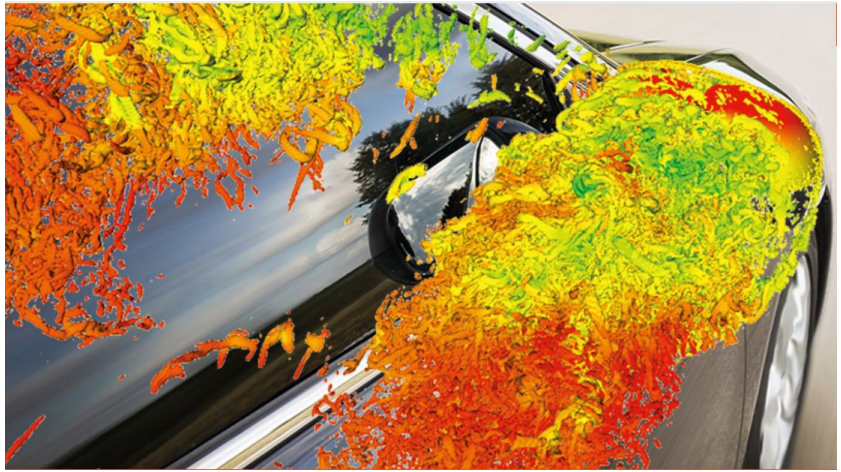


OpenFOAM CONFERENCE

HPC Benchmark Project: follow-up

14 Oct. 2020, Industry Session 11:00-11:20 CET

I. Spisso[1]*, i.spisso@cineca.it

R. Da Via'*

R. Ponzini*

S. Bna*

A . Memmolo*

G. Boga*

* SuperComputing Applications and Innovation (SCAI) Department, CINECA, Italy

1) Chairman of OpenFOAM HPC Technical Committee

8th OpenFOAM Conference 2020

13 October 2020 - 15 October 2020

Worldwide, Online

09:00 am - 06:00 pm CEST (Central European Summer Time)

CINECA

get it right®

www.esi-group.com

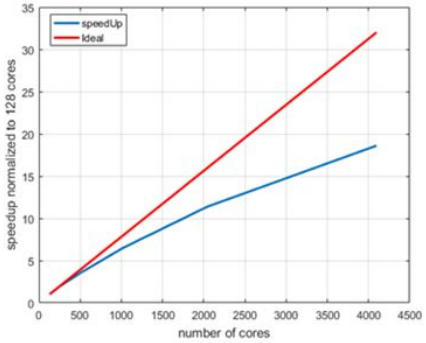
Outline of the presentation

OpenFOAM: HPC Performances

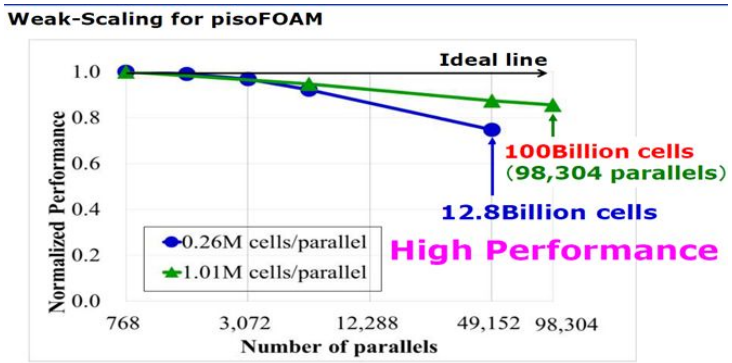
- OpenFOAM scales reasonably well up to thousands of cores, upper limit order of thousands of cores. We are looking at achieving radical scalability of cases of 100's of millions / billions of cell on 10K-100K cores.

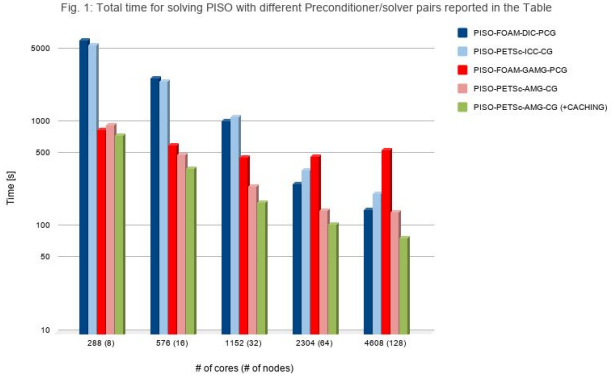
- A custom version by Shimuzu Corp., Fujitsu Limited and RIKEN on old K computer) was able to achieve high performance on 100 thousand MPI tasks on a large scale transient CFD simulation up to 100 billion cell mesh [1].

- Recent add-on: PETSc4FOAM, a liibrary to plug-in PETSc into the OpenFOAM Framework. It provides a plug-in for embedding PETSc and its external dependencies (i.e. Hypre, ML) into arbitrary OpenFOAM simulations [2]. Available in OpenFOAM-v2006



1 – Code scalability on Marconi KNL system on a 64 Million mesh



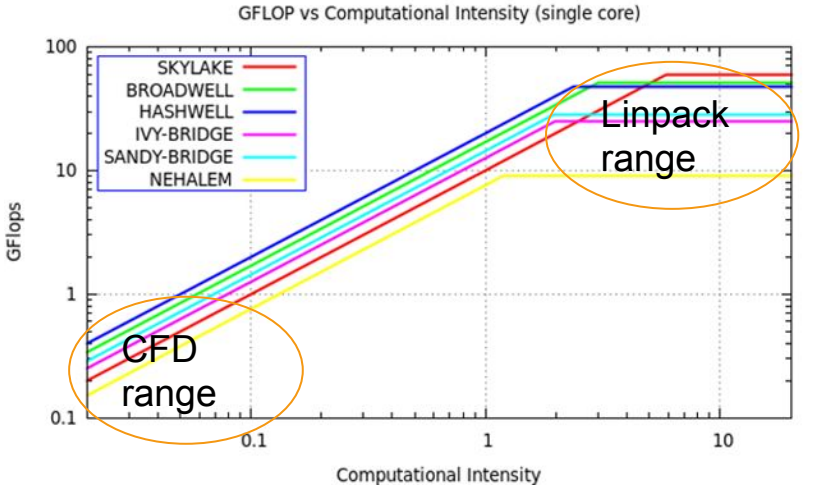


HPC Roof-line Model and Computational intensity

The roofline model: is a visual performance model.

Performance bound (y-axis, Flops) ordered according to arithmetic intensity (x-axis, FLOPS/byte)

- Memory bandwidth versus cpu speed
- FEA/ CFD are typically memory bounded codes
- Peak flops: two order of Magnitude
- Linpack: Dense Linear Algebra



- Suggestion: use less cores than the maximum available in nodes
- Switch to memory bound to MPI bound, move to the x-right of the plot

OpenFOAM: HPC bottlenecks and exaFOAM

- The technological trends of exascale HPC is moving towards the hybrid CPUs+GPUs clusters (not exclusively, see #1 Top500,ARM based) with orders of millions or cores, June 2020 Top 500 List [3]. Example: Summit (OKNL) [4], Marconi100 (CINECA) [5]: One node consists of 2 IBM Power9 procs+ 4/6 Nvidia V100 gpus
(*exaFLOPs= 10^18 FLOPS or a billion of billion calculations per seconds)*

- The well known bottlenecks for enabling OpenFOAM to perform on massively parallel clusters are:
  - Scalability of **the linear solvers** and their limits in the parallelism paradigm.
  - **Sparse Matrix storage** format: The **LDU** sparse matrix storage format used internally does not enable any cache-blocking mechanism (SIMD, vectorization).
  - The **I/O data storage** system: when running in parallel, the data for decomposed fields and mesh(es) has historically been stored in multiple files within separate directories for each processor, which is a bottleneck for big simulation.

- To overcome such issue a community effort has been collect in exaFOAM
  - It is a Consortium (12 Partners + Stakeholders + Supporters) consisting of a well-balanced group of experts to work on the **co-design** of **OpenFOAM** targeting (pre)-exascale HPC architectures.
  - Grant Funded by EuroHPC-03-2019: Industrial software codes for extreme scale computing environments and applications
  - Consortium led by ESI-OpenCFD. Expected start date: Jan 2021. Duration: 3 years

OpenFOAM HPC Technical Committee (TC)

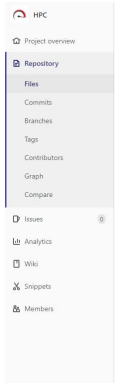
https://www.openfoam.com/governance/technical-committees.php#tm-hpc

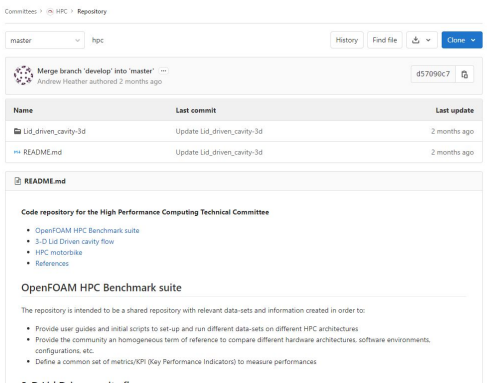
- The Technical Committees cover all the key focus areas for OpenFOAM development; they assess the state-of-the-art, need and status for validation, documentation and further development.
-
- Remits of the HPC TC
  - OpenFOAM recommendations to Steering Committee in respect of HPC technical area
    - Work together with the Community to overcome the actual HPC bottlenecks of OpenFOAM
    - Scalability of linear solvers
    - Adapt/modify data structures of Sparse Linear System to enable vectorization / hybridization
    - Improve memory access on new architectures
    - Improve memory bandwidth
    - Parallel pre- and post-processing, parallel I/O
  - Strong co-design approach
  - Identify algorithm improvements to enhance HPC scalability
  - Interaction with other the Technical Committee (Numerics, Documentations)
- Priorities of HPC TC:
  - HPC Benchmark
  - GPU enabling of OpenFOAM
  - Parallel I/O (to be tested)
    - The *adiosWrite* function object has been rewritten to use the ADIOS2 library for parallel IO and is now available as a regular OpenFOAM module
    - Collated file format in openfoam.org

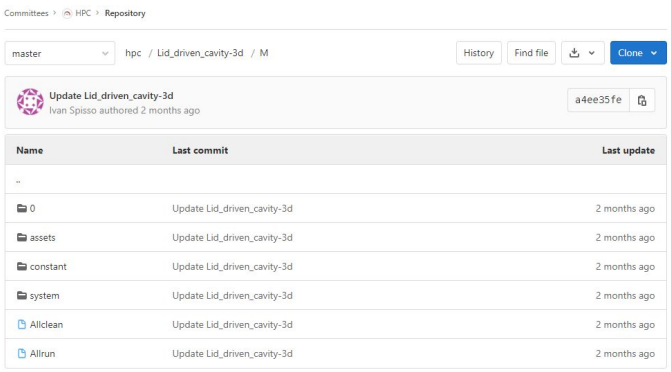
Code repository for HPC Technical Committee

https://develop.openfoam.com/committees/hpc

- Create an open and shared repository with relevant data-sets and information
- Provide an User-Guide and initial scripts to set-up and run different data-sets
- Provide to the community a homogeneous term of reference to compare different HW architectures, configurations and different SW environments
- Define a common set of Metrics/KPI (Key Performance Indicators) to measure performances
- Data-sets are public availables in the repo (work in progress)

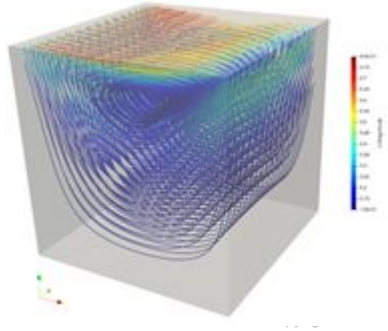


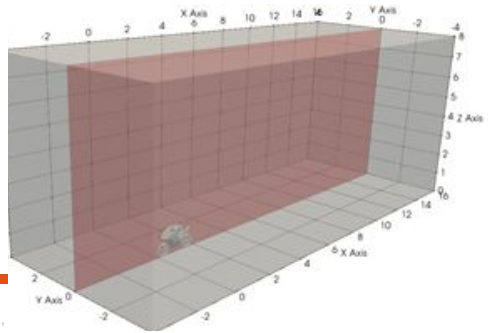


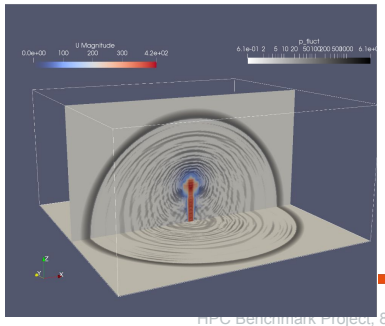


List of test-cases (work in progress)

Test-cases	Problem sizes (millions of cells)					Physical features	Relation with HW/SW infrastructure			Bottleneck(s)	KPIs (Key Performax Index)
	S	M	L	XL	XXL		cpu intensive	memory intensive	I/O intensive		
3D Lid driven cavity flow	1	8		64	216	incompressible laminar flow regular and uniform grid	yes	yes	no	Linear algebra solvers Data structure	Time to solution memory bandwidth bound
HPC motorbike	8.6	17.2	34.4	68.8		external aero incompressible turbulent flow non-uniform grid	yes	yes	no	Linear algebra solvers Data structure	Time to solution memory bandwidth bound
INGV test-case	2	16.3			131	fully compressible (shock waves) transient, unsteady, turbulent	yes	yes	no	Matrix assembly Linear algebra solvers	Time to solution SP vs DP gain
ExaFOAM test-cases											
.....											







INGV Test-case, provided by EU Project COE in Solid Earth

ChEESE
Center of Excellence for Exascale in Solid Earth

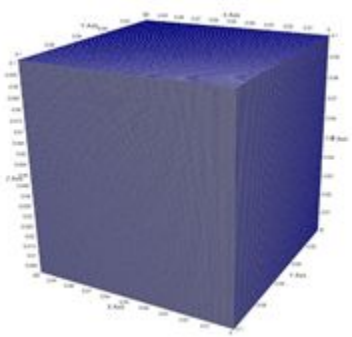
To be integrated in exaFOAM
WP2: Validation and Assessment

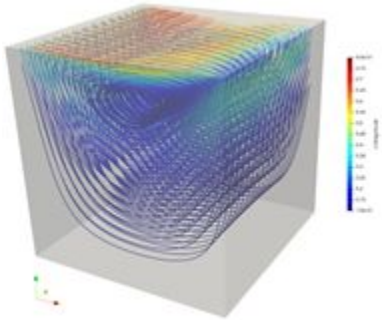
oup,

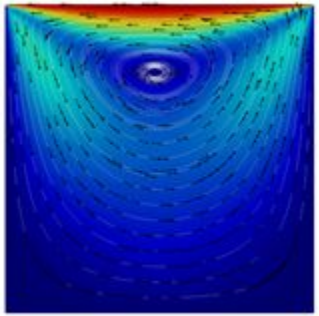
HPC Benchmark Project, 8th OpenFOAM Conference 2020, I. Spisso et al., CINECA

Initial Benchmark test case:

3-D Lid Driven Cavity







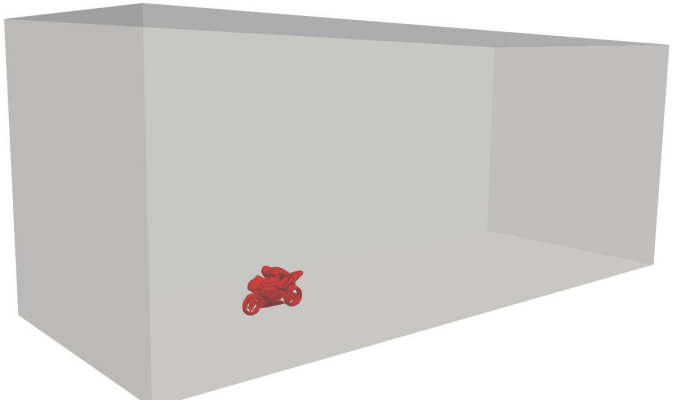
- Derived from 2-D Lid driven cavity flow of OpenFOAM tutorial
- Simple geometry. B.C.: inflow, outflow, no slip walls
- Increasing sizes, from 1 M up to 216 M. (XXL work in progress)

- Stress Test for the linear solver algebra, mainly pressure equation
- KPIs to be monitored: Wall time, Memory Bandwidth, Bound
- Bound = The metric represents the percentage of Elapsed time spent heavily utilizing system bandwidth (available only for Intel architectures)

Test-case	S	M	XL	XXL
dΔx (m)	1.00E-03	5.00E-04	2.50E-04	1.25E-04
Cube side length d (m)	0.1	0.1	0.1	0.1
N of cells tot. (millions)	1.0	8.0	64.0	216.0
n of cells lin (on cube's edge)	100	200	400	600
kinematic viscosity v (m^2/se)	1.0E-02	1.0E-02	1.0E-02	1.0E-02
Co	1	0.5	0.25	0.125
Physical final Time	0.5	0.5	0.5	0.5
delta T (sec.)	1.00E-03	2.50E-04	6.25E-05	1.56E-05
Reynolds	1.0E+01	1.0E+01	1.0E+01	1.0E+01
Top wall velocity U (m/s)	1	1	1	1
num. of Iterations	5.00E+02	2.00E+03	8.00E+03	3.20E+04

Initial Benchmark test case:

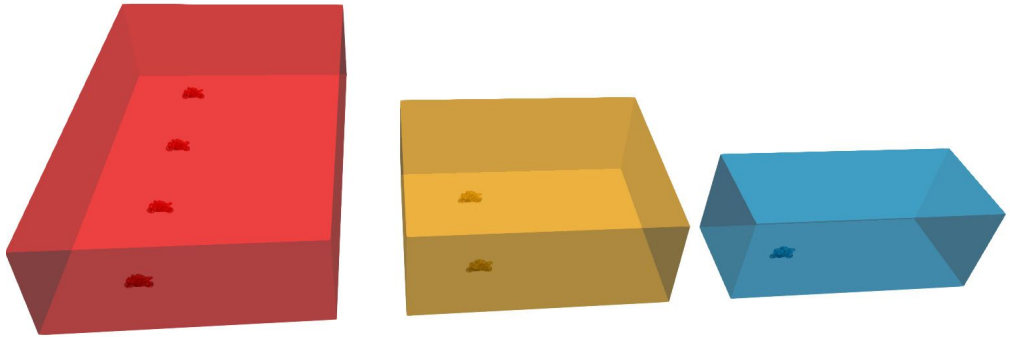
HPC Motorbike



- The test case has been developed from the well known motorbike tutorial
- Mesh generated from *blockMesh* is 3 times finer, along each axis, with respect to the tutorial one
- The snappyHexMeshDict hasn't been touched
- The base mesh (size S) is about 8.6 * 10^6 cells (the tutorial one is 3.2 * 10^5)
- Sizes: XL (34 M), M (17.2 M), S (8.6 M) # of cells

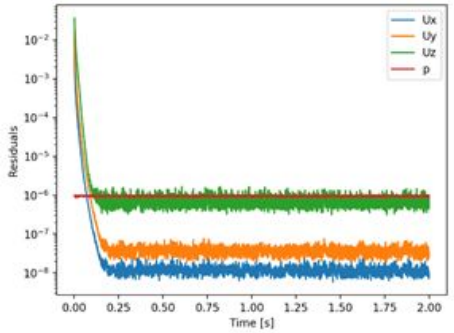
To further increase the number of cells, without changing the mesh topology, the mesh can be mirrored using the *mirrorMesh* tool

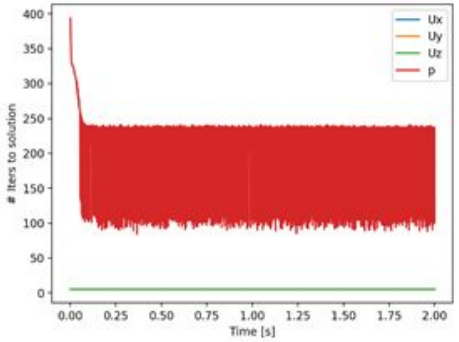
- Cell sizes are kept the same size
- No need to change the setup (b. c. for turbulence fields - yPlus values are the same on each mesh)
- Suitable to perform weak scaling



Set-up of linear algebra solvers

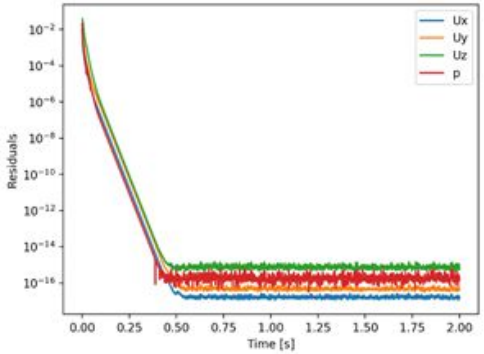
```
solvers
{
    p
    {
        solver          PCG;
        preconditioner  DIC;
        tolerance       1e-04;
        relTol          0;
        maxIter         3000;
    }
pFinal
    {
        $p;
        relTol          0;
    }
    U
    {
        solver          PBiCGStab;
        preconditioner  DILU;
        tolerance       0;
        relTol          0;
        maxIter         5;
    }
}
```

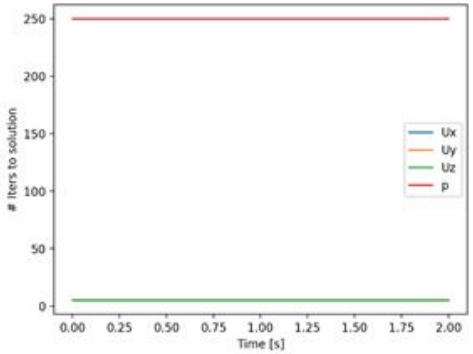




fixedNORM: exit norm is fixed. Only a subset of iteration can be used for benchmarking

```
solvers
{
    p
    {
        solver          PCG;
        preconditioner  DIC;
        tolerance       0;
        relTol          0;
        maxIter         250;
    }
    .....
```





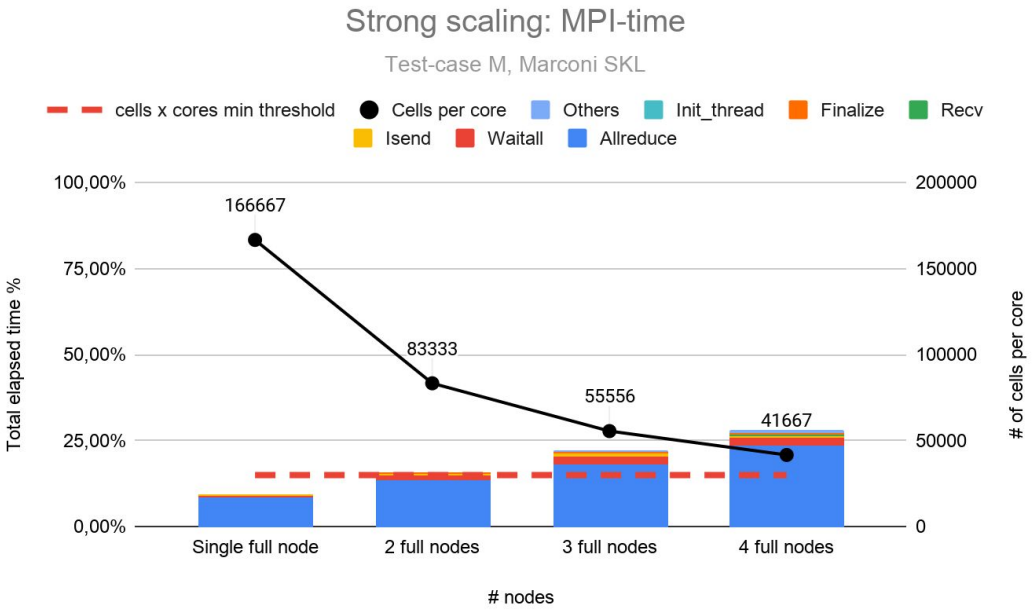
fixedITER: comp. load is fixed. It is NOT representative of real set-up. Used only for preliminary tests in the dev. phase or stress HPC architectures

HPC hardware comparison

Architectural Technical resource specs

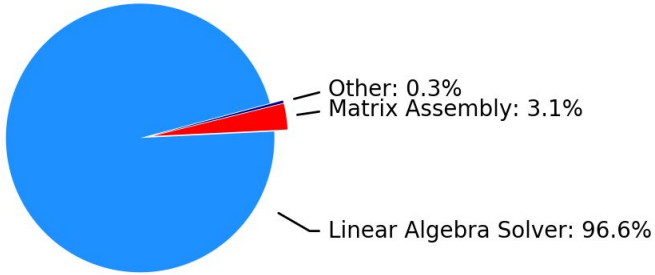
Cluster	computes				memory / node (GB/node)	network	factory memory bandwidth (GB/sec)
	proc. type	cores per node	tot. number of nodes	Accelerators			
Galileo (CINECA)	Intel Xeon E5-2697 v4 @ 2.30 GHz (Broadwell)	2 x 16	1022	60 nodes with 1 nVidia K80 GPU 2 nodes with 1 nVidia V100 GPU	128	Intel OmniPath, 100 Gb/s	153.6
Marconi (CINECA)	Intel Xeon 8160 @ 2.10 GHz (Skylake)	2 x 24	2188	n.a.	192	Intel OmniPath, 100 Gb/s	256
Marconi100 (CINECA)	IBM Power9 AC922 @ 3.1 GHz	2 x 16	980	4 x Nvidia V100 GPUs, Nvlink 2.0, 16 GB	256	Mellanox Infiniband EDR DragonFly +	300 (Power9)
ARMIDIA (E4)	Marvell TX2@2,2/2.5 GHz	2 x 32	8	Nvidia Tesla V100 PCIE 32GB	256	Mellanox Infiniband EDR 100 Gb/s	341.34
more clusters to be add							

Profiling of fixedIter setup: Lid Driven cavity - M

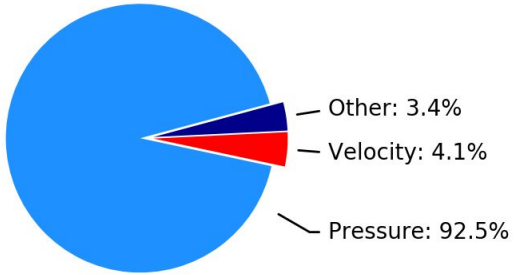


CPU time distribution - 1 node

Discretization and solution - cumulative values (u,p)

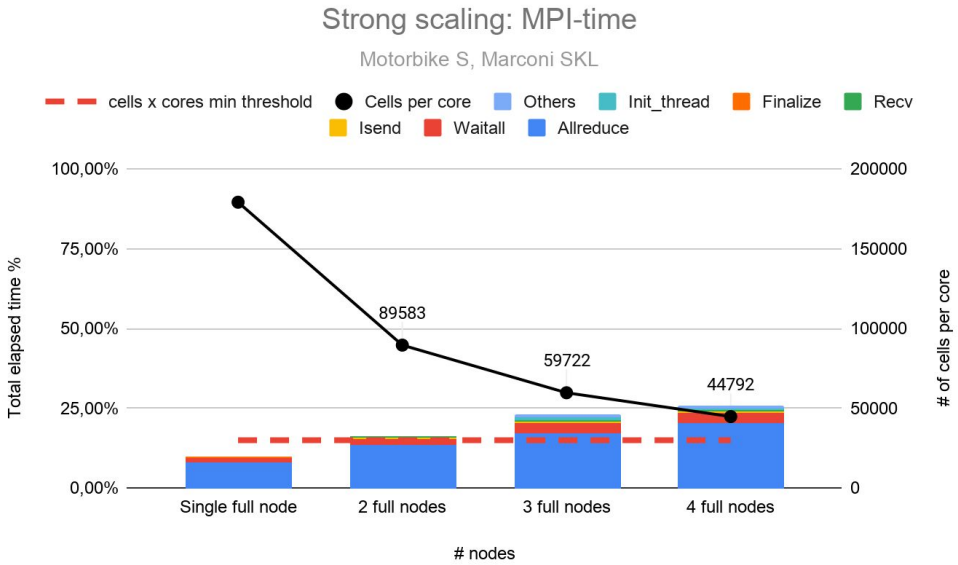


Distribution between the solvers (u,p,other)



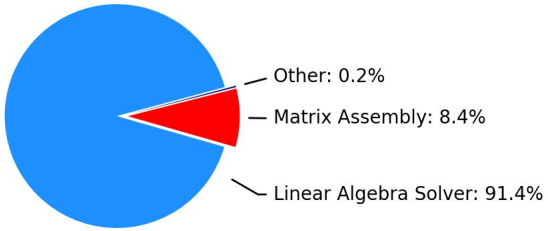
- Simulations run on Marconi SKL
- OpenFOAM v1912, compiled with Icc18 OptSKL + Prof options
- Profiling tools: APS from Intel Vtune 2020 + HPC toolkit (spack build 2020).
- Similar CPU time distribution when using up to 4 nodes
- Special queue to access to hardware's counter (perf_event_paranoid value to 0 or less)

Profiling of fixedIter setup - HPC Motorbike - size S

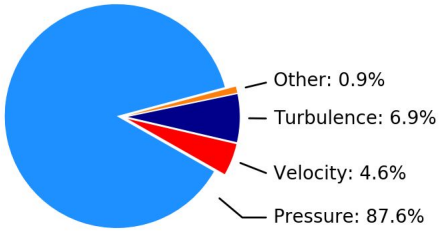


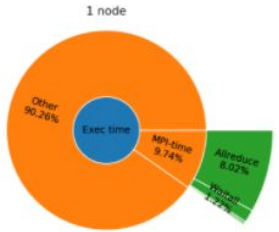
CPU time distribution - 1 node

Discretization and solution - cumulative values (u,p,turb)



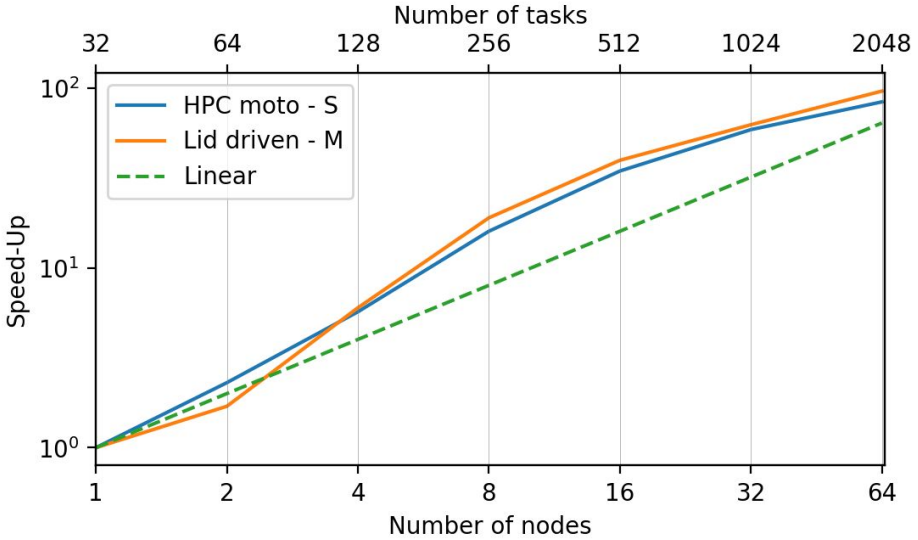
Distribution between the solvers (u,p,turb,other)

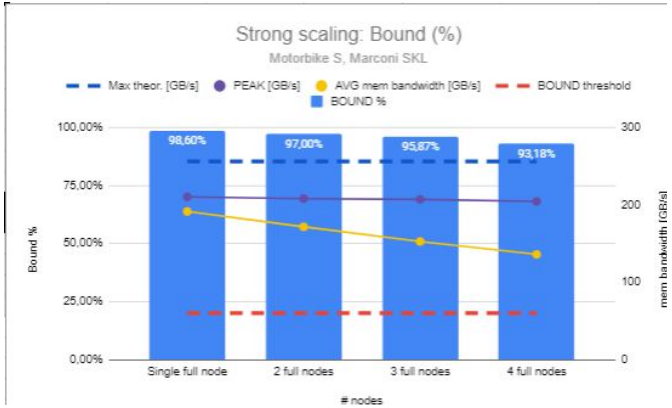


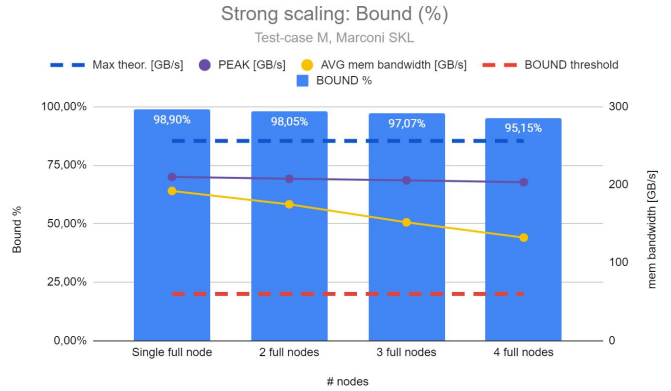


- Simulations run on Marconi SKL
- OpenFOAM v1912, compiled with Icc18 OptSKL + Prof options
- Profiling tools: APS from Intel Vtune 2020 + HPC toolkit (spack build 2020).
- Similar CPU time distribution when using up to 4 nodes
- Special queue to access to hardware's counter (perf_event_paranoid value to 0 or less)

Memory bound & Strong scaling- fixedIter setup



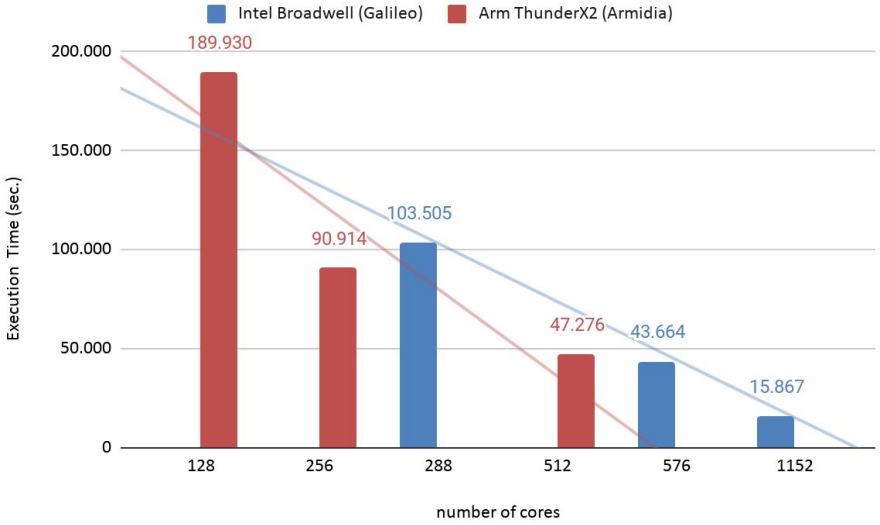


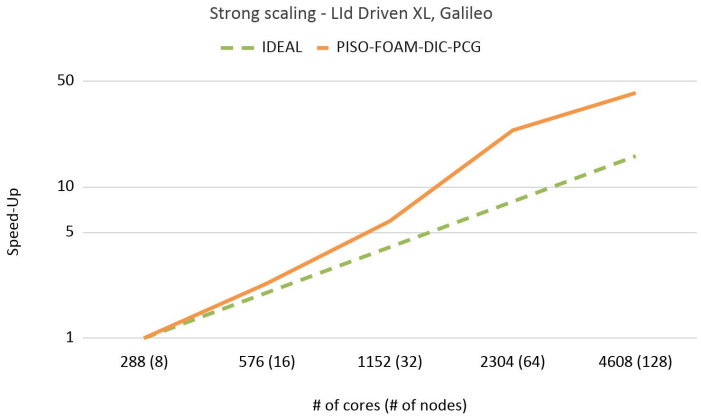


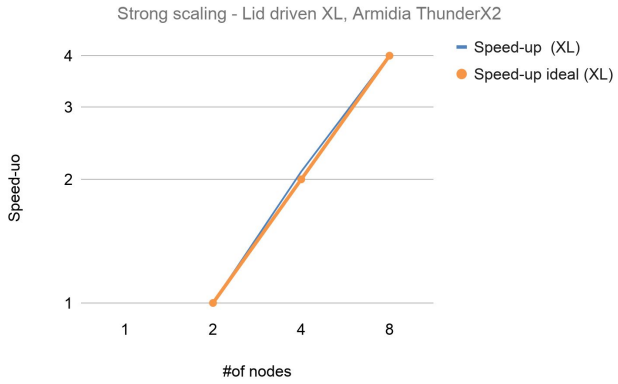
- Top Figure: Strong scaling comparison between Lid Driven M and HPC motorbike on Galileo (Broadwell). Superlinear effect
- Same fvSolution setup, fixedIter
- Right figures: bound (%) on Marconi SKL, up to 4 nodes.
- Same strong scaling behavior. These cases are heavily memory bound, with low number of nodes, explaining the superlinear scaling for number of nodes > 8

HPC comparison: Armidia and Galileo cluster

Lid Driven XL test-case







- Galileo: strong scaling, same superlinear trend of M case
- Armidia: strong scaling linear up to 8 nodes (ref 2 nodes)
- Comparison by using the same number of cores
  Runs by using the full number of tasks-per-nodes 36 per Broadwell, 64 for ARM
  Continuous line, trend line

Conclusion / Further work / Acknowledgment

- Conclusion
  - Follow-up on OpenFOAM HPC Benchmark project
  - For the test-cases studied, the DIC-PCG solver is highly mem. bounded with low number of nodes (full node configuration)
  - As a consequence, superlinear strong scalability is observer for high number of nodes
  - Linear algebra solver is where most of the time is spent (around 90%, mainly pressure)
  - Present preliminary results on HPC architectures

- Further work
  - Finalize test-cases (XXL Lid-Driven, XL motorbike)
  - Run weak-scaling motorbike with suitable
  - Low level profiling of linear algebra (preconditioner + solvers), extraction of roofline of full-code/functions
  - Profiling with hardware independent tool (HPCtoolkit, score-p, LIKWID)
  - Profiling with PoP Methodology (WP5 exaFOAM)
  - Add energy to solution as KPI
  - Monitoring the developments of the European Processor Initiative

- Acknowledgment

  - G. Amati, M. Valentini, *SuperComputing Applications and Innovation (SCAI) Department, CINECA, Italy*
  - F. Brogi, *INGV Istituto Nazionale di Geofisica e Vulcanologia, Pisa Italy*
  - M. Cerminara, *INGV Istituto Nazionale di Geofisica e Vulcanologia, Pisa Italy*
  - F. Magugliani, *E4 Italy*
  - G. Rossi, *Intel, Italy*
  - N. Ashton, *Amazon Web Service, UK*
