Accelerating the FlowSimulator:

Mesh Deformation Performance Enhancement through Mixed Precisions

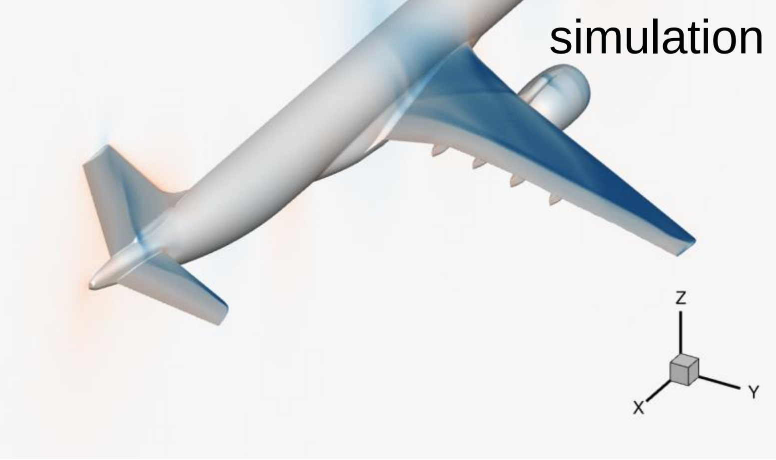
Cristofaro, Wendler, Huismann, Rempke

German Aerospace Center (DLR) – Dresden

Motivation

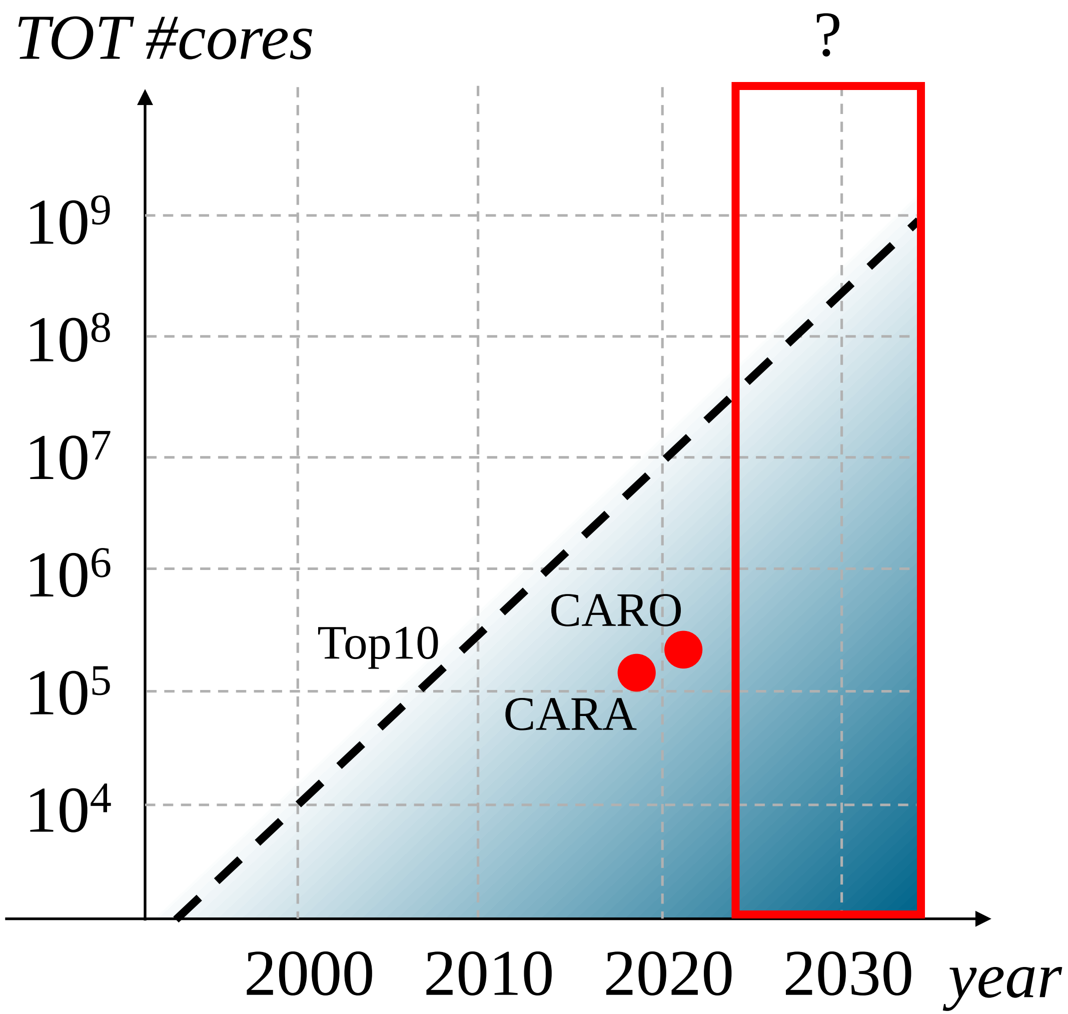
- Simulations with acceptable accuracy and performance *may* replace costly testing in the aeronautical industry

- Aeroelastic problems can be modelled with fluid-structure interaction simulations:
  - CFD solver
  - CSM solver
  - Interpolation
  - Mesh deformation



- **High-performance computing** can be exploited to reach **acceptable time-to-solution**

Trend in HPC computational resources



Increase in resources

↓

shorter time-to-solution

&

larger simulations

2006 A380: ~50 10^6 elements*

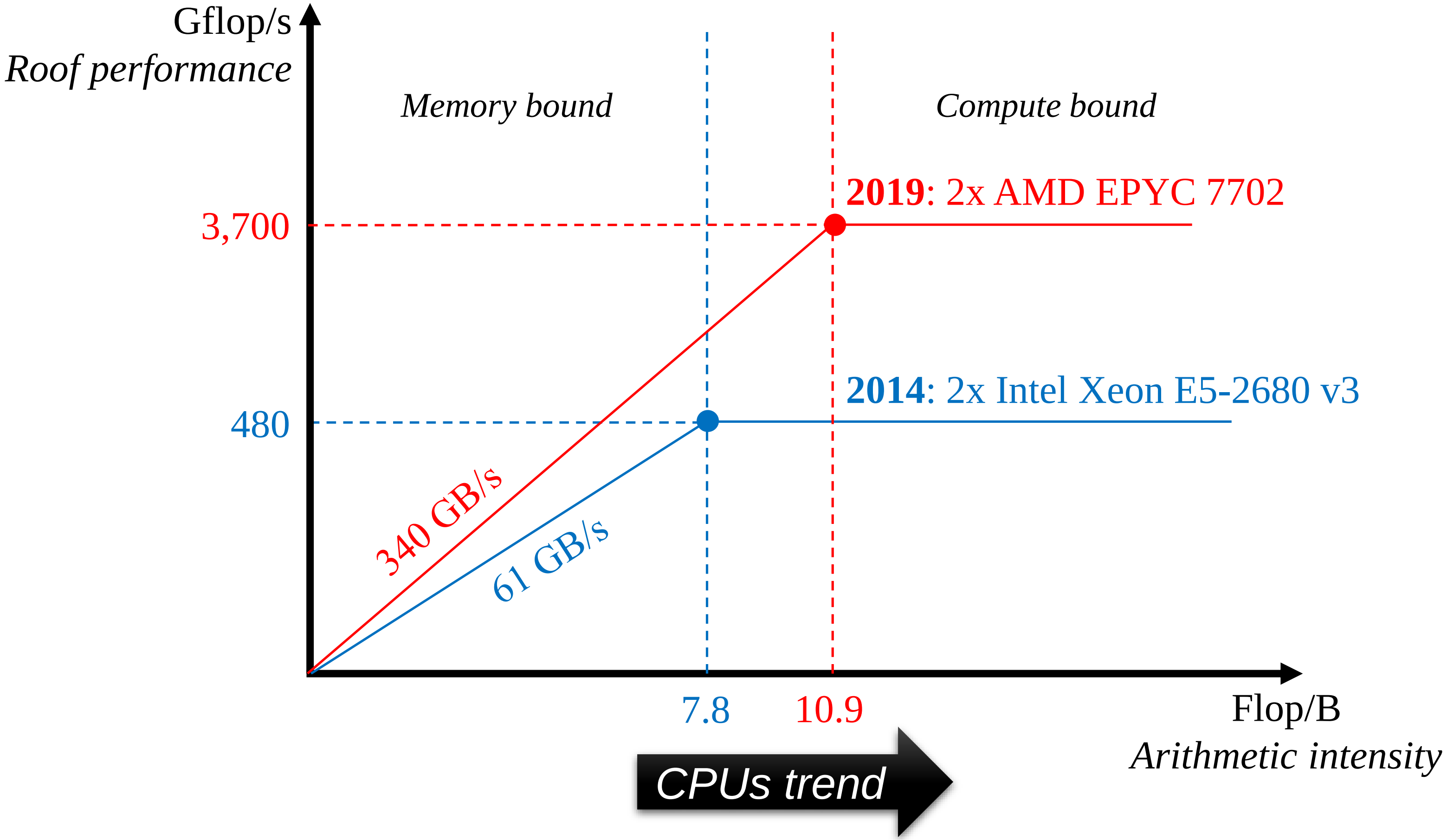
*2022 HLPW4**: ~700 10^6 elements*

BUT WE NEED **SCALABLE SOFTWARE!**

**The DLR TAU-code: recent applications in research and industry. D. Schwammborn et. al., ECCOMAS 2006*

***https://commonresearchmodel.larc.nasa.gov/*

Roofline model and trend in CPUs



Same algorithm with 9 Flop/B
(72 Flop/double):
- Compute bound in 2014
- Memory bound in 2019

More algorithms become memory bound!

We can boost performances
by increasing the arithmetic intensity

Possible solution
Reduce RAM memory access
for floats:
mixed precision

Measurement platform

DLR HPC System: CARO

- 174,592 cores
- #135 Top500 (11/2021)
- Göttingen (DE)
- each node:
  - 2x AMD EPYC 7702 (64 cores)
  - RAM: 256 GB DDR4
  - 16 cores per NUMA domain
  - 16 MB L3 cache shared among 4 cores

CARO node

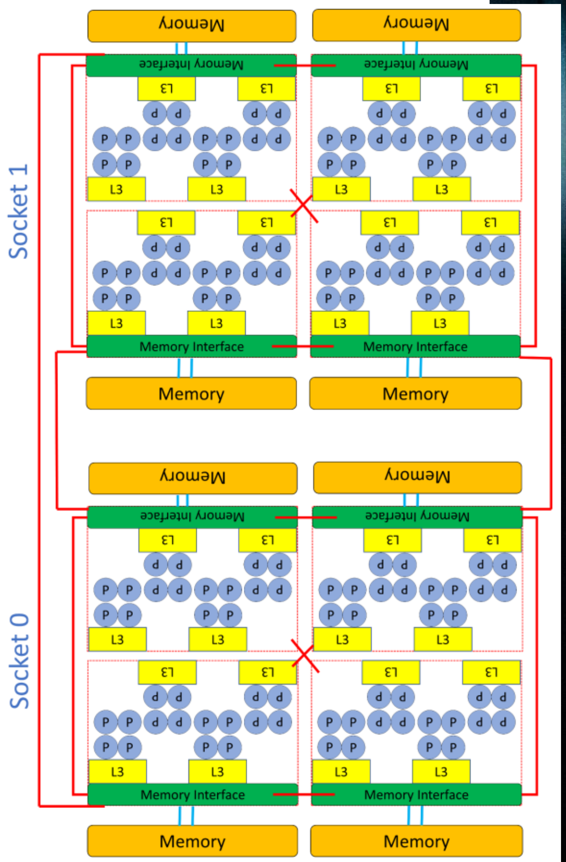


Image by Hirav Patel

Simulation enviroment

FlowSimulator

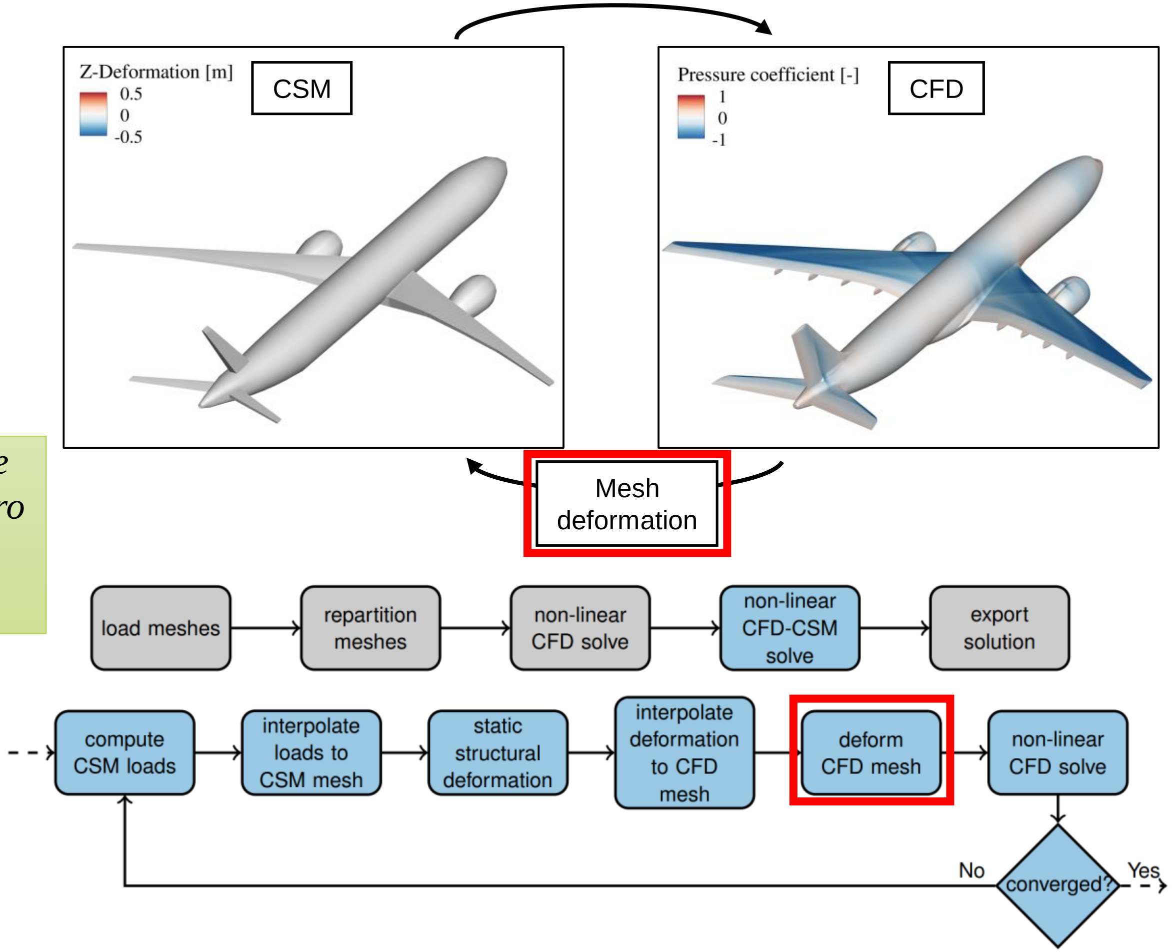
- simulation environment
- cooperation of DLR & AIRBUS

integrates:

- CFD solvers
  - TAU, CODA, Trace, HYDRA
- CSM solvers
  - Nastran, b2000++
- linear solvers
  - PETSc, Spliss
- utility components
  - e.g. mesh deformation
- predefined simulation toolchains
  - e.g. FSI

Node-level performance analysis of the structural mechanics solver b2000++pro
Ebrahimi Pour, Klimach
Presentation later in this session

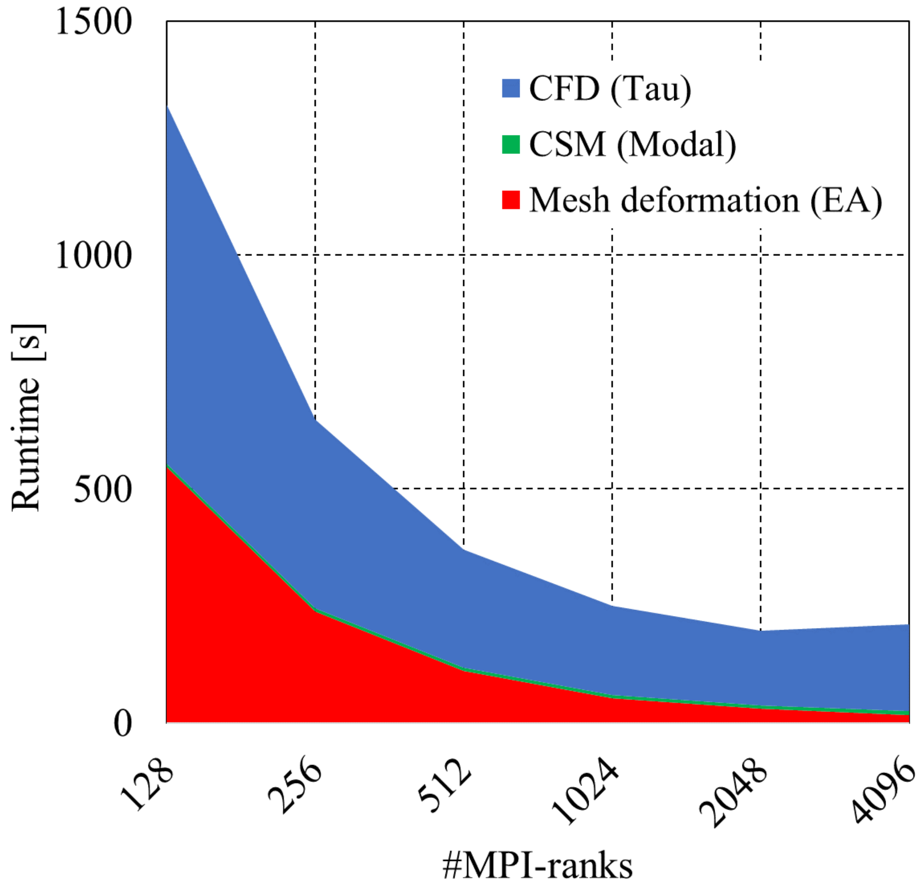
Steady aeroelastic simulation



Huismann et al. "Accelerating the FlowSimulator: Profiling and scalability analysis of an industrial-grade CFD-CSM toolchain," Coupled 2021

Strong scaling of steady aeroelastic simulation

LANN wing (CFD mesh: 1.2 10^6 nodes, CSM mesh: 1,260 nodes)



Mesh deformation (elastic analogy):

- up to 40% of total runtime
- > 80% runtime spent in linear solver
- good test bench also for CFD

Cristofaro, et al. "Accelerating the FlowSimulator: Improvements in FSI simulations for the HPC exploitation at industrial level," Coupled 2023

Mixed precision in linear system solvers

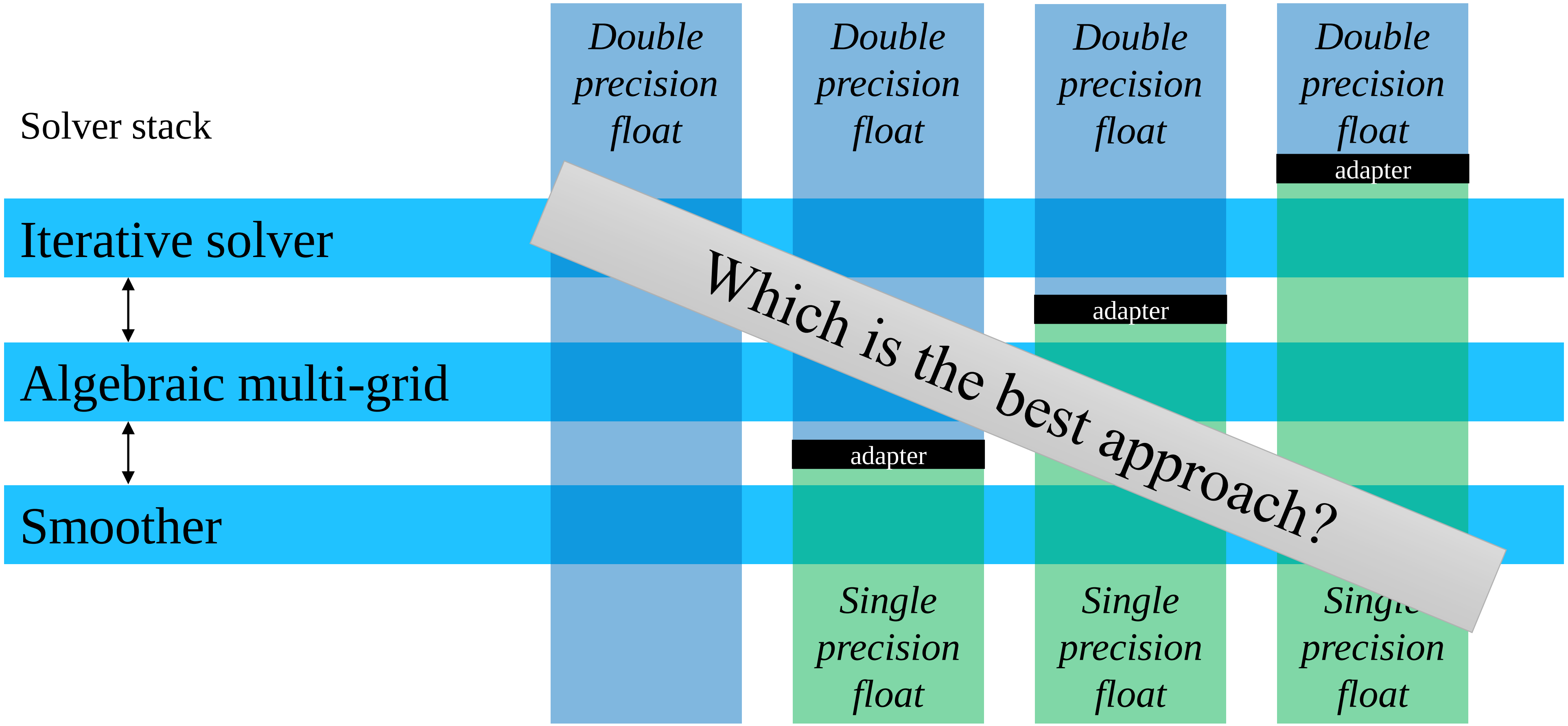
Mixed precision can **reduce the runtime** by **reducing the RAM access**:

- use of different precision levels (single, double) within linear system solver
- adapter casts values between precision levels
- improve performance by reducing memory footprint (8 to 4 Byte/float)
- ideal speed-up up to x2 for memory bound
- beneficial to most time consuming simulation blocks
- convergence rate may be affected by lower precision

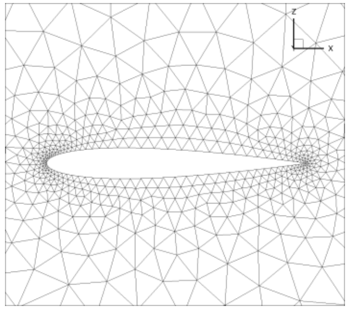
- Recently implemented in Spliss:
  - DLR Sparse Linear System Solver Library
  - used within CFD solver CODA

Accelerating the FlowSimulator: Mixed Precision Linear Solvers in Industrial Grade CFD
Wendler et al.
Fri, 07/06/2024, 10:30 - 12:30, Room 3A

Mixed precision approaches

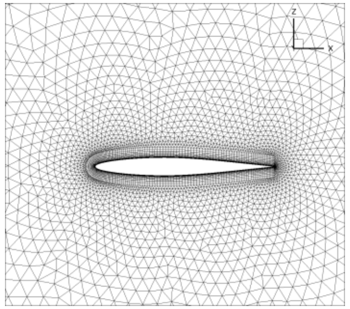


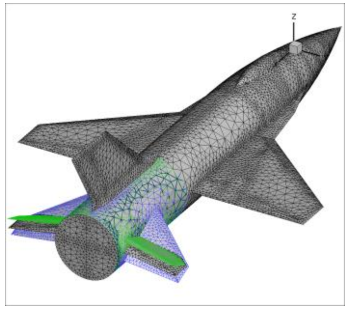
Test cases



naca0012
2k vertices

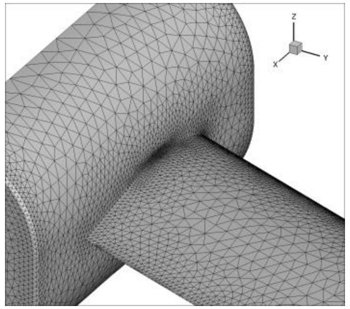
naca64A010
21k vertices

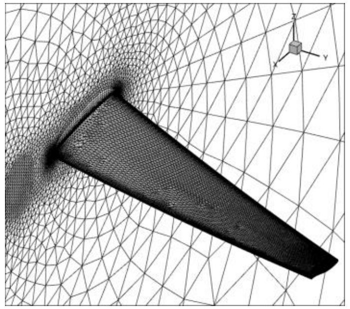




SDM
60k vertices

wing_body
231k vertices





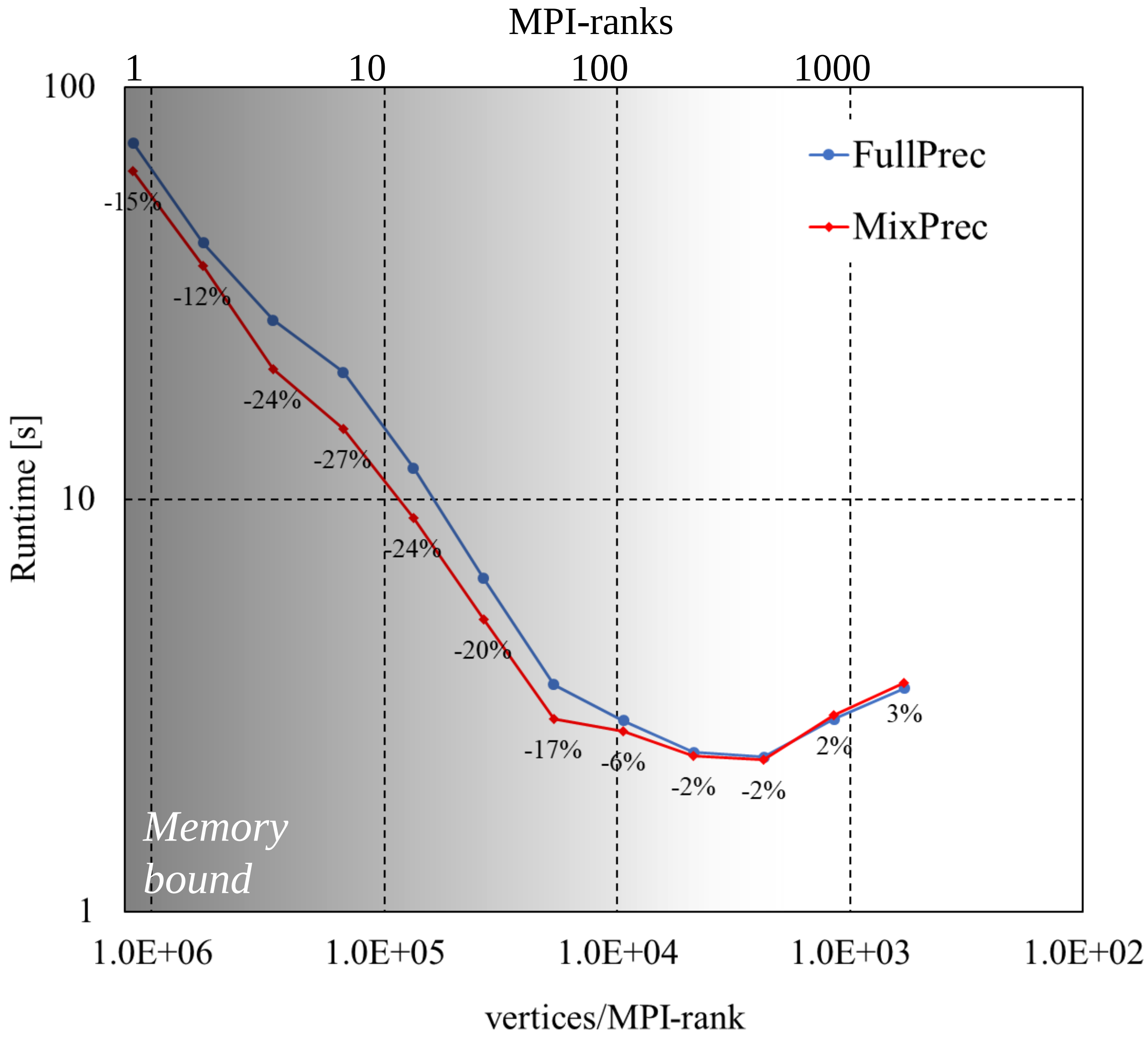
LANN
1.2M vertices

XRF1
30M vertices

- Meshes: 2k – 30M vertices
- Parallelization: 100 – 1M vertices / MPI-rank
- Computing resources: 1 – 600 MPI-ranks (x 4 OpenMP threads)
- Solvers (Spliss library):
  - GmRes / BiCGStab + MG + J / GS + LI / BI
- Runtimes: 0.2 s – 6 minutes
- TOTAL simulations: 128

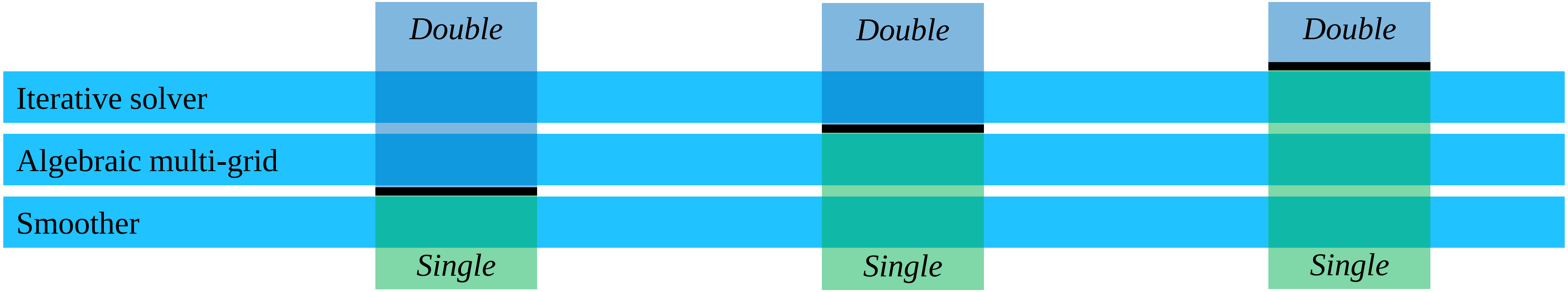
Mixed precision results - example

Test case: LANN – GMRes-MG-GS-Lines



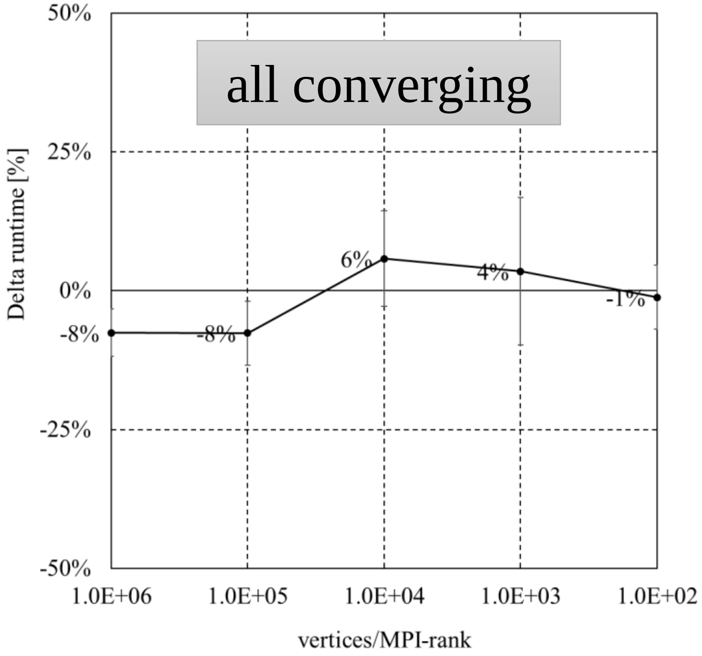
Same analyses done for all cases and runtime reduction is then averaged

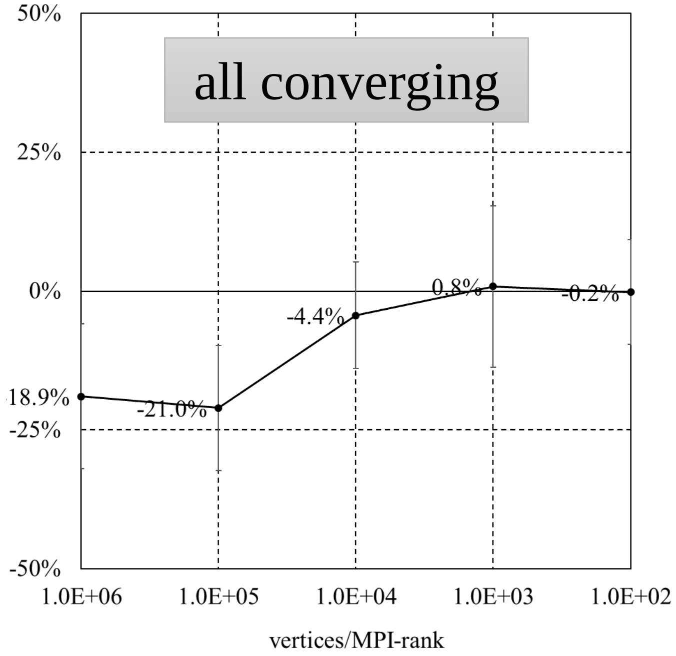
Mixed precision approaches comparison

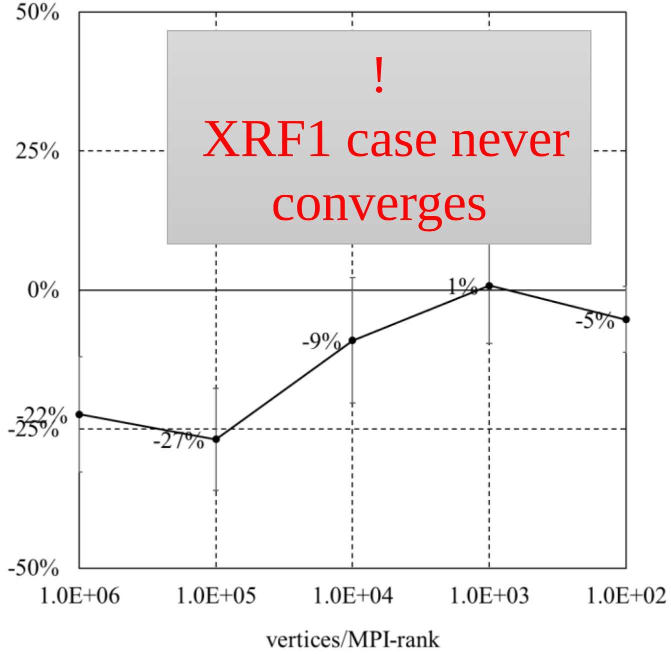


Runtime difference
wrt full prec

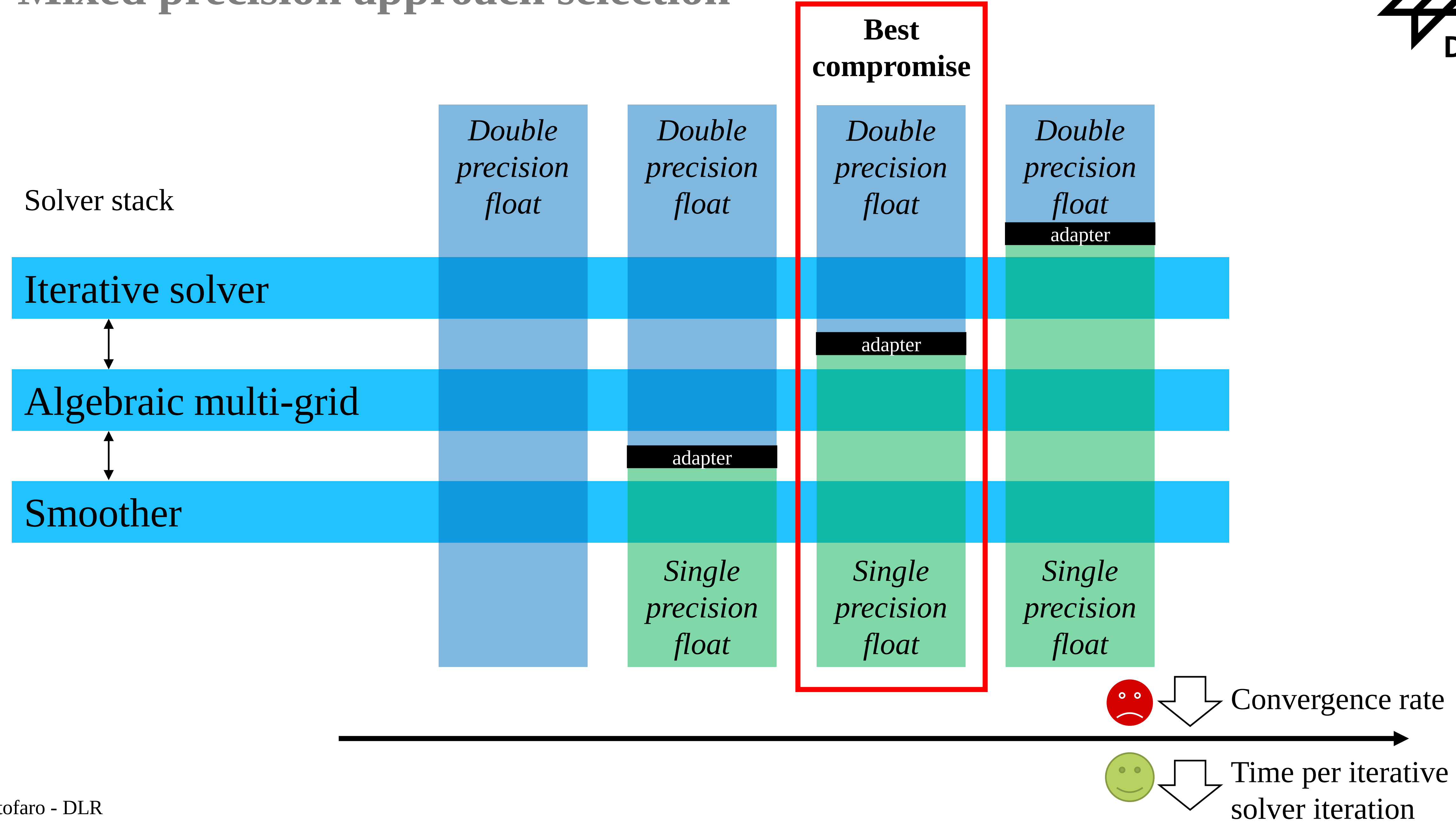
Results averaged over 124 cases



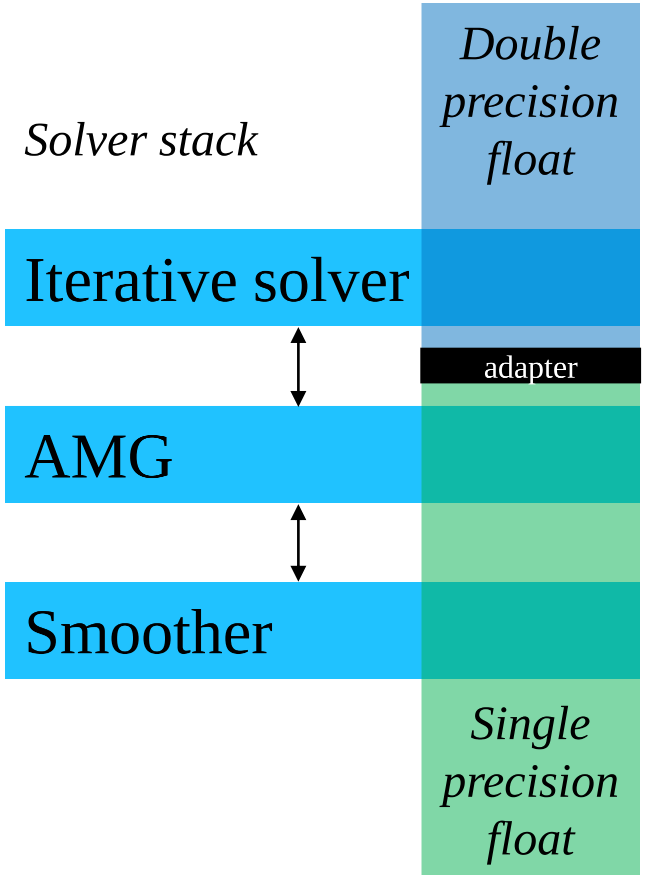




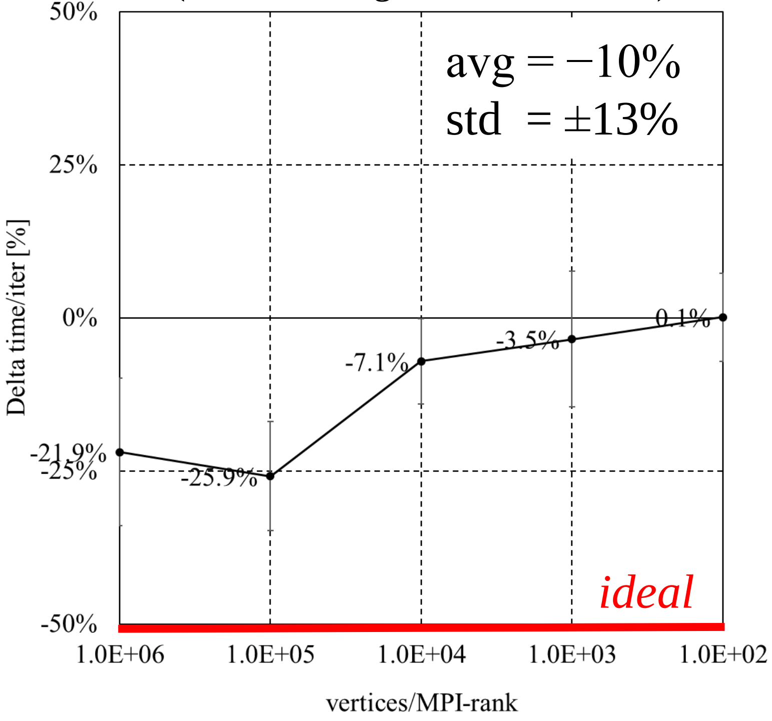
Mixed precision approach selection



Mixed precision time per iteration and #iterations

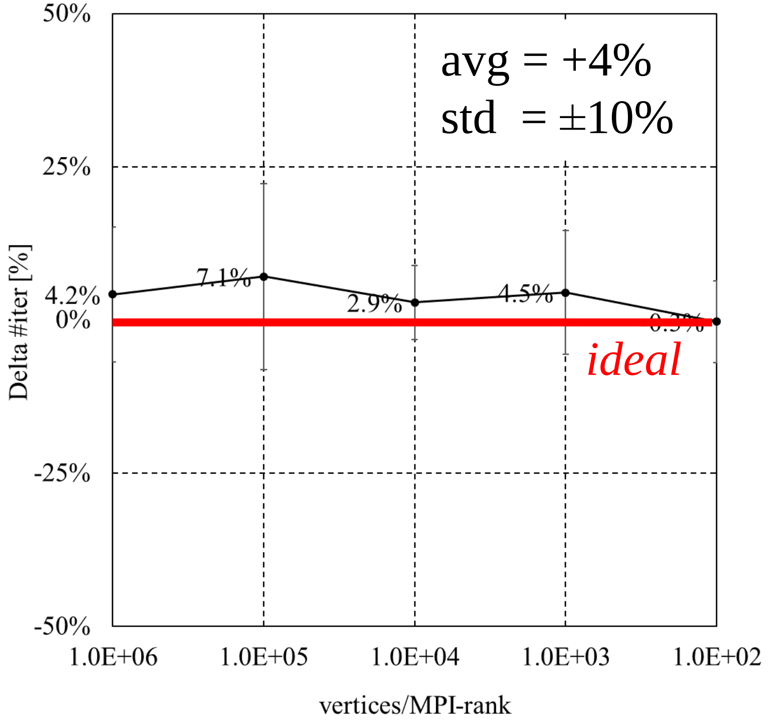


Time/iter difference
wrt full prec
(results averaged over 124 cases)



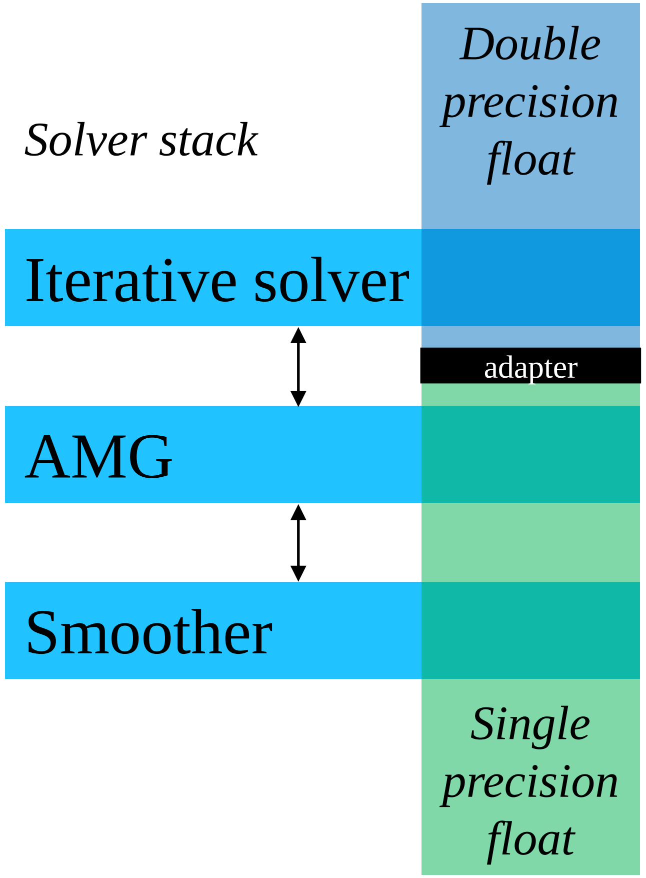
Time per iteration benefit:
- up to 26%
- negligible for low parallelization levels

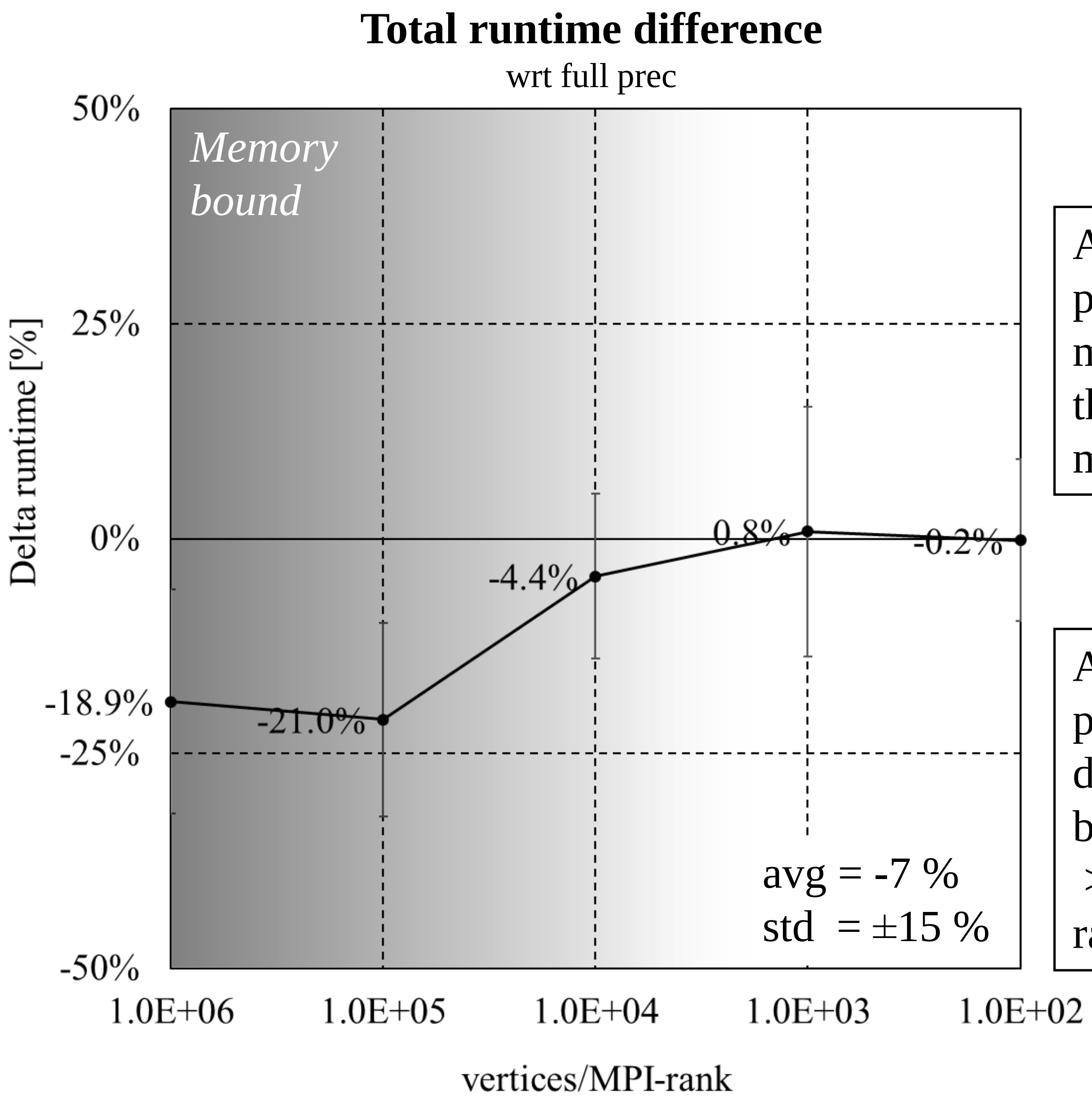
#iter difference
wrt full prec
(results averaged over 124 cases)



Slower convergence rate leads to +4% more iterations needed on average

Mixed precision runtime benefit





Advantage of mixed precision is to reduce memory access, thus beneficial only in memory bound regions

Application of mixed precision to mesh deformation is mostly beneficial for > 10,000 vertices / MPI-rank

Conclusion

- Take advantage from HPC for fast fluid-structure interaction simulations
  - mesh deformation is critical component
  - good test bench also for CFD (same linear solver)
- Trend in HPC hardware
  - ridge point moves towards higher arithmetic intensity
  - performance more often limited by RAM memory bandwidth (memory bound)
- Mixed precision
  - implemented in Spliss (DLR linear solver library)
  - reduces memory requirements of floats by 50%
  - best compromise: **mixed precision between iterative solver and AMG**
    - ~20% runtime benefit for low parallelization levels (memory bound)
    - negligible difference for large parallelization levels (MPI-comm bound)

Q&A

Back-up slides

Supercomputers at the German aerospace center

CARA

- 145,920 cores
- #221 Top500 (11/2019)
- Dresden

CARO

- 174,592 cores
- #135 Top500 (11/2021)
- Göttingen

DLR

↓

Software methods
&
Product virtualization

↓

High performance computing

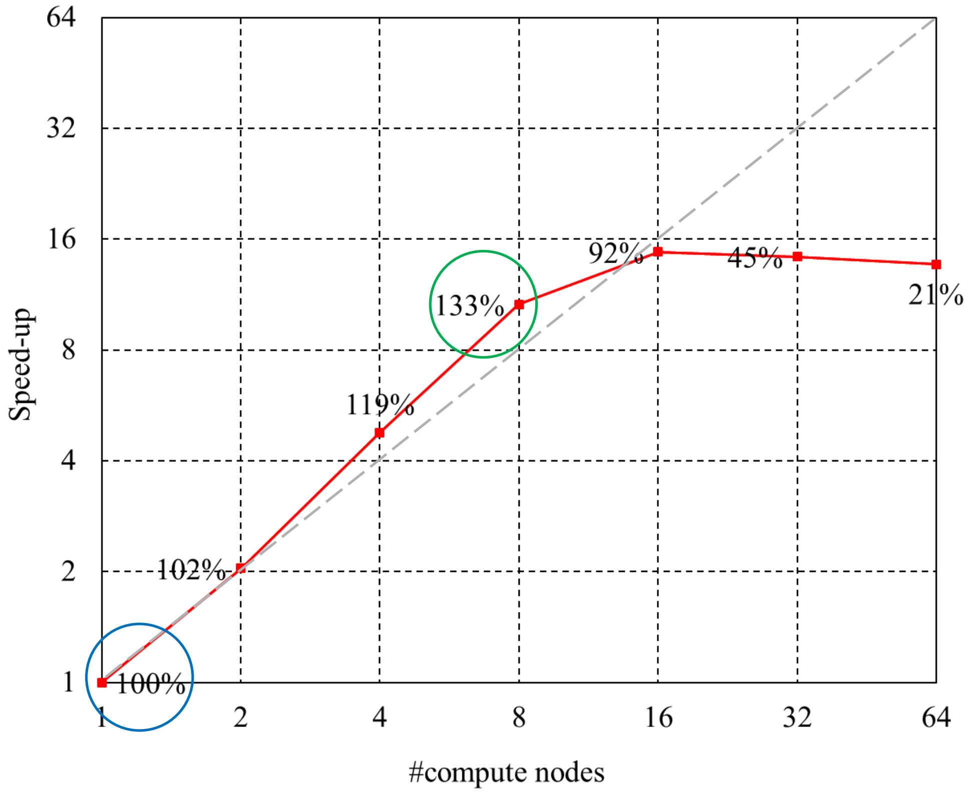
Superlinear scaling

Mesh deformation:

memory bound (RAM bandwidth limits execution speed)*

→ when problem fits in L3 cache (i.e. small MPI-domains)

→ RAM access reduced → execution becomes much faster



- 1 compute node: 54.7 ms/iter
  32 MPI-ranks, 40,000 vertices/MPI-rank

- 8 compute nodes: 5.1 ms/iter
  256 MPI-ranks, 5,000 vertices/MPI-rank

Computing resources x8 → speed-up x11

*Ebrahimi Pour, Cristofaro et al, "Accelerating the FlowSimulator: Performance Analysis of Finite Element Methods on High–Performance Computers", IPTW 2023

Superlinear scaling considerations

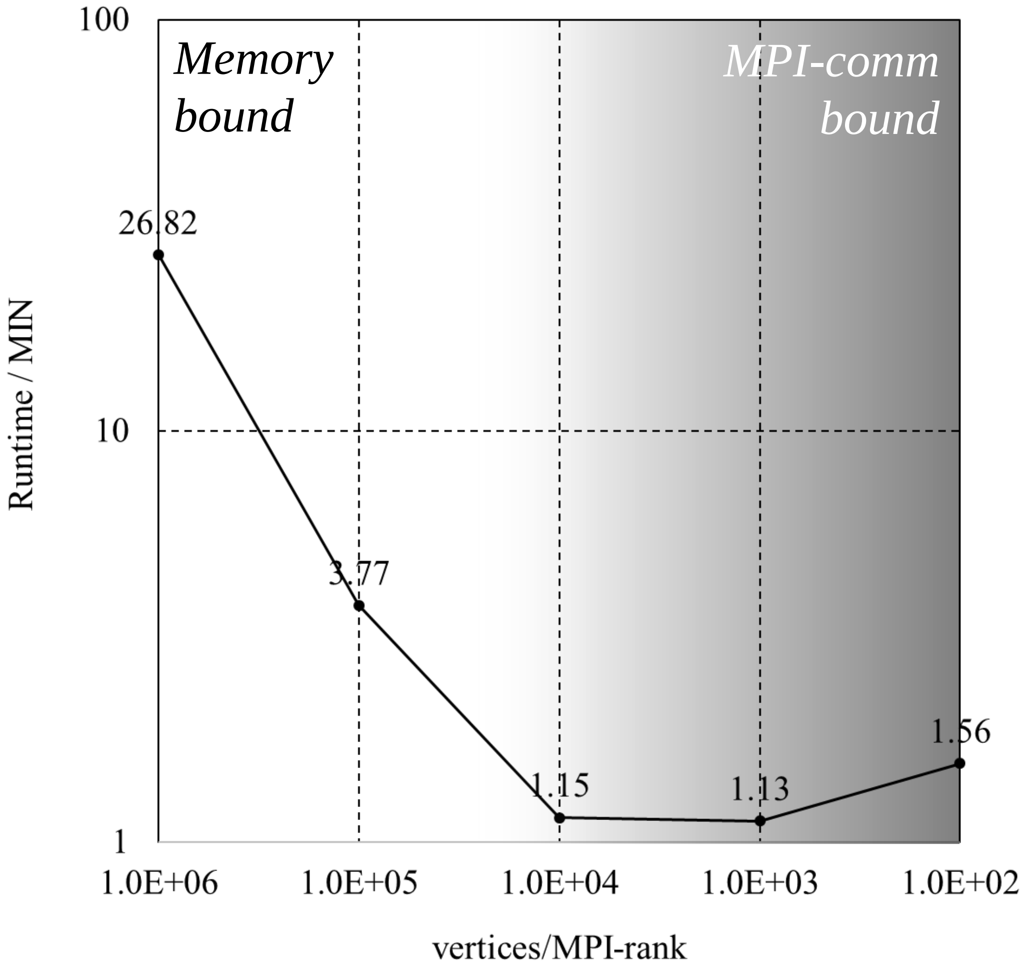
Superlinear scaling has **zero implementation cost**

but it is **hard to achieve**:

- problem size depends on settings (e.g. #eqs in CFD)
- differs between simulation blocks (e.g. CFD, CSM, Mesh Deformation)
- may be close to sudden drop in performances due to MPI-comm overhead
- may need large computing resources
- Strongly related to hardware architecture and L3 cache size

Absolute scaling

The minimum runtime for each case is used to adimensionalize the results
the scaled results are then averaged



Combined effects:

- mesh size
- mesh topology
- number of linear solver iterations
- number of MPI-ranks
- number of compute nodes

general indication of runtime sweet spot:
1,000 - 10,000 vertices/MPI-rank

- same region as superlinear scaling, problem fits in L3 cache
- MPI-comm overhead still acceptable