  - ⇨ **Frequent checkpoint write:** hardware failure, wall time
  - ⇨ **Fault Tolerance in future MPI norm**
  - ⇨ **Longer submission queue (QoS)**

Efficient strategy

- **neptune_cfd considered technologies to access Exascale:**
  - Hybrid MPI/OpenMP parallelization
  - GPU acceleration
  - Asynchronous MPI and I/O operations
  - Co-processing ⇨ already implemented

Efficient strategy

- **neptune_cfd considered technologies to access Exascale:**
  - Hybrid MPI/OpenMP parallelization
  - GPU acceleration
  - Asynchronous MPI and I/O operations

Presented mesh refinement / checkpoint interpolation strategy tested from 1 to 64 billion cells

⇨ **saving hundreds of millions of CPU hour for transient stage**

⇨ **limits “quickly” reached beyond tens of billions of cells:**
software, MPI library and hardware limits reached

Grands Challenges computations == Frontiers computations >>> usual research ones

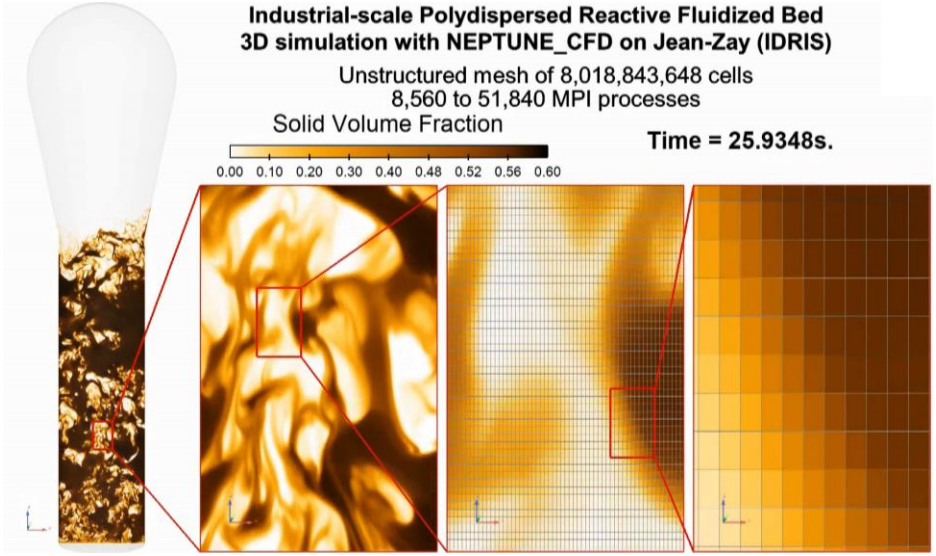
⇨ **easy and efficient strategy for production computations to reduce the cost of transient stage**

code_saturne

neptune_cfd

Thanks to your attention

HPC and computer codes:
a constantly evolving ecosystem transition



IMFT INSTITUT DE MÉCANIQUE DES FLUIDES DE TOULOUSE

calmip

cnrs dépasser les frontières

LABORATOIRE DE GÉNIE CHIMIQUE TOULOUSE • UMR 5503

TOULOUSE INP

Université de Toulouse

EDF R&D