Multicore/manycore with a scientific computing culture: experience feedback with special focus on performance modeling and redesign of computational methods

Florian De Vuyst

CMLA, ENS Cachan Université Paris-Saclay et CNRS

37è Forum ORAP – 17 mars 2016, CNRS Michel-Ange, Paris

Acknowledgements

The team :

- Marie Béchereau (PhD candidate, CMLA)
- Thibault Gasc (PhD candidate, MDLS, CEA, CMLA)
- Mathieu Peybernes (CEA Saclay, DEN)
- Raphael Poncet (CGG)
- Renaud Motte (CEA DAM DIF)

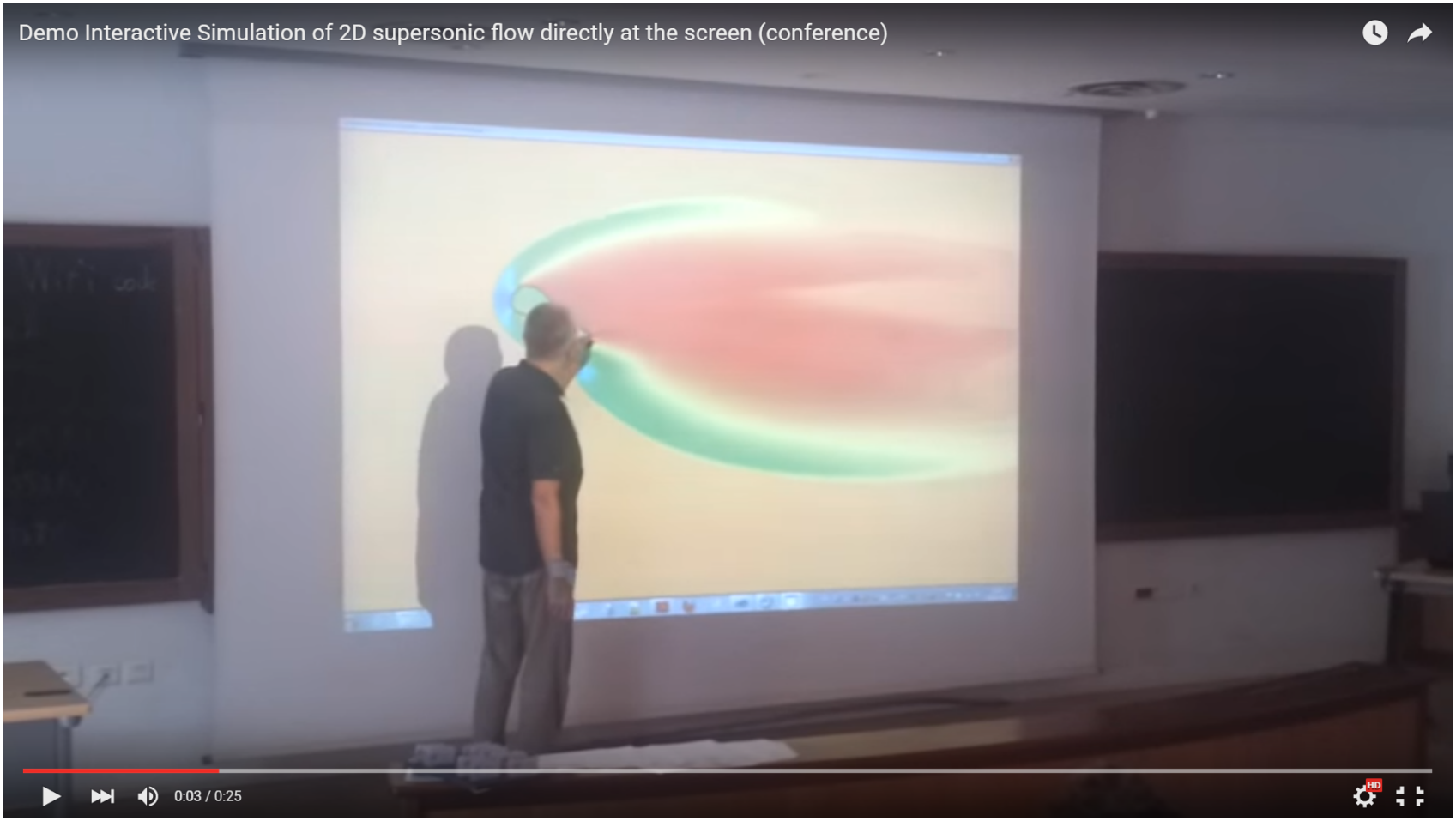
Acknowledgements :

- Thomas Guillet, Philippe Thierry (INTEL)
- NVIDIA (CUDA Research Center, 2013)
- MDLS for hardware support

1. GPU computing experience

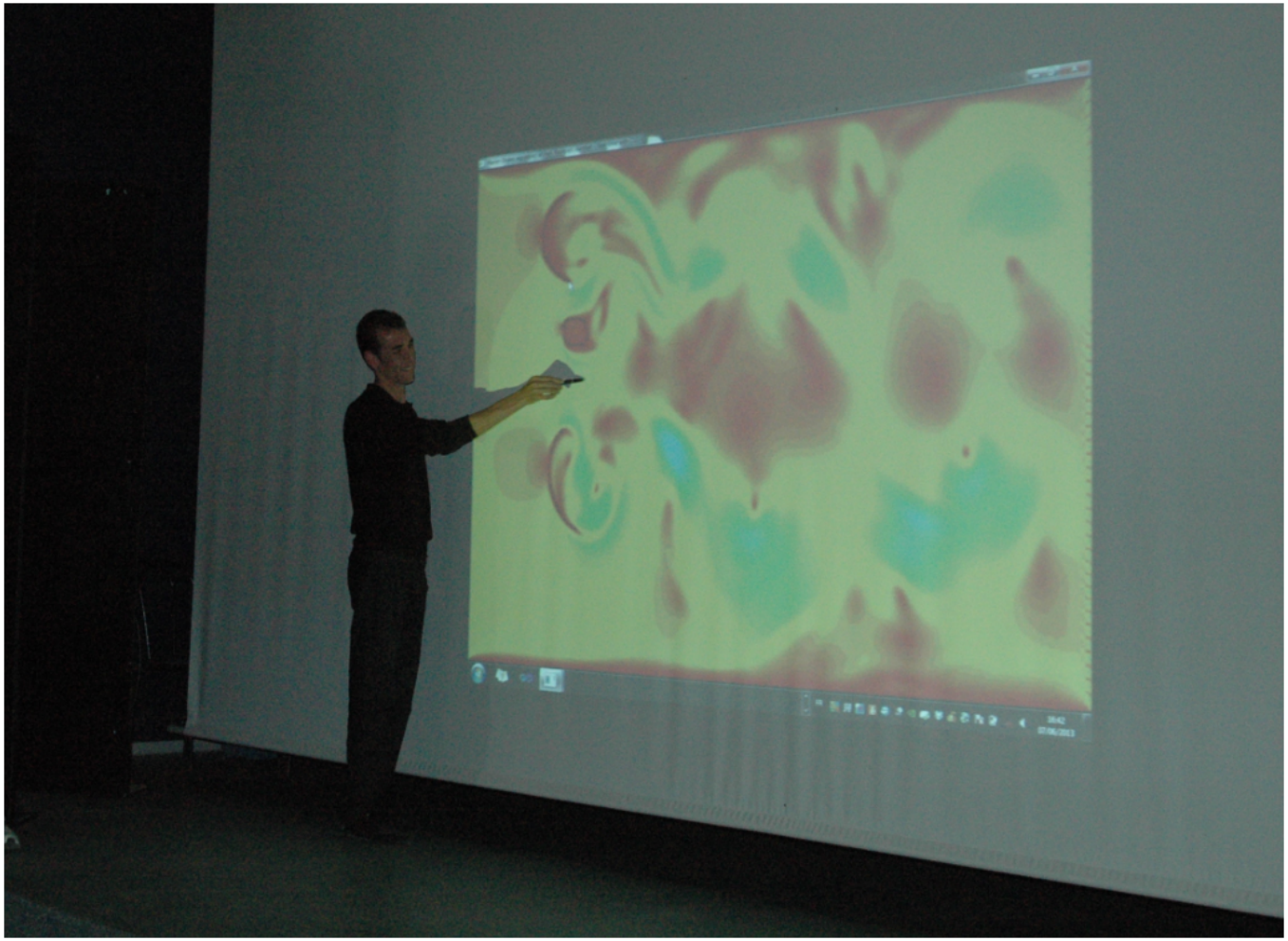
Beginning of the story ...

Real-time interaction of 2D flow dynamics on boards/screens (2012)



Beginning of the story ...

Real-time interaction of 2D flow dynamics on boards/screens (2012)



GPU computing experience

- NVIDIA CUDA Language, TESLA boards
- Subtantial speedup observed
- Nice binding between computation and visualization

- Classical 2D FV CFD Cartesian code & LBM codes implemented

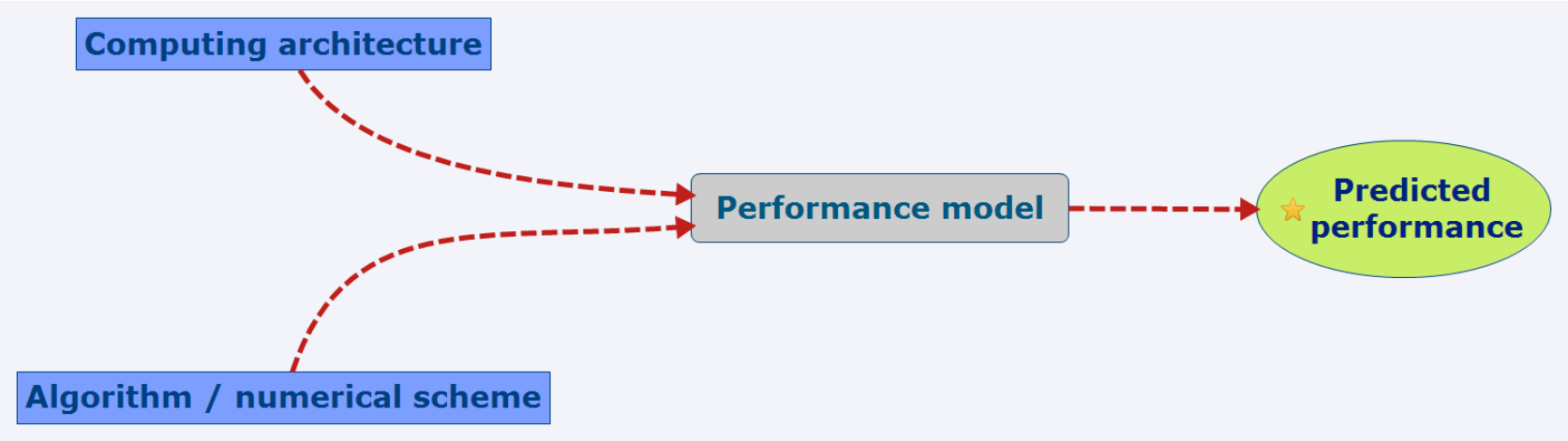
- Performance issue : sometimes hard to choose the adequate data structures
- Global performance : (most of the time) hard to understand

2. Performance modeling on (multicore) CPU

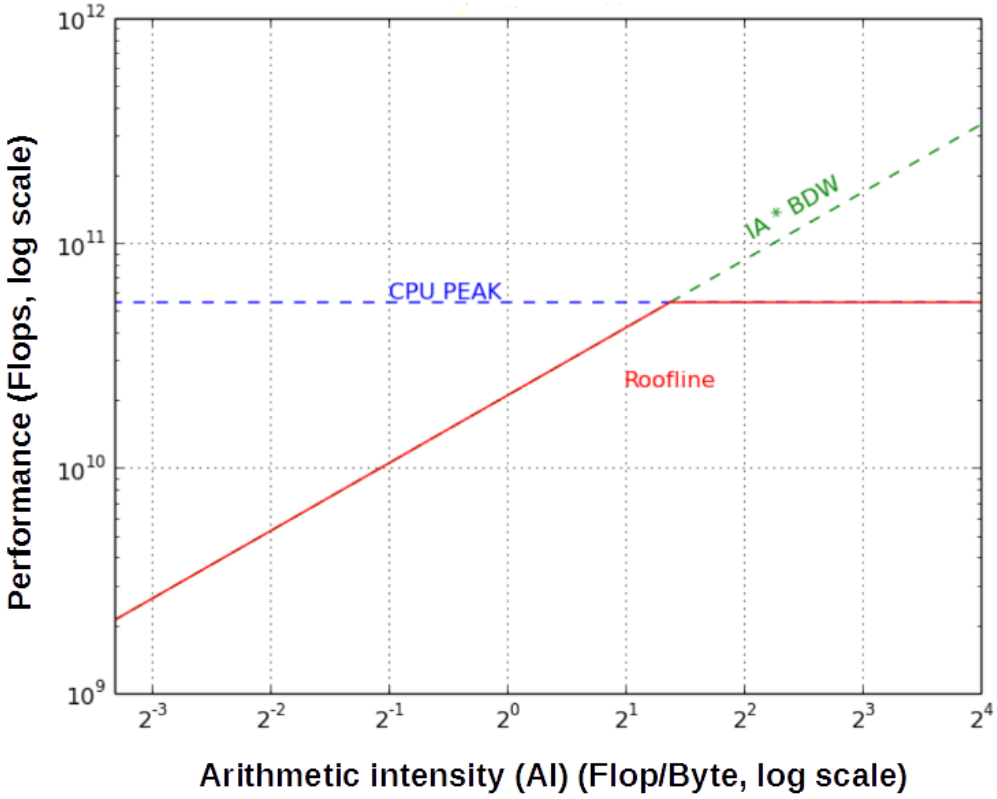
Performance modeling : objectives

- Derive a model which is able to predict the performance of a given solver on a given CPU-based machine.

Elements of a performance model



Reference node-based performance model : Roofline



[S. Williams, A. Waterman & D. Patterson, Roofline : an insightful visual performance model for multicore architectures, Communications ACM 55(6) : 121-130 (2012)]

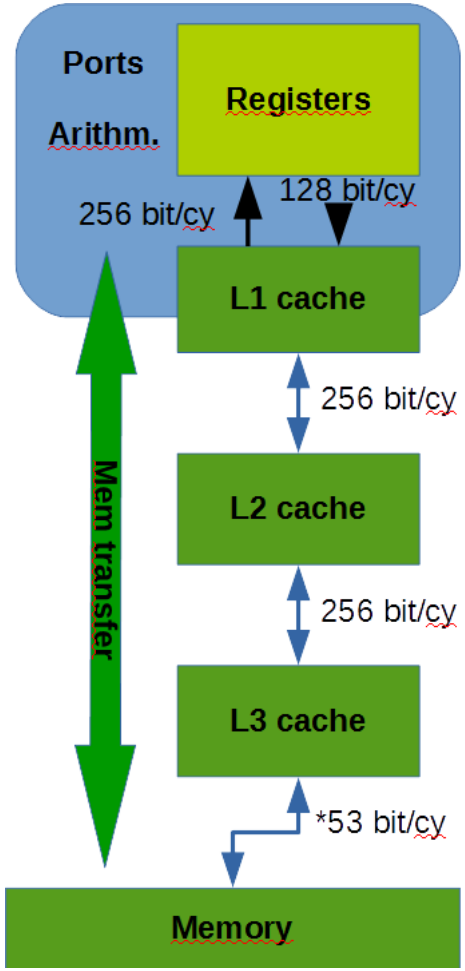
Execution Cache Memory (ECM) model

[Hager & Treibig 2010]

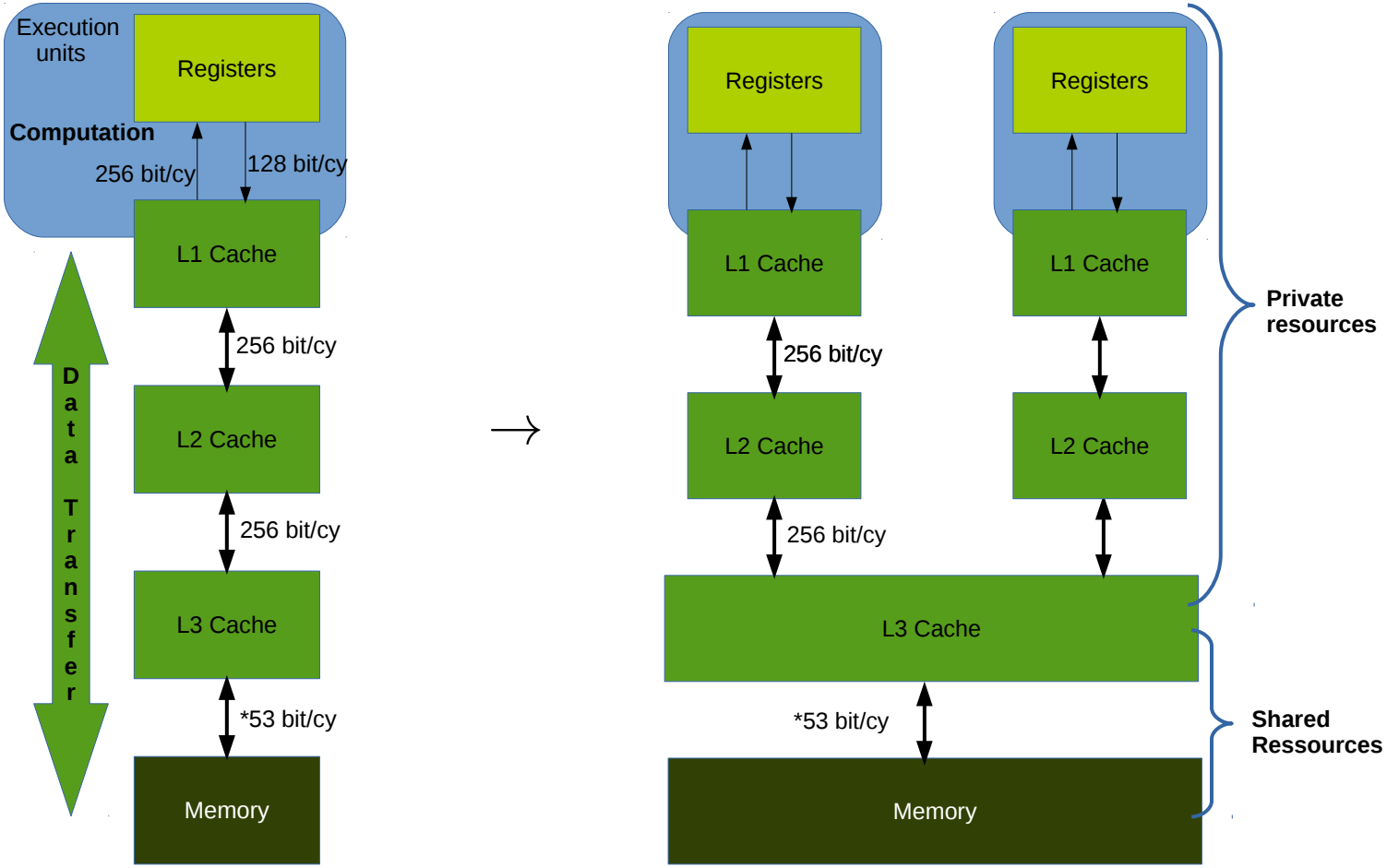
Metrics : time [cycle / cache line update]

- Execution times at (ALU+L1 cache) level
- Transfer speeds between Lx-L(x+1) cache/memory

[G. Hager et al., Concurrency and Computation : Practice and Experience, DOI : 10.1002/cpe.3180 (2013).]

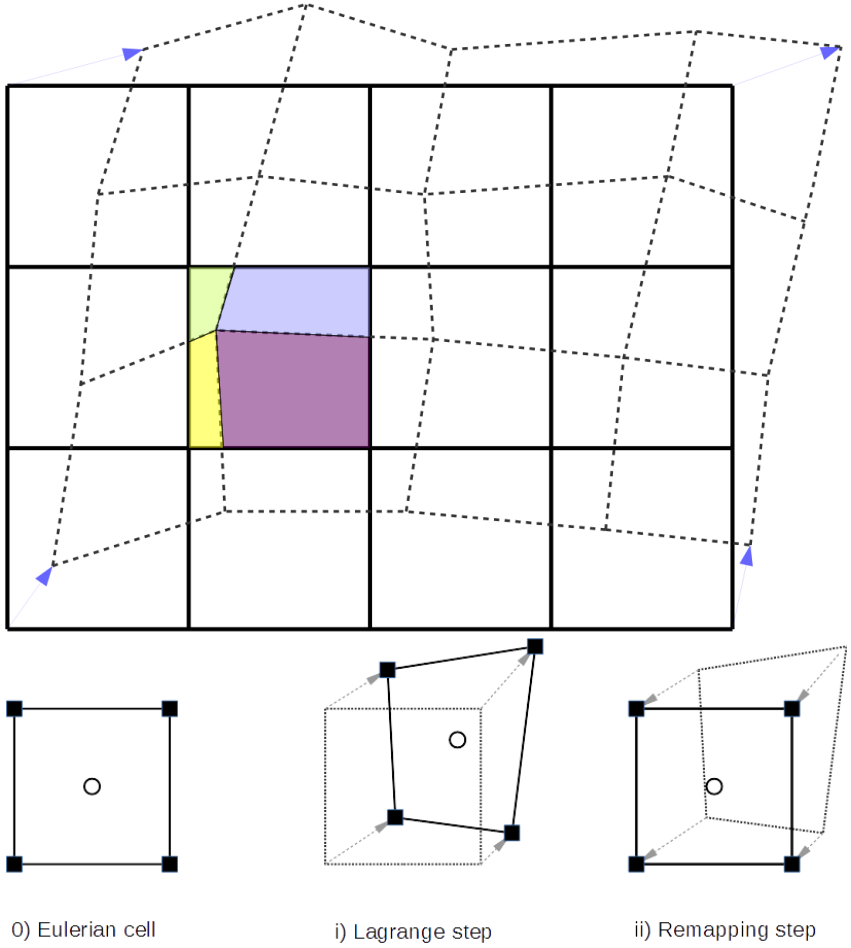


ECM : from one-core to multi-core CPU



3. Application : Lagrangian remap Hydrodynamics solver

Lagrangian remap algorithm



Dataflow diagram

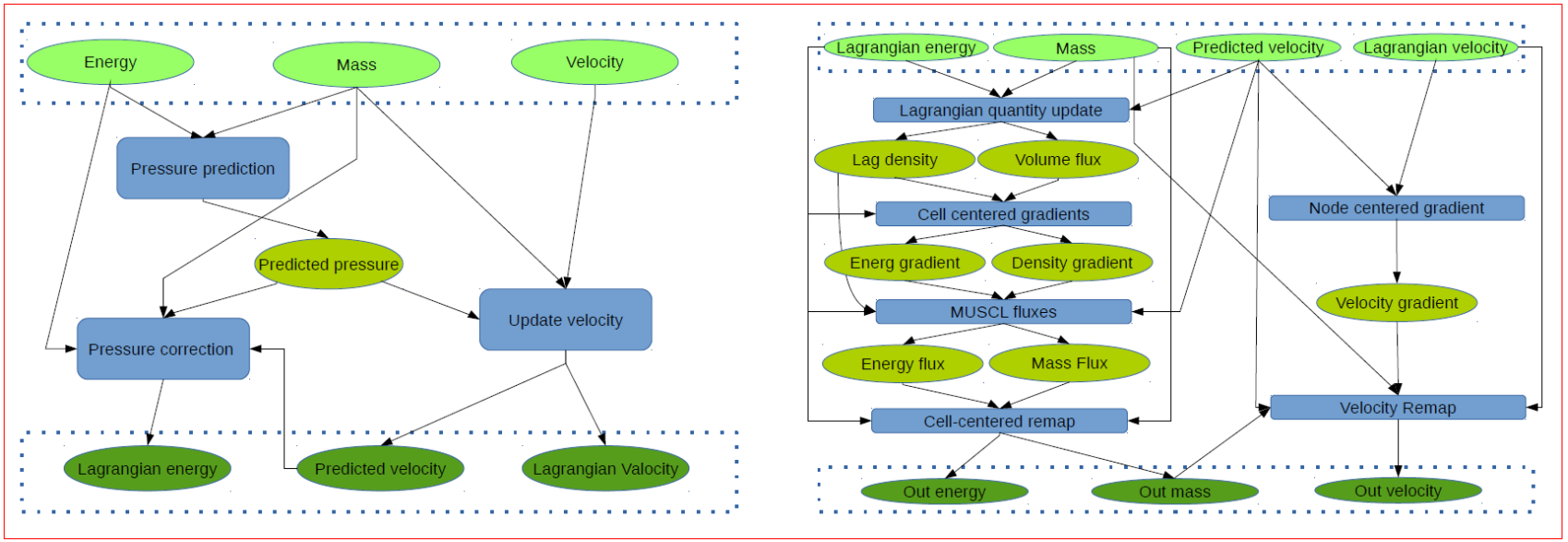


Figure: LEFT : Lagrangian step & RIGHT : remapping step dataflow diagrams (input data, kernels, output data)

Results

- Roofline is able to predict performance of compute-bound kernels (O(5%) errors)
- ECM is necessary for good performance prediction of memory-bound kernels (O(5%) errors)

Redesign of Lagrangian remapping

Why is it necessary to redesign ?

- Staggered velocity variables involve more communications (memory-bound)
- Alternating direction strategies multiply communications too
- The remapping process is not SIMD

Focus : remapping process

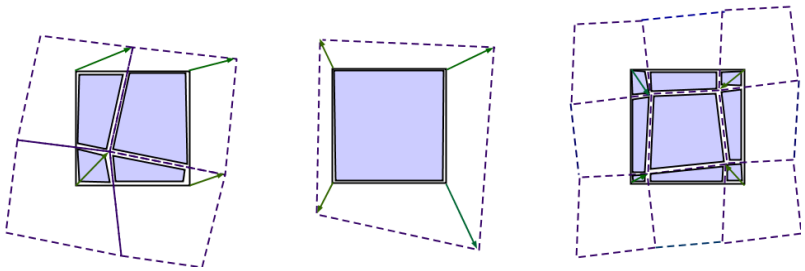


Figure: Geometric remapping : bad for SIMD

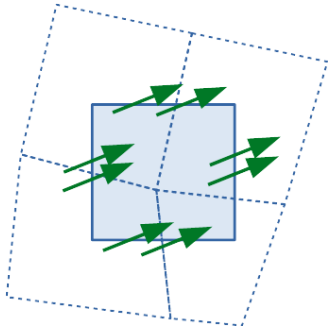
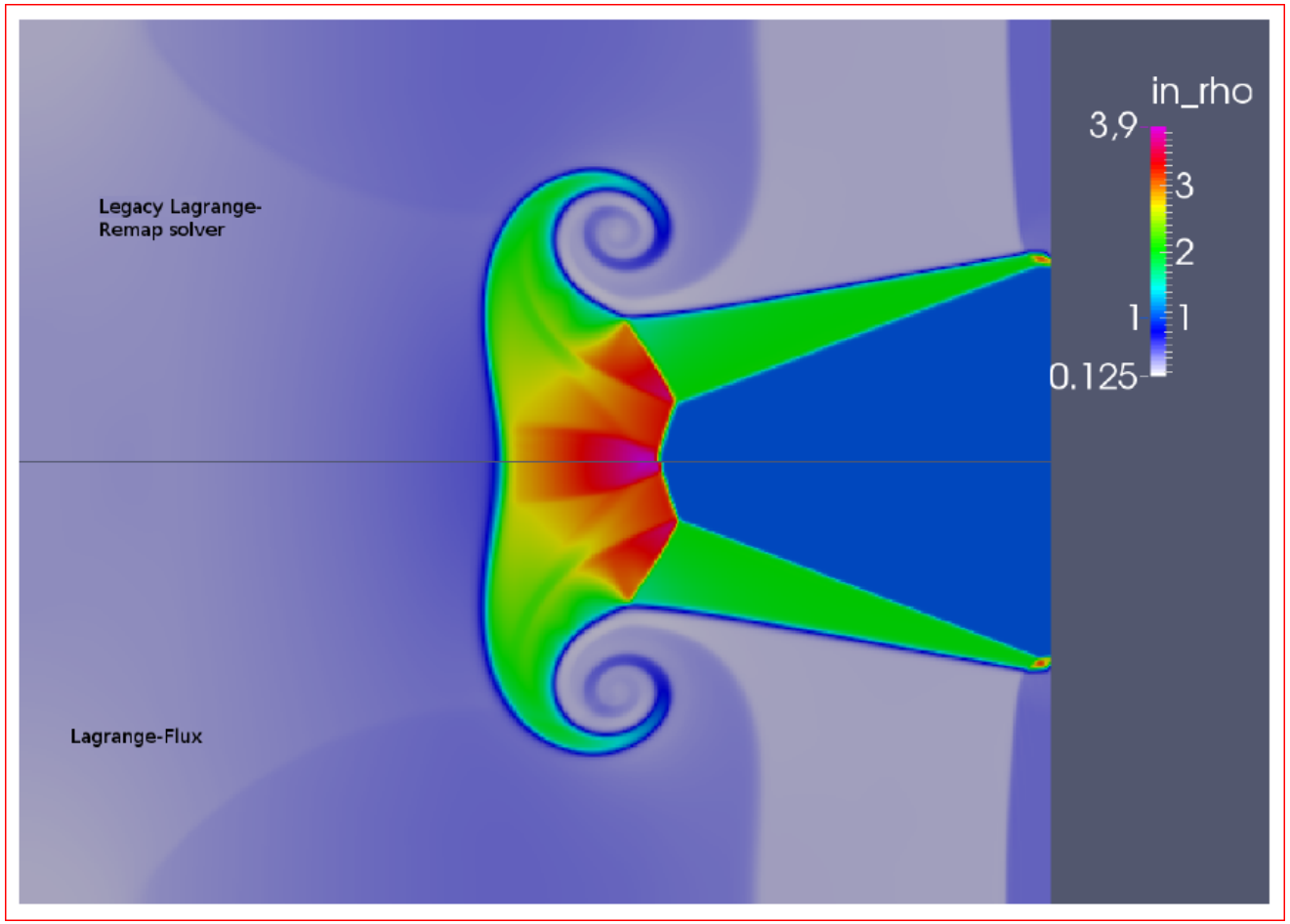
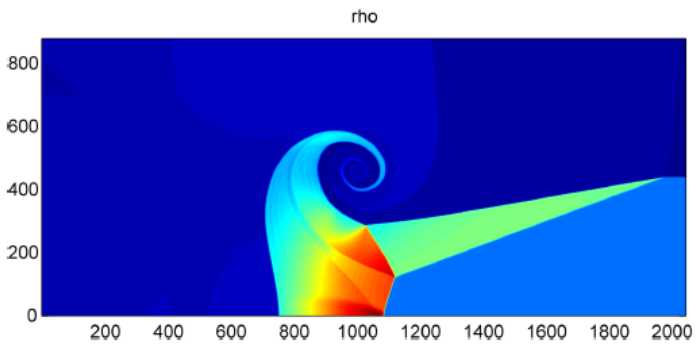


Figure: Geometry-free reformulation by balance of advection fluxes : SIMD-ready ⇒ Lagrange-flux schemes

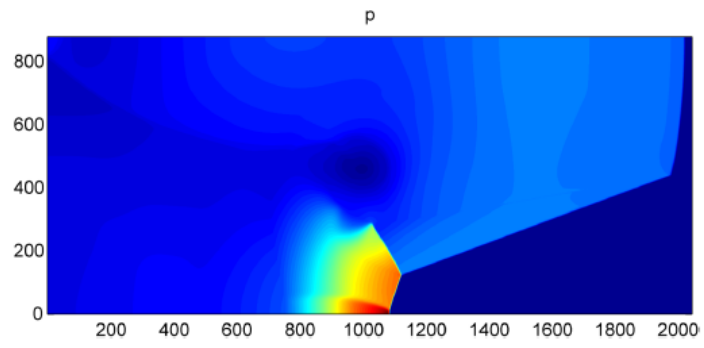
Comparison of results between legacy Lagrange-remap and Lagrange-flux scheme



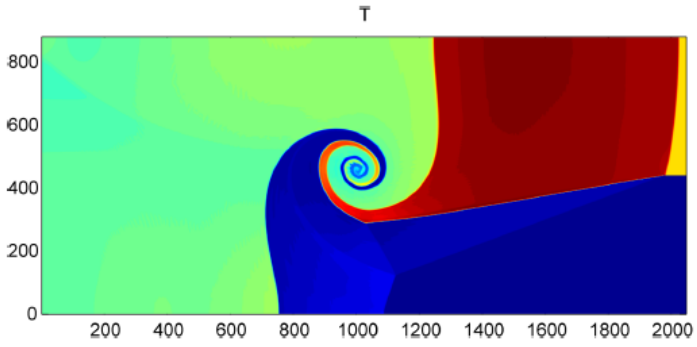
Multimaterial extension of Lagrange-flux schemes



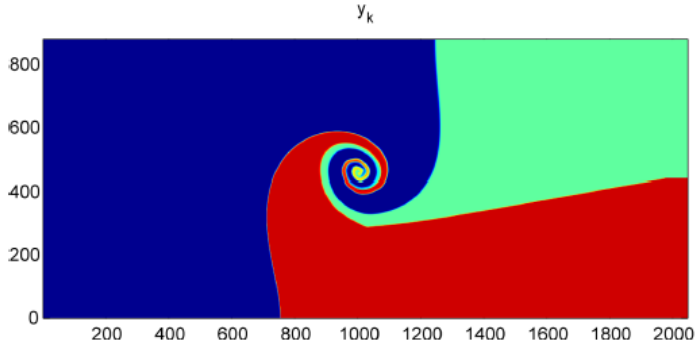
(a) Density field



(b) Pressure field

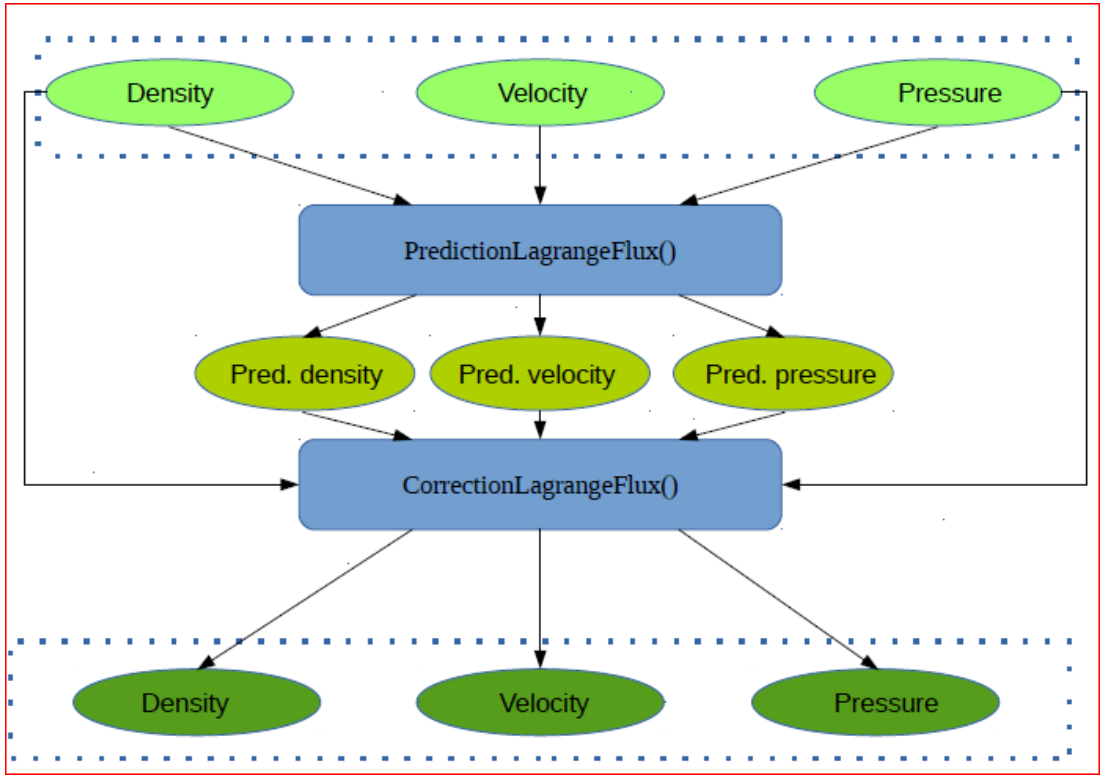


(c) Temperature field



(d) Colored representation of material indicators

Dataflow diagram for the Lagrange-flux scheme



Rem : easy parallelization & vecto by #pragma directives.

Performance prediction vs measurements with ECM

Kernel name	type	prediction	measure	error
Full Lagrange-Remap	MB	10697 cy/CL	10092 cy/CL	6%
Prediction Lagrange-Flux	CB	5712 cy/CL	5548 cy/CL	2.8%
Correction Lagrange-Flux	CB	5712 cy/CL	5690 cy/CL	0.3%

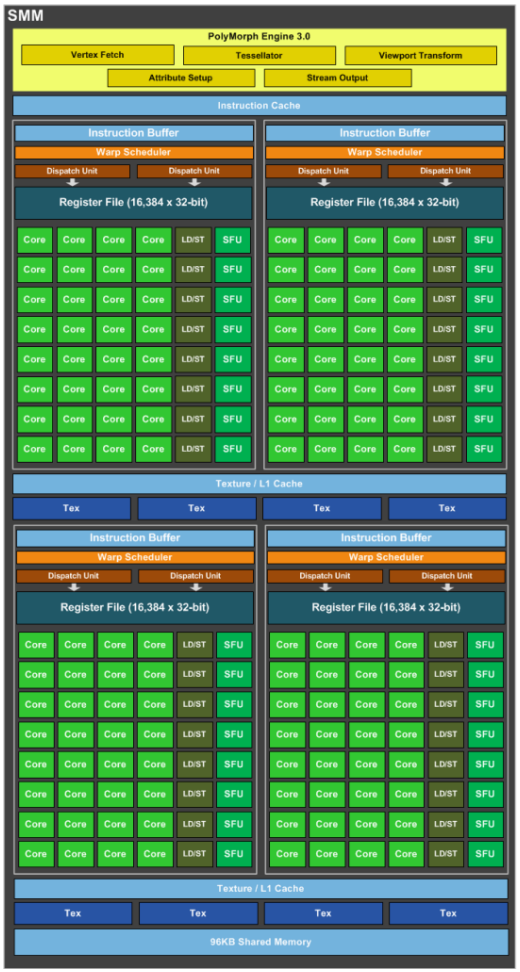
Figure: Performance metrics : nb. of cycles (cy) per cache line (CL)

Comparison of vecto + multicore scalability performance

Scheme	1 core	1 core AVX	16 cores AVX	scalability
Lagrange-Flux	1.9	3.9	52.0	27.1
Lagrange-Remap	2.4	3.7	36.5	14.6

Figure: Performance comparison in millions of cell updates (MCUPs), CPUs INTEL SandyBridge 2×8 cores. Performed on a fine mesh.

Research perspective



- Extend the performance modeling methodology to GPU/manycore-based
- (NB Some literature already exists on this subject)

Concluding remarks

- Nice speedups with GPU but sometimes unclear performance issues
- Performance modeling is a powerful tool to understand global performance
- Performance analysis is insightful for bottleneck analysis
- Some parts of the solver may be redesigned for achieving better arithmetic intensity, leading to new computational approaches
- Future work : accurate performance models for accelerators/manycore/GPU proc. ?

References

1. T. Gasc, F. De Vuyst, M. Peybernes, R. Poncet, R. Motte, Building a more efficient Lagrange-remap scheme thanks to performance modeling, ECCOMAS 2016, Paper P12210, Proc. of the Conference ECCOMAS 2016, Minisymposium "MS 414 - New trends in numerical methods for multi-material compressible fluid flows", submitted (2016).

2. F. De Vuyst, T. Gasc, R. Motte, M. Peybernes, R. Poncet, Lagrange-Flux Eulerian schemes for compressible multimaterial flows, ECCOMAS 2016, Paper E8851, Proc. of the Conference ECCOMAS 2016, Minisymposium "MS 414 - New trends in numerical methods for multi-material compressible fluid flows", submitted (2016).

3. F. De Vuyst, M. Béchereau, T. Gasc, R. Motte, M. Peybernes, R. Poncet, Stable and accurate low-diffusive interface capturing advection schemes, submitted to IJNMF, Special issue for the MULTIMAT 2015 Conf., submitted (2016).

4. R. Poncet, M. Peybernes, T. Gasc and F. De Vuyst, Performance modeling of a compressible hydrodynamics solver on multicore CPUs, Proc. of the PARCO 2015 Conference, Edinburgh, accepted (2016).