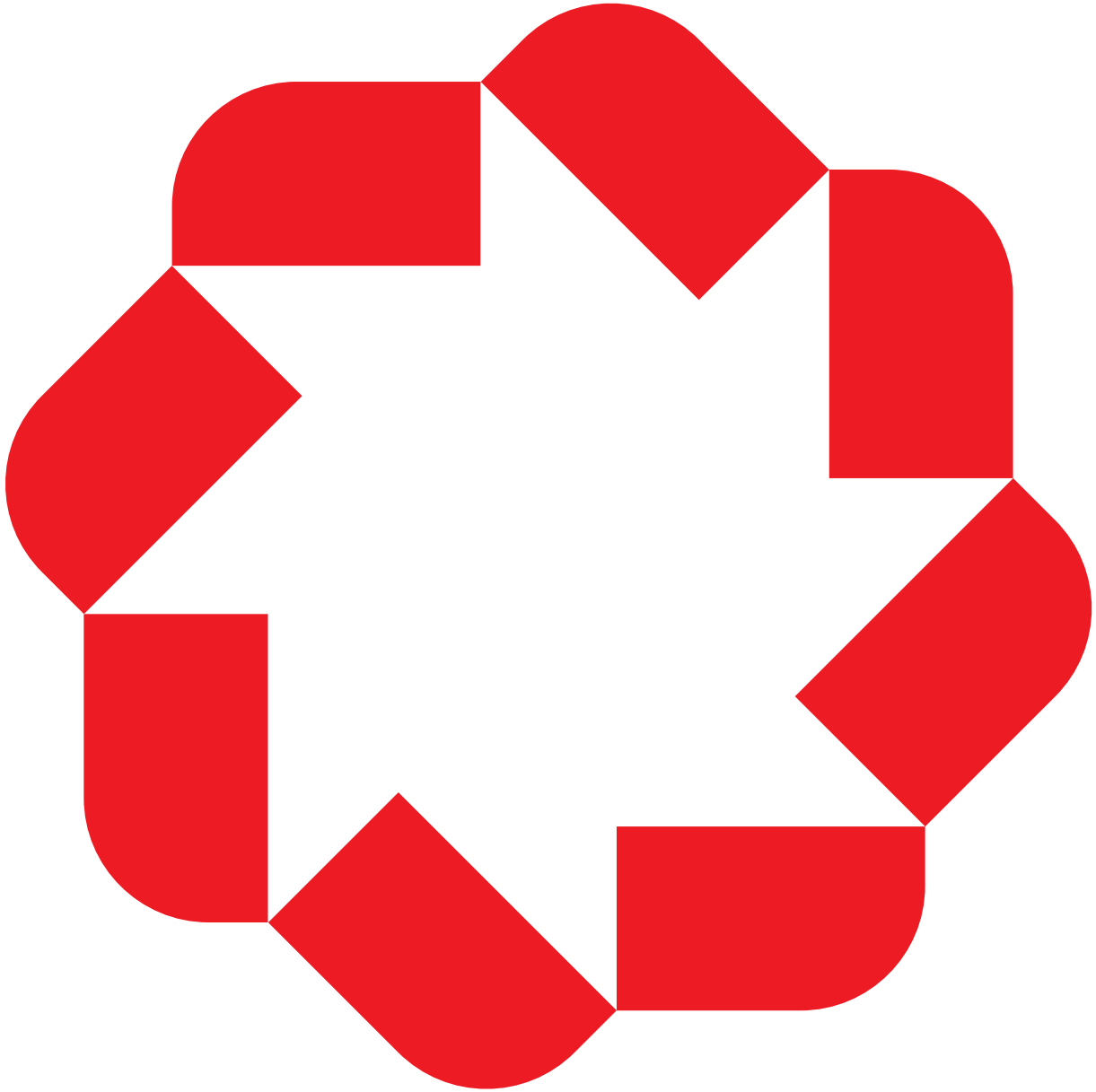
Recent developments about GPU acceleration in HELYX

1st OpenFOAM – HPC Challenge

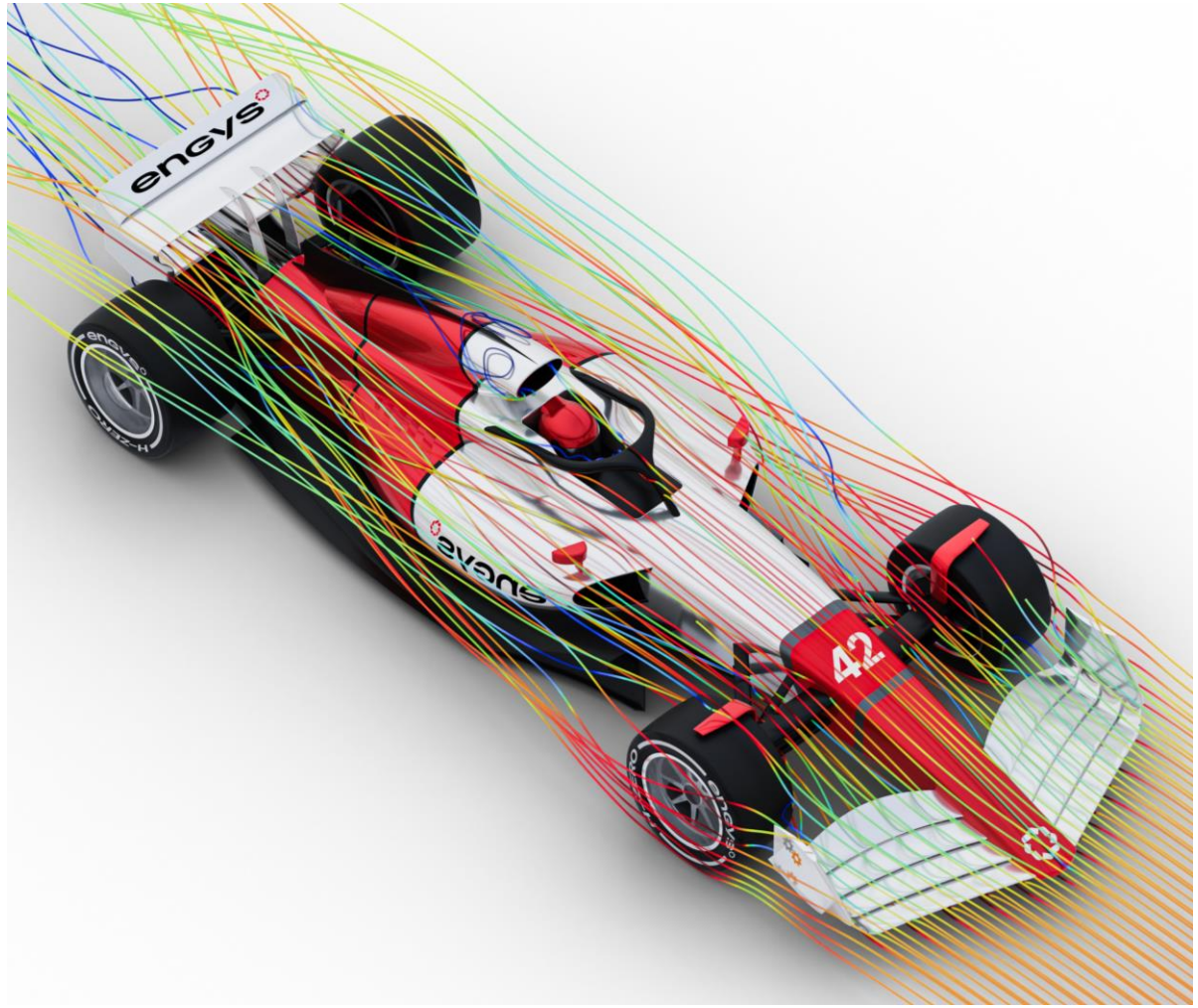
Stefano Oliani, ENGYS

July, 1st 2025



Introduction

- Native GPU HELYX core is the ultimate goal
- Completed a prototype to showcase capabilities
- Support from F1 teams consortium



Recent Developments | GPU Status

Prototype native GPU – operational

- Matrix assembly + linear algebra on GPU
  - Most common discretization schemes available
  - Segregated and block-coupled steady/transient solvers
  - Most common linear solvers available
    - ✓ Segregated : PBiCGStab, GAMG with multicolor smoothers/preconditioners (GS, DILU)
    - ✓ Coupled: FGMRES, AMG with block-coupled smoothers (GS, DILU)

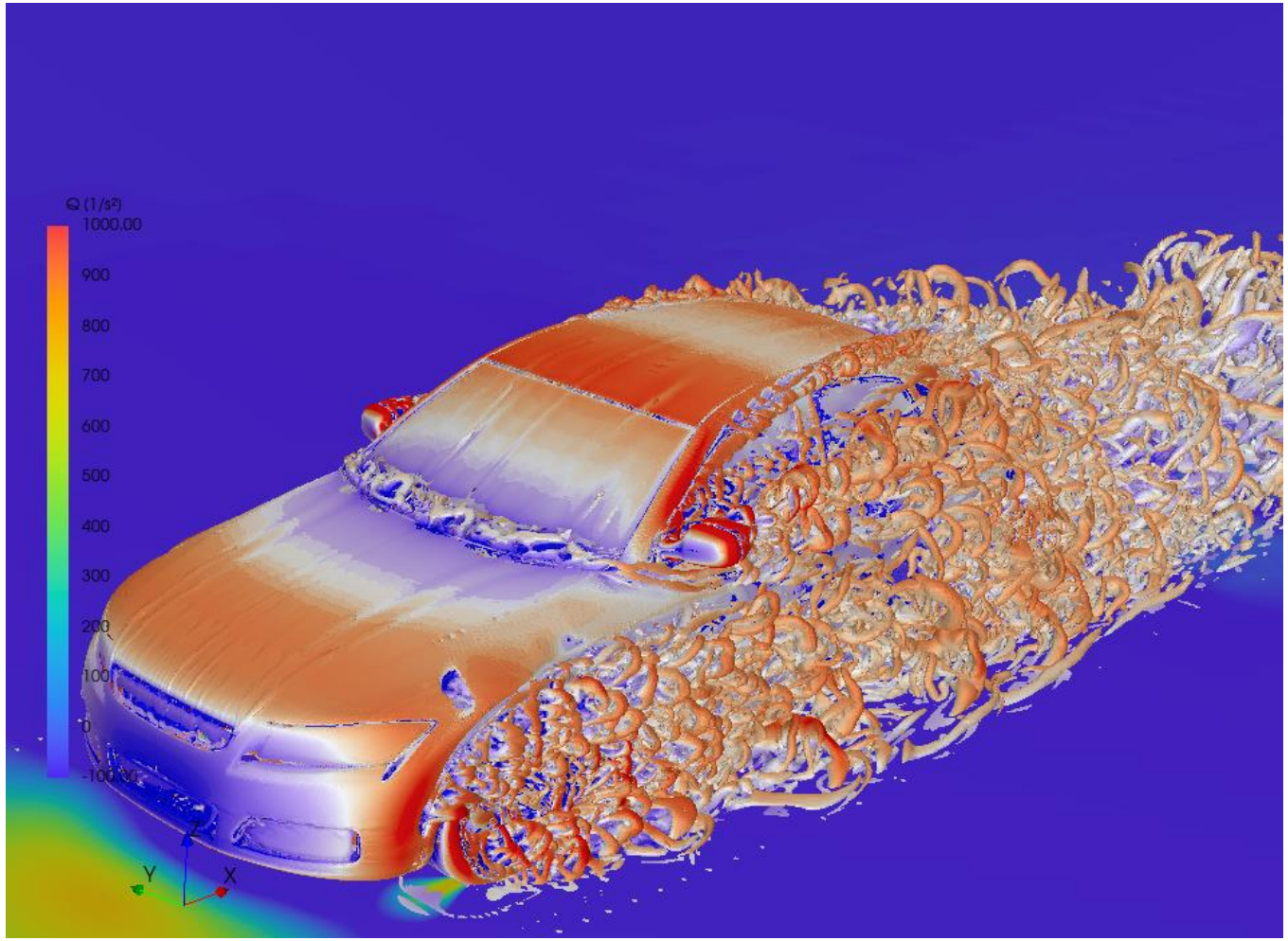
Recent Developments | GPU Status

Turbulence:

- RANS closure (k-Omega SST),
- DES, DDES (k-Omega SST, SA)
- LES (k-Eqn, WALE, Smagorinsky)

Postprocessing:

- Most common ext aero FO (forceCoeffs, probes, fields etc)



Recent Developments | GPU Status

Hardware

- CUDA (Nvidia hardware)
- Compiled and tested also on AMD – looking into performance gap

OFHPC Challenge - Results

Grids:

- Used fine grid from committee (236M)

Methods:

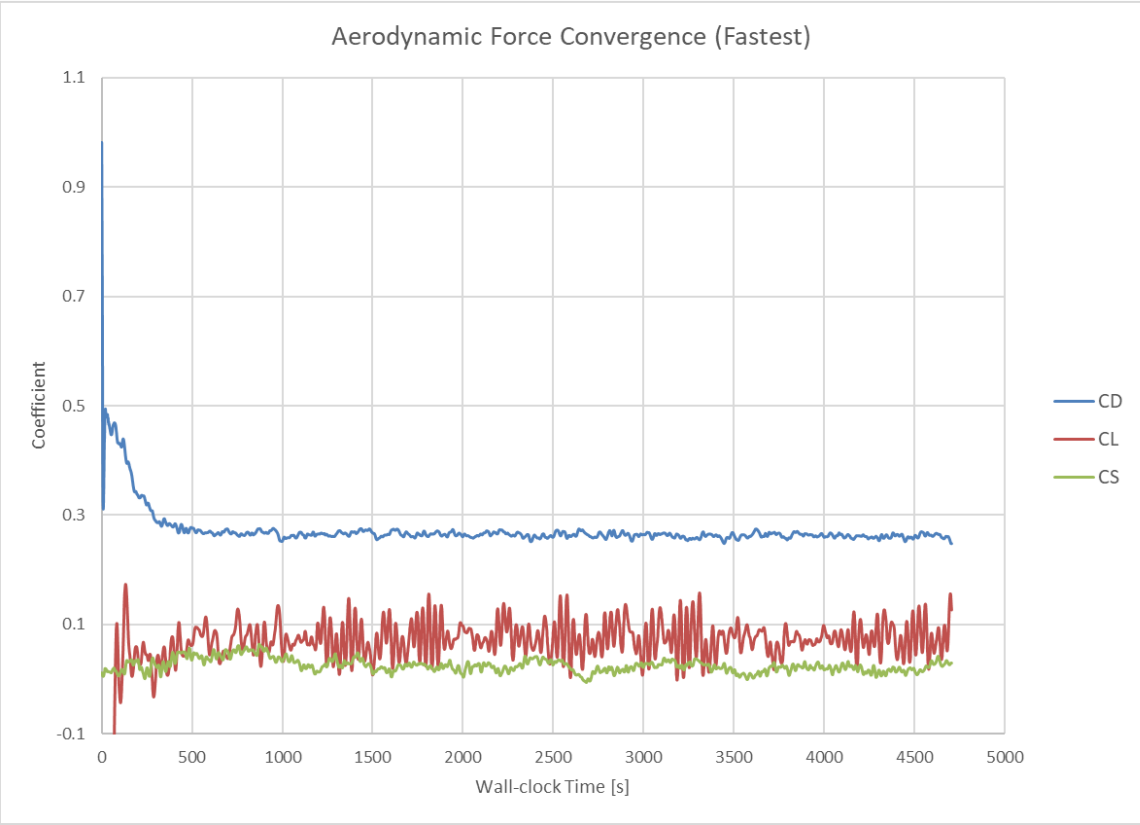
- Software Track, simulations run with standard HELYX and HELXY-GPU
- Setup is the same as specified for the Hardware track (fvSchemes, fvSolution, FO, initialization etc)
- Convergence assessment with meanCalc

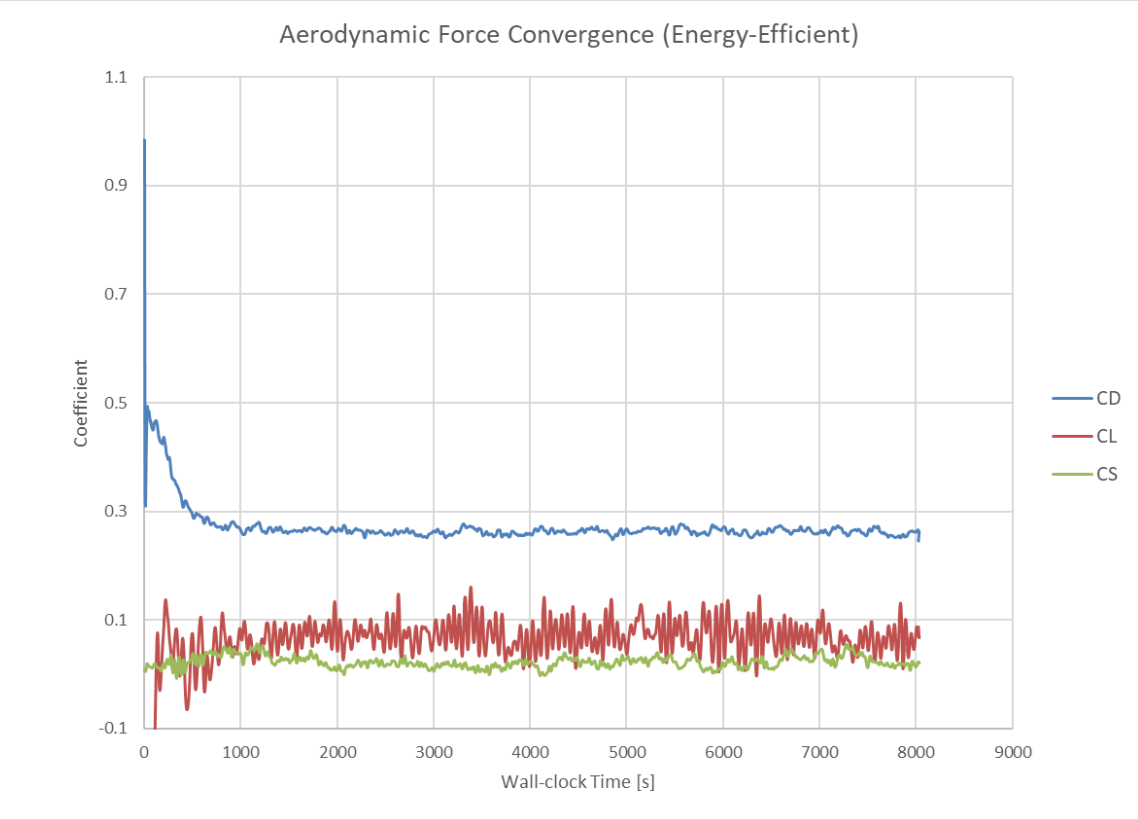
Hardware:

- Simulations run on Leonardo (CINECA) – Booster partition.
- Nodes used at full capacity

No Focus on Pre-processing

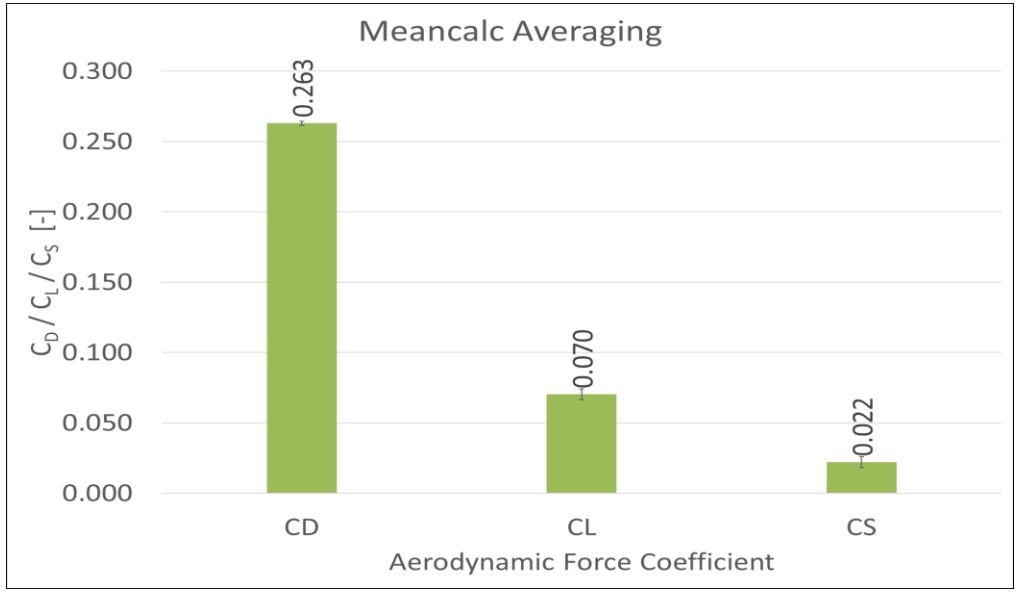
OFHPC Challenge - Results

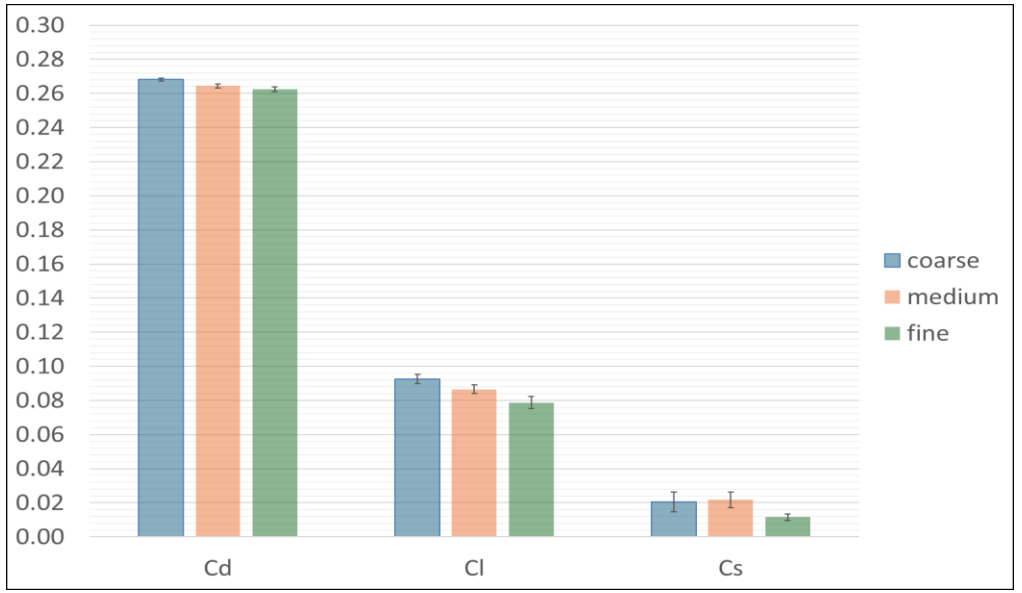




OFHPC Challenge - Results

- Drag coefficient is in good agreement with vanilla OpenFOAM
- Lift coefficient convergence is more sensitive to simulation settings (e.g. schemes and linear solver tolerances/multigrid setting)
- Side coefficient is more similar to coarse/medium grid
- All coefficients are in good agreement with data from AutoCFD workshop





OFHPC Challenge - Results

ENERGY EFFICIENT

FASTEST

Software	Simulation #1	Simulation #2	Simulation #3
Solver Type:	FVM	FVM	FVM
Compressibility:	Incompressible	Incompressible	Incompressible
Pressure/ Density Based:	Pressure	Pressure	Pressure
Segregate/Coupled:	Segregated	Segregated	Segregated
Velocity convection scheme:	linearUpwindV	linearUpwindV	linearUpwindV
Order of accuracy (space/time):	2nd	2nd	2nd
Steady-State/Transient:	Steady State	Steady State	Steady State
Time for pre-processing [s]:	0	0	0
Wall-clock time to completion excl. pre-processing [s]:	8042	4705	24271
Wall-clock time per timestep/iteration [s]:	2.01	1.18	6.07
TDP of system (CPU+Accelerator) [W]:	3700 W	7400 W	4000 W
Total energy to completion [kW*h or J]:	7.15 kWh	8.36 kWh	27.0 kWh
Decomposition Method:	Scotch	Scotch	Scotch
Renumbering Method:	RCM	RCM	RCM
Hardware:			
Server Type:	GPU	GPU	CPU
Hardware Spec (CPU):	Intel Ice Lake Xeon Platinum 8358	Intel Ice Lake Xeon Platinum 8358	ntel Ice Lake Xeon Platinum 8358
Hardware Spec (GPU):	Nvidia A100 custom 64GB	Nvidia A100 custom 64GB	/
# of nodes used:	2	4	16
# of CPU cores used:	8	16	512
# of GPUs used:	8	16	0
Memory type:	HBM2e	HBM2e	DDR4-3200
Memory capacity per node:	256 GB	256 GB	512 GB
Last-level Cache (LLC):	40 MB	40 MB	48 MB
Network Interconnect:	NVLink 3.0 + Mellanox HDR Infiniband	NVLink 3.0 + Mellanox HDR Infiniband	Mellanox HDR Infiniband
Storage Device:	SSD	SSD	SSD
Storage File-system:	Lustre	Lustre	Lustre
Network Topology:	Dragonfly	Dragonfly	Dragonfly

OFHPC Challenge - Results

- In terms of runtime: 1 A100 ≈ 195 Intel Xeon Platinum 8358 cores
- In terms of energy consumption: 8 A100 are almost 4 times more energy efficient than 512 Intel Xeon Platinum 8358 cores
- Cost (from AWS) : 8 A100 are 3 times cheaper than 512 Intel Xeon Platinum 8358 cores
- In terms of memory bandwidth: 1 A100 ≈ 300 Intel Xeon Platinum 8358 cores → lots of room for improvement (theoretically!)
- Moving from 8 to 16 GPUs the parallel efficiency is only 85 % → has to be improved (mainly due to poor resources utilization at coarse grid levels for multigrid)

Need to explore new algorithms/ techniques to achieve speedup ≈ MB ratio

Conclusions

- Prototype of HELYX native GPU solver has been shown
- GPU offload can help reducing turnaround time and energy consumption
- Testing on a broader spectrum of hardware is ongoing
- Significant work still needed to improve performance (especially linear algebra)
- Multigrid solvers tend to scale worse compared to other parts of the code

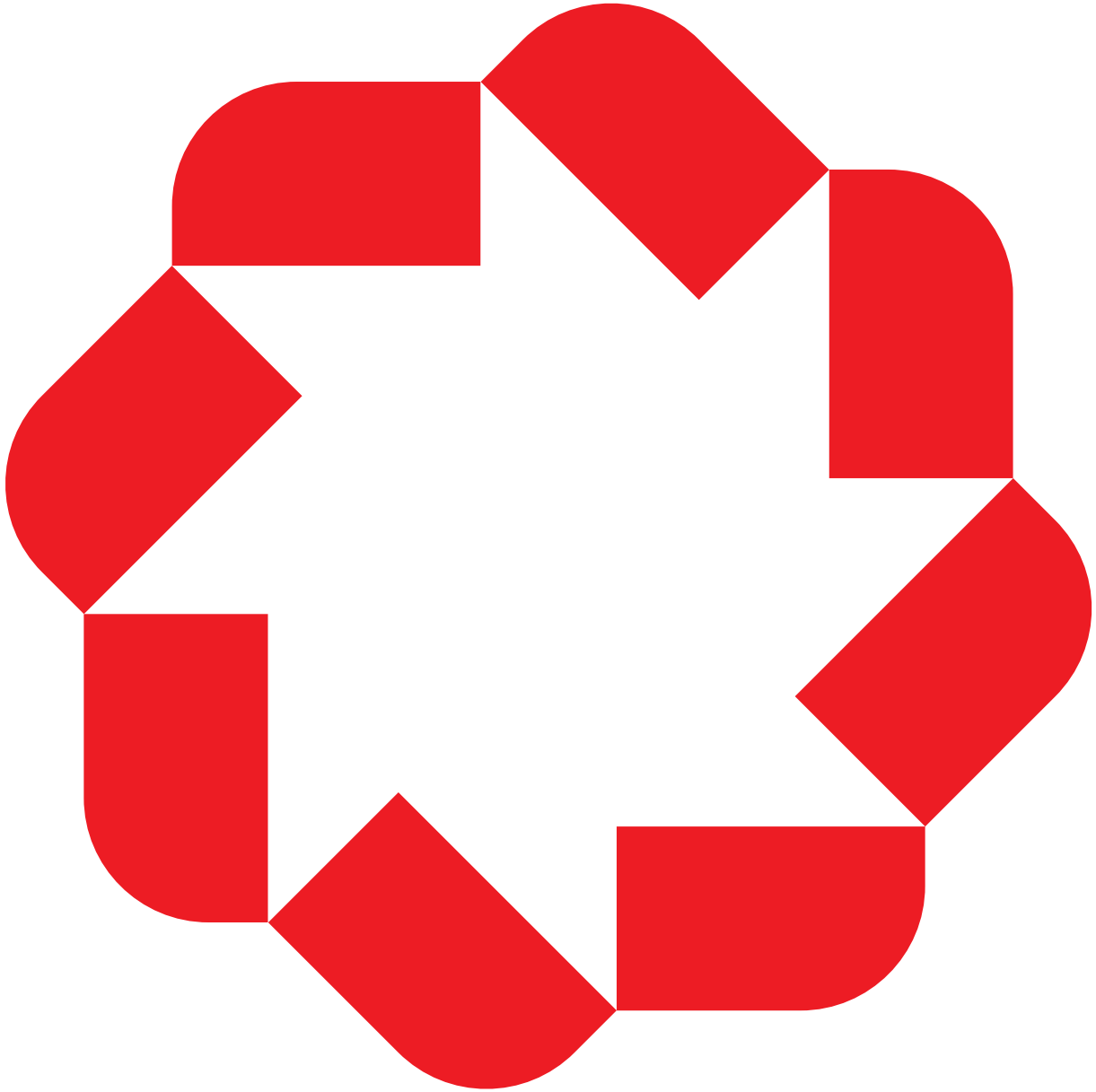
Acknowledgements

Universidad Rey Juan Carlos

SPECTRAL COMPUTE

NVIDIA

EuroHPC Joint Undertaking



Thanks for the attention!

engys
Innovation through CFD

Stefano Oliani
Senior Core Developer
s.oliani@engys.com

Disclaimer