Developments for CAE Applications on GPUs and Arm CPUs

Stan Posey, Program Manager, CFD Domain, NVIDIA

DOE HPC4EI Workshop
17-18 Oct 2023

TOPICS OF DISCUSSION

- HPC for Mfg Updates
- Applications in CAE

Exascale AI Systems apply CAE Workloads

NVIDIA (EOS)
18 EFLOPS AI Perf
Quantum-2 400Gb/s InfiniBand

LANL (Venado)
10 EFLOPS AI Perf

BSC (MareNostrum 5)
18 EFLOPS AI Perf
Quantum-2 400Gb/s InfiniBand

CSCS (ALPS)
20 EFLOPS AI Perf

CINECA (Leonardo)
10 EFLOPS

KAUST (Shaheen-III)
7 EFLOPS AI Perf

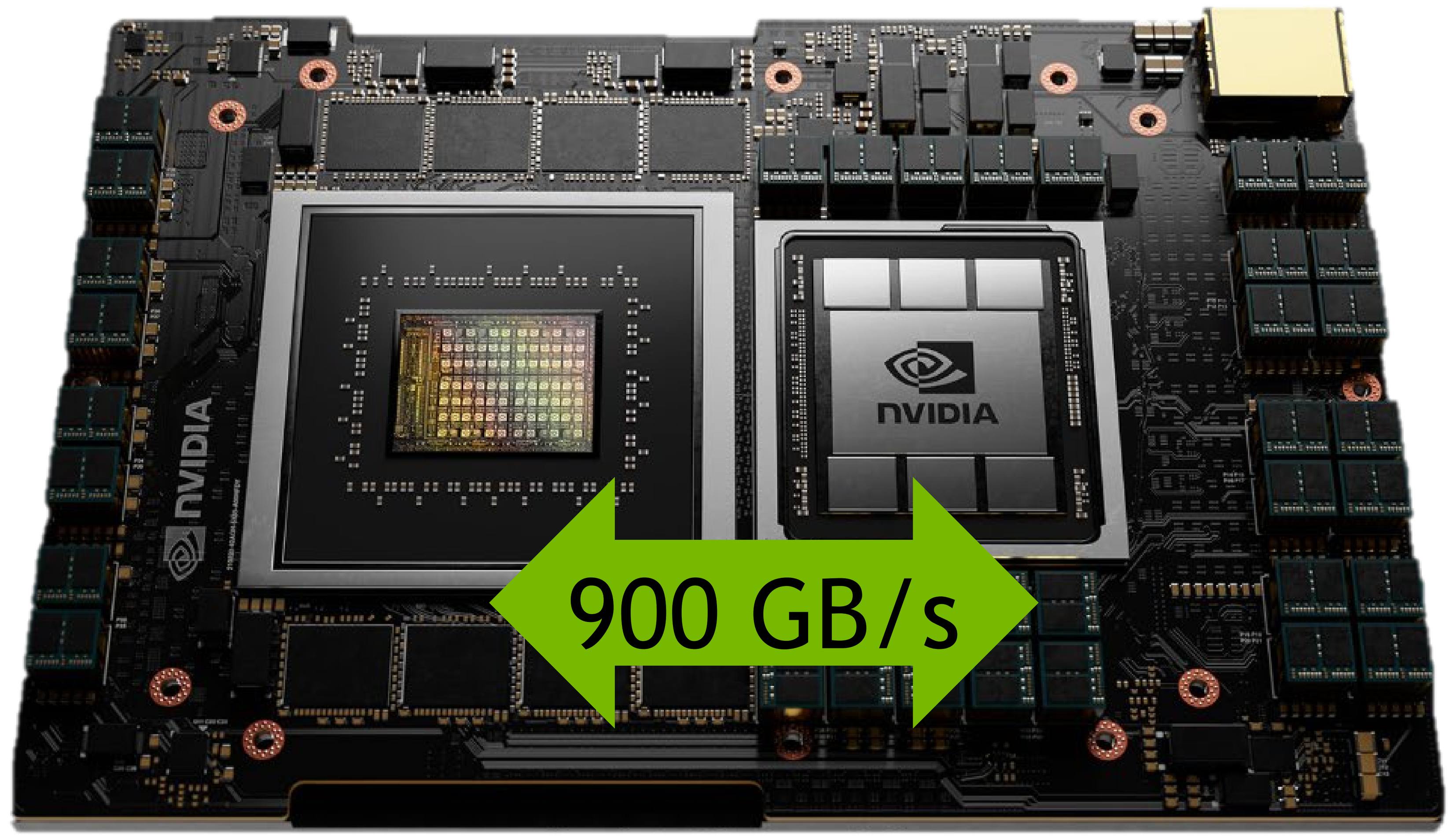
Feature Progression of NVIDIA GPU Architectures

		H100 (2023)	A100 (2020)	V100 (2017)
Peak FP64 TF/s	**High order dense ops**	**26 / 34** 3x	9.7 1.3x	7.5
Peak FP64 TC TF/s		**51 / 67**	19.5	
Peak FP32 TFlop/s		**51 / 67** 3x	19.5	15.0
Peak TF32 TC TF/s	**AI-based methods**	**756 / 989**	156	
Peak FP16 TFlop/s		**1513 / 1979** 6.4x	312 2.6x	120
Memory BW (GB/s)	**2nd order sparse ops**	**2000 / 3350** 1.6x	1555 / 2038 2.3x	900
Memory Capacity (GB)		**96**	40 / 80 2.5x	16 / 32
Interconnect		**NVLink: Up to 900 GB/s** 2.0x PCIe – G5: 128 GB/s	NVLink: Up to 600 GB/s 2.0x PCIe: 64 GB/s	NVLink: Up to 300 GB/s PCIe: 32 GB/s
Max Power (W)		**300 – 700**	400	250 - 300

NVIDIA Next-gen GPU H100 and Arm CPU "Grace"

Breakthrough Designs for Large-Scale HPC and AI Applications

Grace Arm + H100 (Hopper)



Available
Q3 2023

Grace Arm-only Node

GRACE PERFORMANCE: Superchip Design with 144 high-performance Armv9 Cores

GRACE MEMORY BANDWIDTH: 480GB of LPDDR5x memory with ECC, 500 MB/s memory bandwidth

GRACE INTERCONNECT: NVLink-C2C with 900 GB/s bandwidth coherent connection to CPU or GPU

HIGHEST ENERGY EFFICIENCY: 2X Perf/Watt v. conventional servers, CPU cores + memory in 500W

CAE Accelerated with Modulus and Physics-ML

Modulus open-source platform for developing physics-based machine learning models

Neural acceleration of CAE solvers

Neural acceleration of CAE post-processing

ISVs integrate pre-trained models into their Design optimization SW

- More simulation data generation using Solver → More accurate and greater design space coverage with AI surrogate model for a better Design optimization SW

https://developer.nvidia.com/modulus

TOPICS OF DISCUSSION

- NVIDIA HPC Updates
- **Applications in CAE**

Select Collaborations for GPU Accelerated CAE

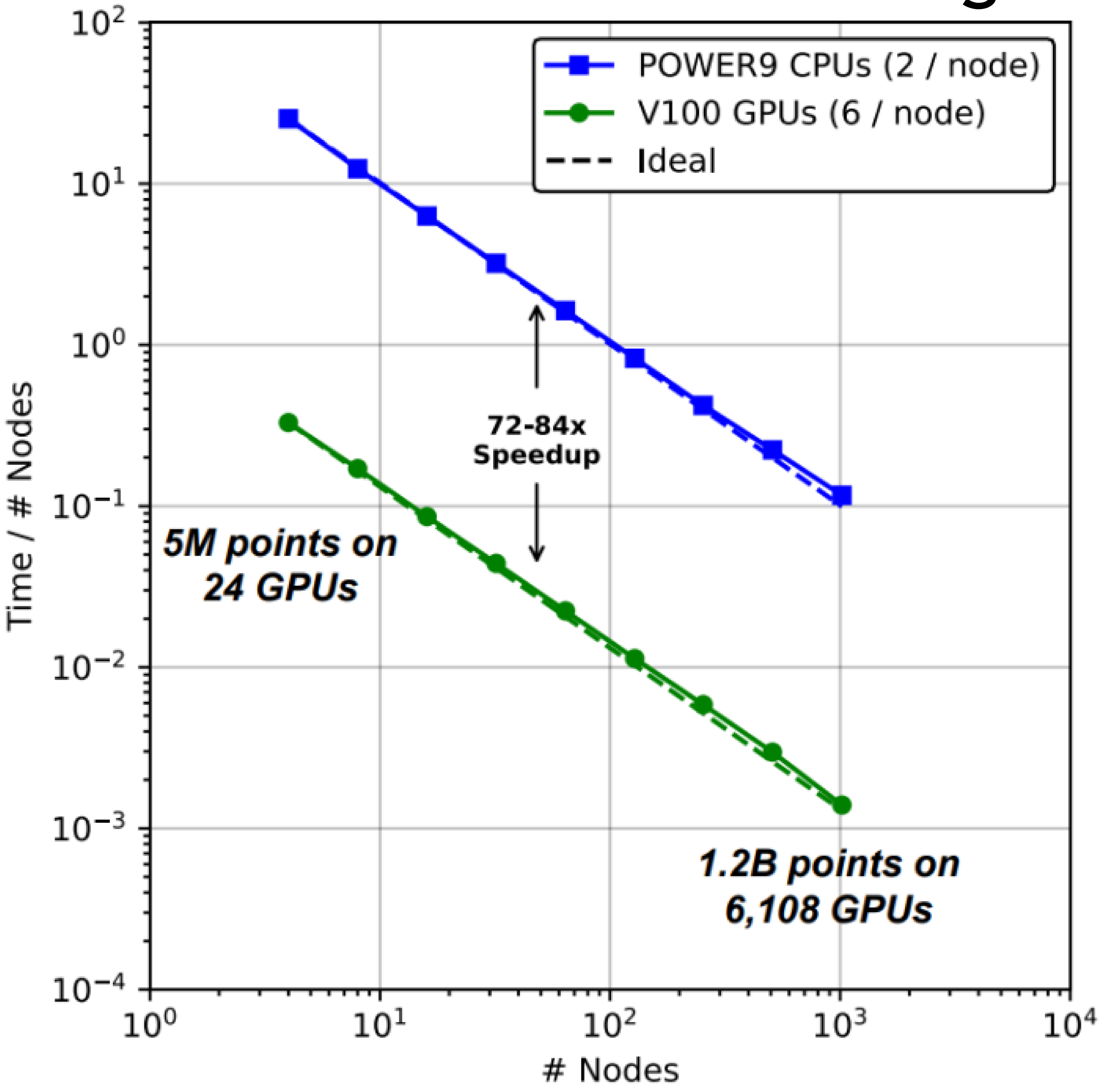
	Org	Software	Method	GPU Features
Community:				
	NASA	FUN3D	CUDA	Full application, RG chemistry
	NASA	OVERFLOW	OpenACC	Central diff schemes, Euler flux, smoothing
	DoD NRL	JENRE	CUDA	Full application, high-order FE LES
	DOE ANL	NekRS	OCCA (CUDA)	Full application, high-order spectral element
	ESI-OpenCFD	OpenFOAM	CUDA Lib	Linear solver only using AmgX lib
Custom:				
	Boeing	BCFD	OpenACC	Full application, 2^{nd} order FV RANS
	GE	GENESIS	CUDA	Full application, high-order LES
Commercial:				
	ANSYS	Fluent	CUDA	Full application, core features
	Siemens	STAR-CCM+	CUDA	Full application, core features
	Altair	ultraFluidX	CUDA	Full application LBM, core features
	Cadence	CharLES	CUDA	Full application, core features

FUN3D Applied to Mars Lander Simulations

2 x 64-core AMD 7742	1.0x
NVIDIA V100 32 GB	4.0x
NVIDIA A100 40 GB	7.0x

https://www.nvidia.com/en-us/on-demand/session/gtcspring22-s41286//

ORNL Summit Scaling

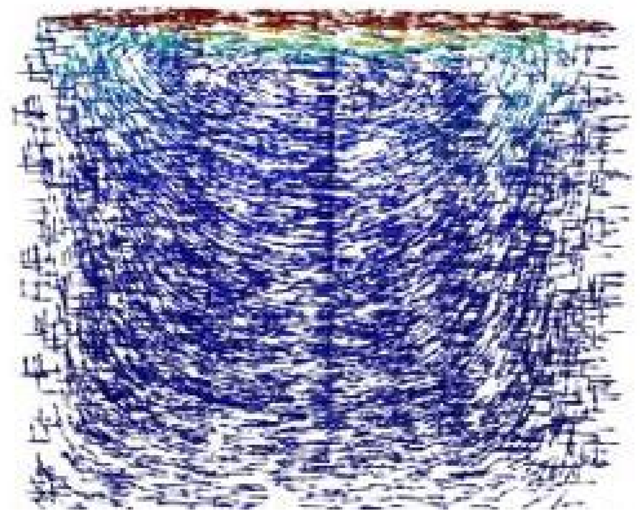


NVIDIA and OpenFOAM GPU Collaboration

- **OpenFOAM collaboration among members of the HPC Technical Committee**
- Contributions from ESI-OpenCFD, Leonardo, CINECA, and AWS
- GPU evaluations at DOE ORNL, General Motors, VW, others
- NVIDIA also member of the data-driven modeling SIG on mlfoam

- **GPU solution for standard OpenFOAM release – required no source changes**
- Linear solver GPU off-load using plug-in of external solver from NVIDIA AmgX lib
- External solvers possible from the introduction of PETSc4FOAM lib[1] by HPC TC

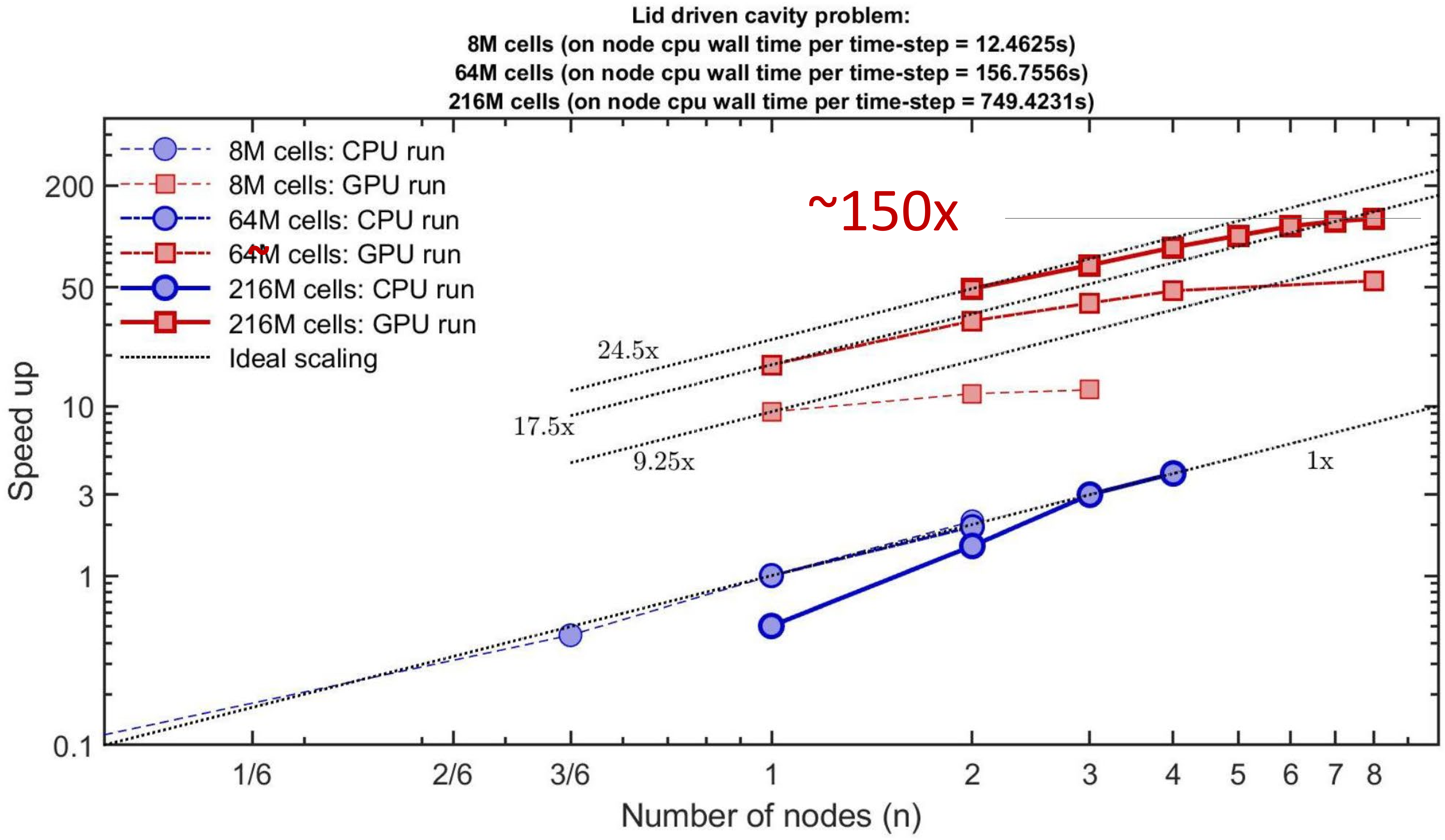
- **GPU Speedups:**
- Typical linear solve speedups (AMD Milan 64c + H100) **~8x**
    - Overall speedups limited by % of total time in linear solve
- ORNL Summit scaling: 216M cells on 8 GPU nodes (V100) **~150x**



[1] S. Bnà, I. Spisso, M. Olesen, G. Rossi *PETSc4FOAM: A Library to plug-in PETSc into the OpenFOAM Framework* PRACE White paper

OpenFOAM GPU Strong Scaling on ORNL Summit

Strong scaling on Summit



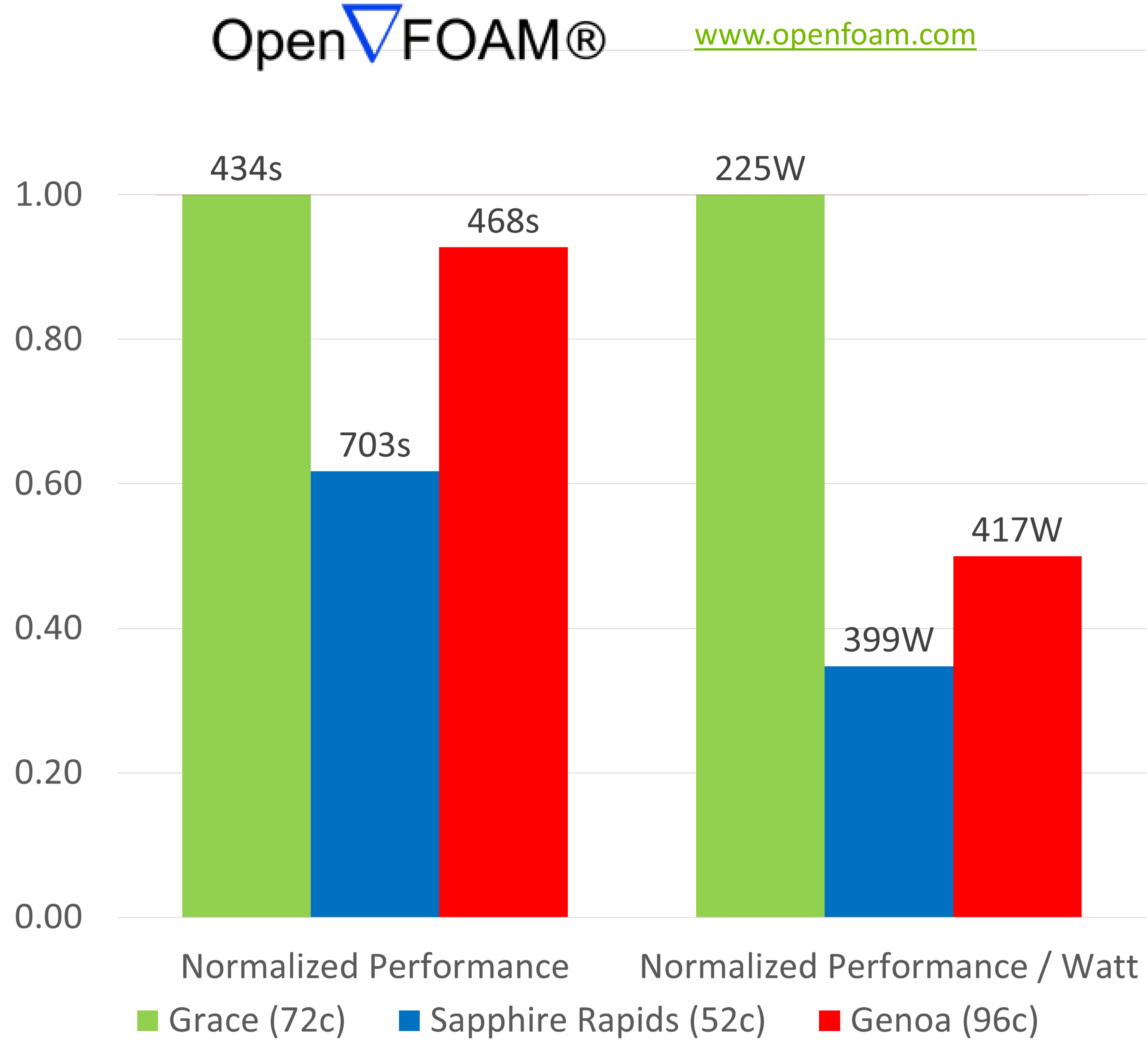
Running GPU-enabled OpenFOAM on Summit

FERMI Project

Dr. Arpan Sircar, Dr. Vittorio Badalassi

DOE Oak Ridge National Laboratory

OpenFOAM Performance with Grace Arm vs. x86



NOTE: All results single socket proccesor

- **Version:** OpenFOAM v2212
- **Compiler used:** GCC 12

(NVIDIA / AMD / Intel)

- **Compile options:** `WM_COMPILE_OPTION=Opt,`

`-march=native`

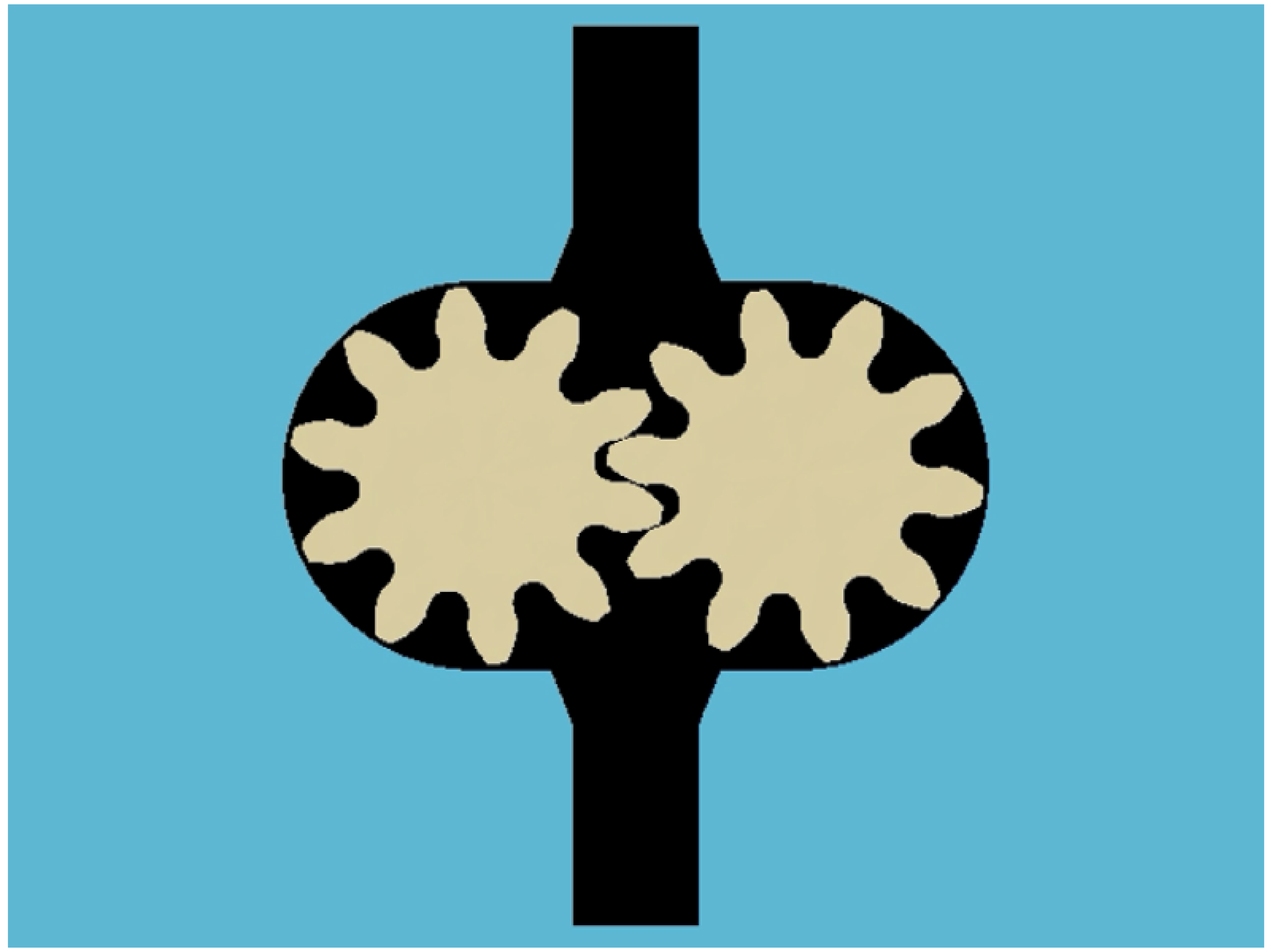
- **Benchmark details:** Motorbike Large (34M)
  - Simulates air flow around a complex unstructured geometry describing a motorcycle and rider
  - Configuration from:
  - https://develop.openfoam.com/Development/openfoam/-/blob/master/tutorials/incompressible/simpleFoam/motorBike/system/fvSolution
  - "GAMG" for the pressure solve and "smoothSolver" for momentum, k, and omega
  - 100 iterations

- **Run configuration:** 1 MPI process per core
  - `mpirun --bind-to core --map-by core -n $ncores simpleFoam -parallel`

Cadence CFD Speedups for GPU + Grace Arm

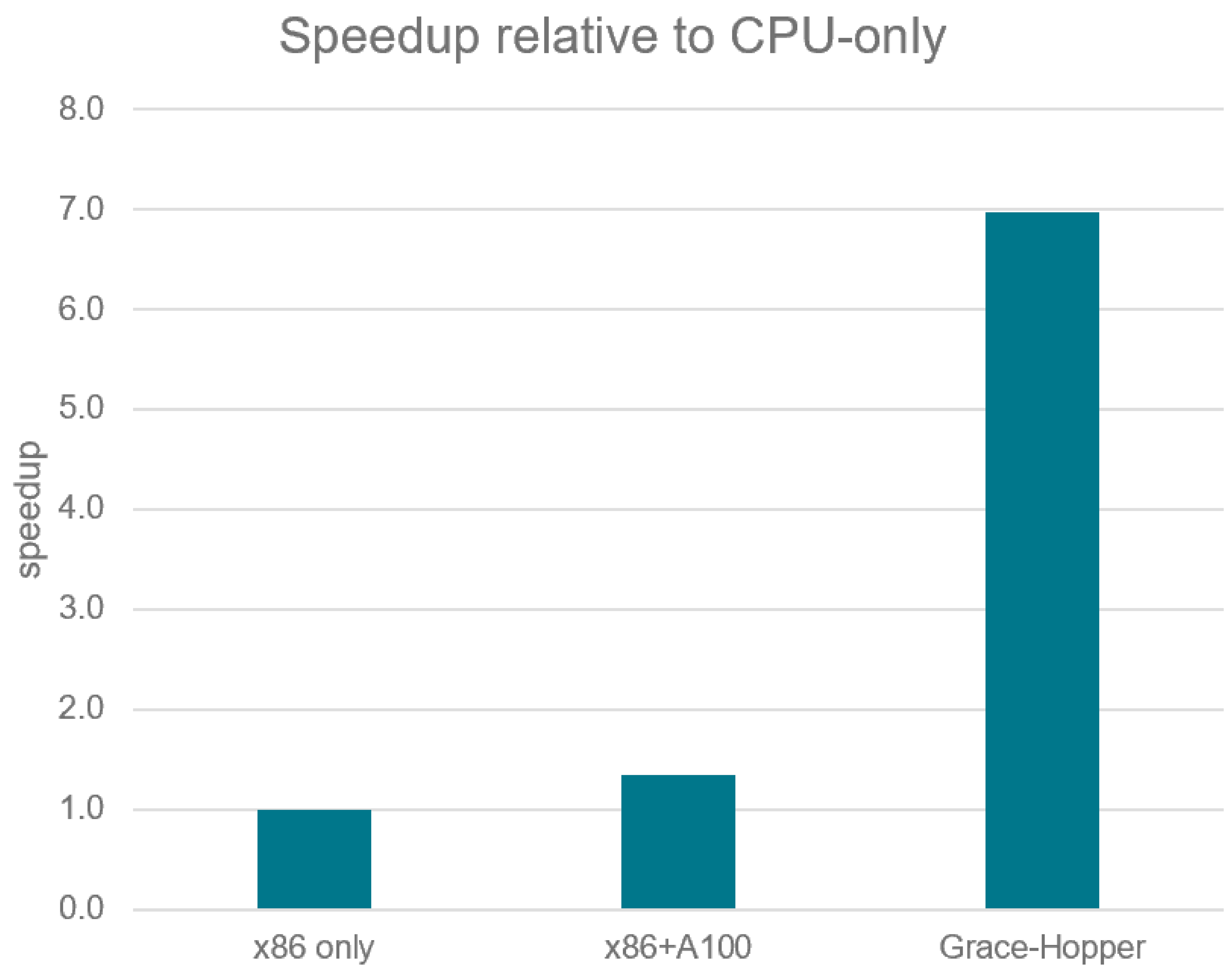
Example : Hybrid CPU/GPU moving solver Fidelity CharLES Grace-Hopper vs X86+A100

cadence®



Complex compressible moving geometry simulation of a gear pump

- conservative moving geometry algorithm
- dynamic data and dynamic connectivity
- Well suited to CPU implementation to build systems + GPU offload to advance solver



Simulation details:

- 2 million control volumes, 1 million of which are active at any given point
- x86 only: 16 cores AMD EPYC 75F3
- x86+A100 80GB, all 32 cores gave best timing
- GH: 72 cores + H100

NVIDIA

Trek Bicycle Apply CFD Workflow on NVIDIA GPUs

Up to 6x performance gain across product development

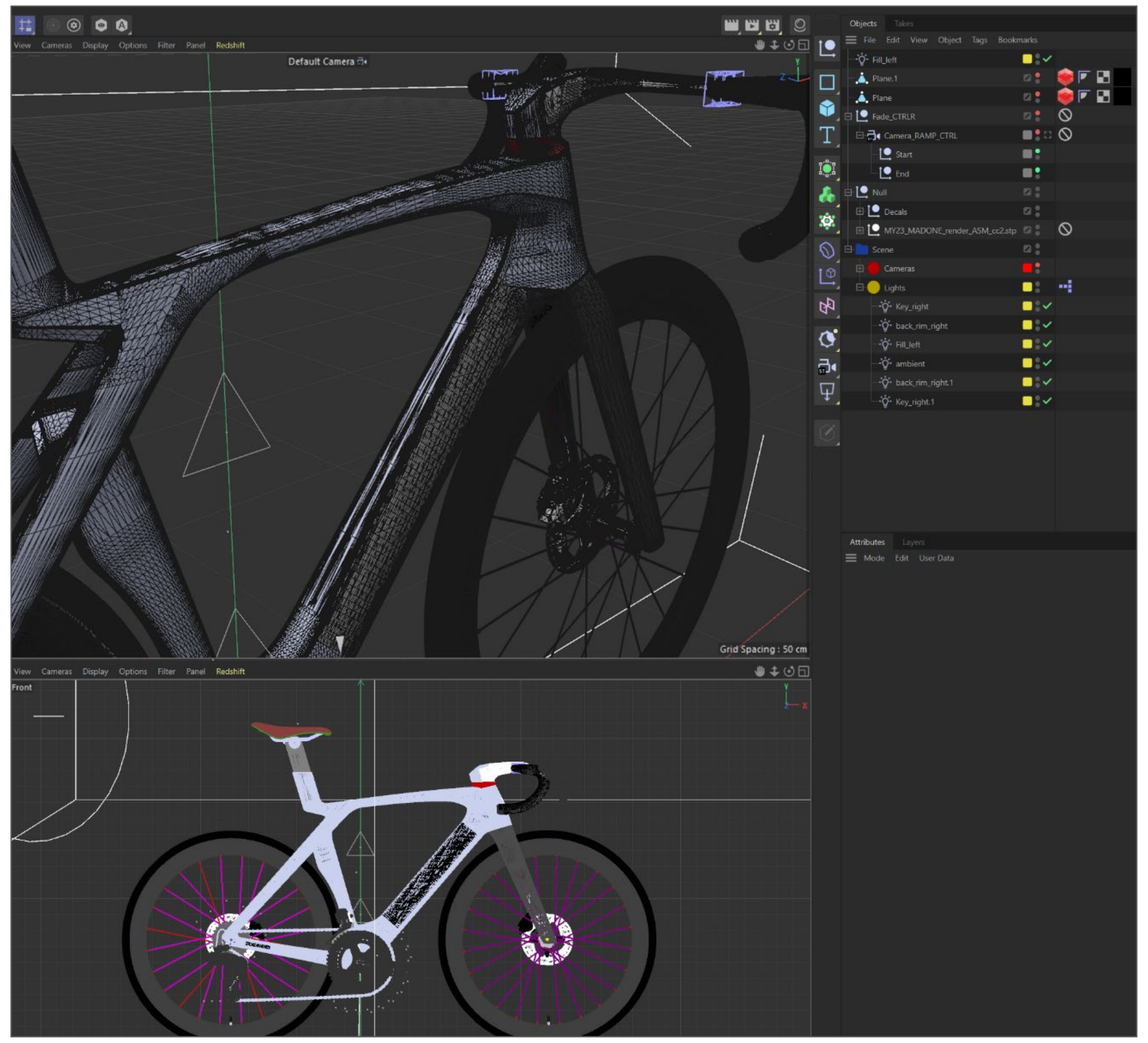


Image courtesy of Trek Bicycles

Preliminary results on previous generation hardware and pre-production hardware and software, final performance may vary.

2X Speedup in Design and Styling

3X Speedup in Engineering Simulations with GPU Accelerated Siemens Simcenter STAR-CCM+

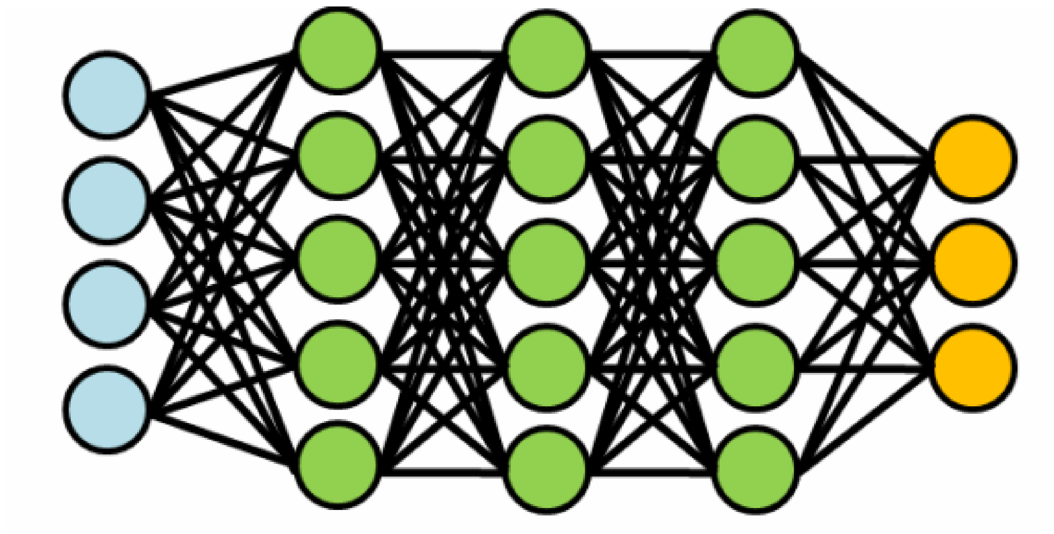
2X Speedup in Generating Photorealistic Renders

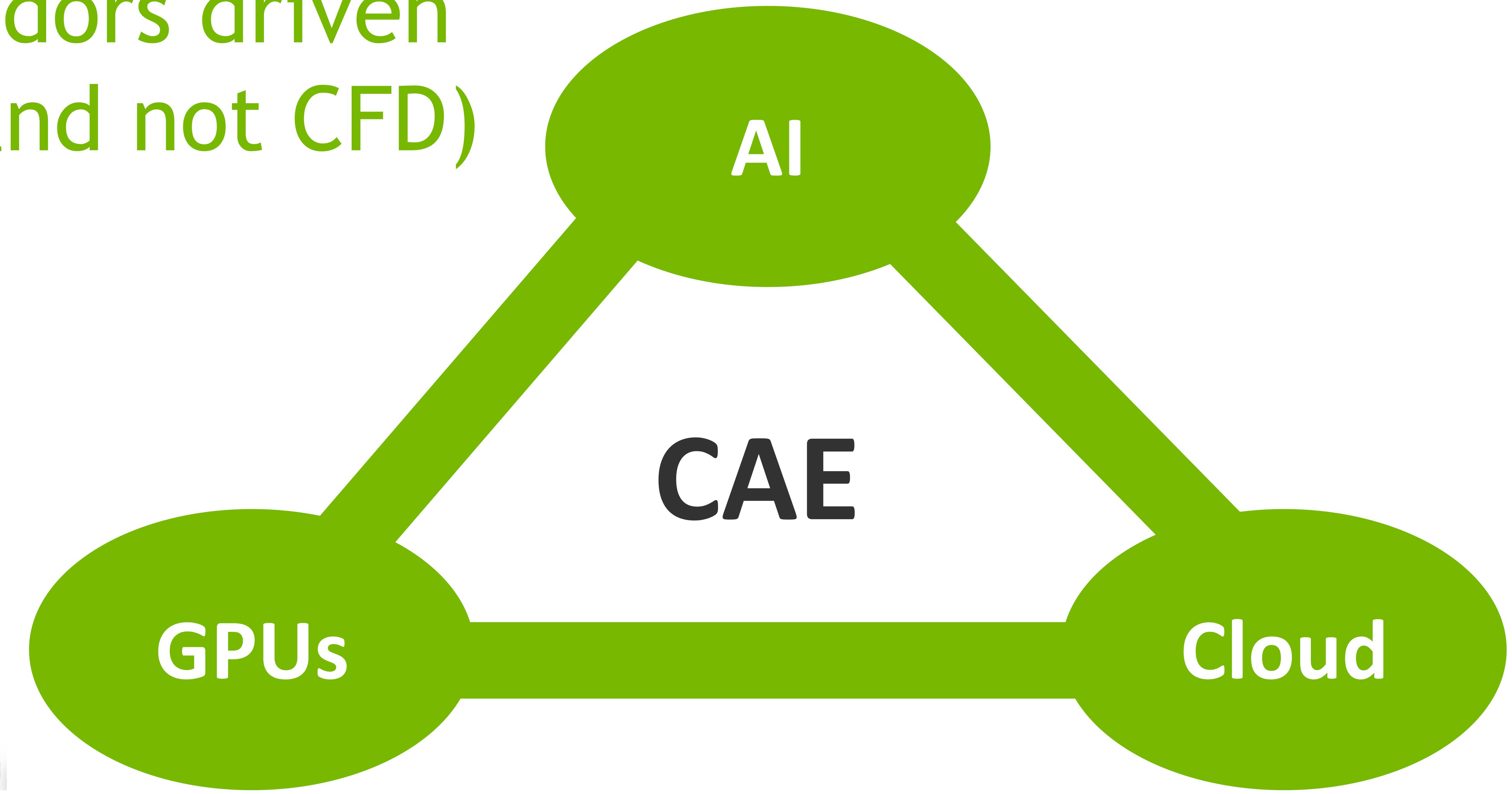
Shorten Development Cycle By 12-16 Weeks

"Now that we can run higher fidelity and more accurate simulations and still meet deadlines, we are able to reduce wind tunnel testing time for significant cost savings," said John Davis, the aerodynamics lead at Trek Bicycle. "Within the first two months of running CFD on our GPUs, we were able to cancel a planned wind tunnel test due to the increased confidence we had in simulation results."

HPC Technologies Driving Novel CAE Trends

Reality: HPC opportunity for system vendors driven by AI market (and not CFD)



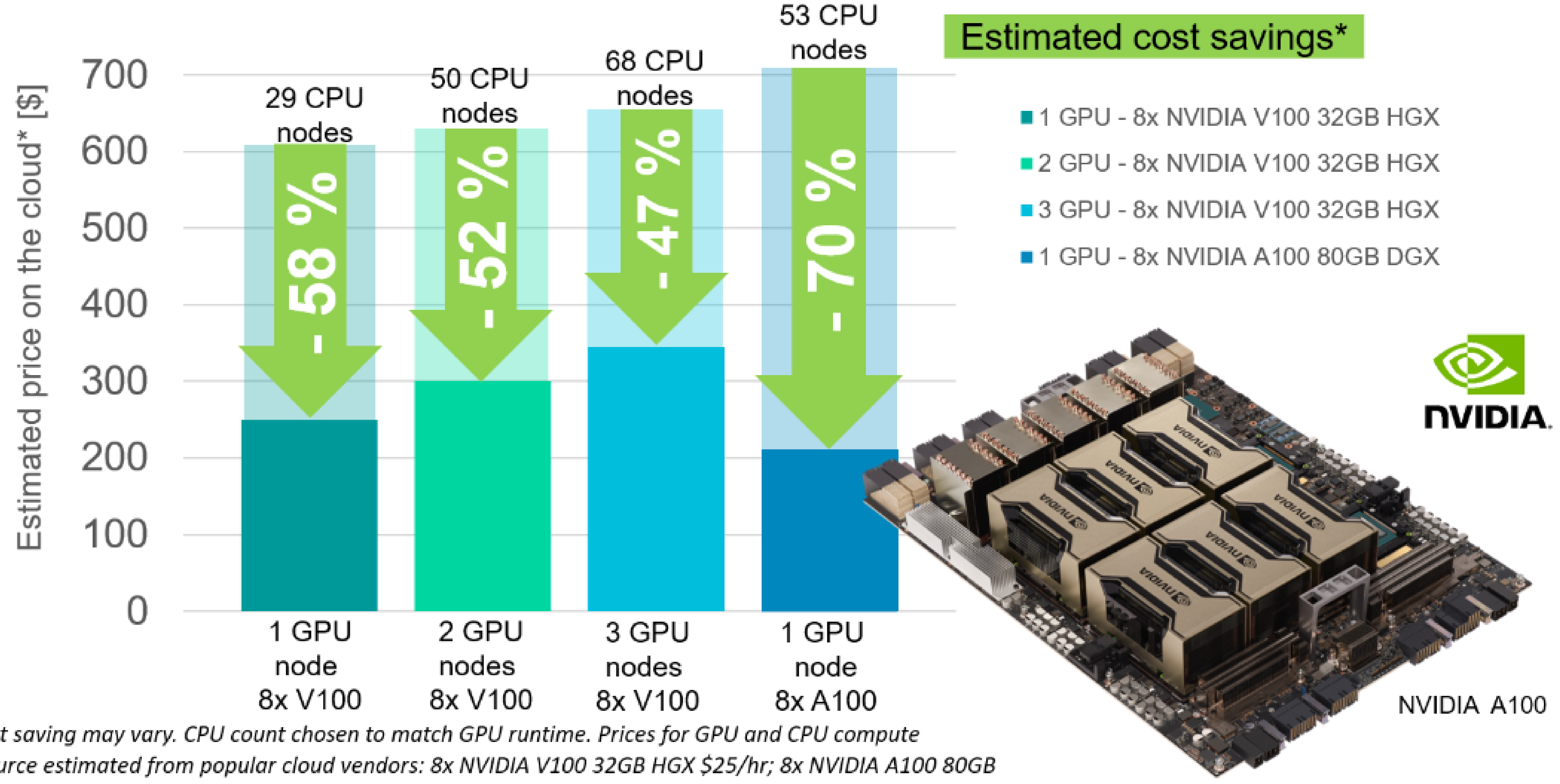


Commercial CFD Software Towards GPUs and Cloud

SIEMENS Digital Industries Software

More bang for the buck with CFD on GPUs on the cloud

If we translate those numbers into the cost of running on the cloud we realize that on NVIDIA GPUs we get a significant cost reduction. An instance of GPUs 8x NVIDIA V100 costs approximately $25, 8x NVIDIA A100 are about $33, while an instance of CPU (1 dual-socket Xeon Gold node) costs $2.10 dollars.



*Cost saving may vary. CPU count chosen to match GPU runtime. Prices for GPU and CPU compute resource estimated from popular cloud vendors: 8x NVIDIA V100 32GB HGX $25/hr; 8x NVIDIA A100 80GB DGX $33/hr; Compute-optimized CPU instance dual-socket Xeon Gold nodes (40 cores per node) $2.10/hr

https://blogs.sw.siemens.com/simcenter/gpu-acceleration-for-cfd-simulation/

COUNTRIES & REGIONS ▾ CONTACT US CAREERS STUDENTS AND ACADEMIC

Why Ansys ▾ Products & Services ▾ Learn ▾ **Ansys**

ANSYS BLOG FEBRUARY 25, 2022

Unleashing the Power of Multiple GPUs for CFD Simulations

32x Performance Gain Featured

Computational fluids dynamics (CFD) engineers are keenly interested in accelerating their simulation throughput, whether that's by automating workflows, upgrading to newer/better methods, or using high-performance computing (HPC).

https://www.ansys.com/blog/unleashing-the-full-power-of-gpus-for-ansys-fluent

CFD startups promote latest HPC trends

- **FlexCompute:** www.flexcompute.com/
- **Luminary Cloud:** www.luminarycloud.com/
- **Volcano Platforms**
- **AI Engineering:** www.ai-eng.com/

AWS Performance-Cost Study for Cloud CFD

GPUs More Favorable Performance-Cost Profile

Source: 18th OpenFOAM Workshop, 11-14 July 2033 – Dr. Neil Ashton, AWS

NVIDIA

GE Engines AI-Based Application for Aerodynamics

Learning with the flow: GE study on Summit could lead to cleaner, greener jet flights

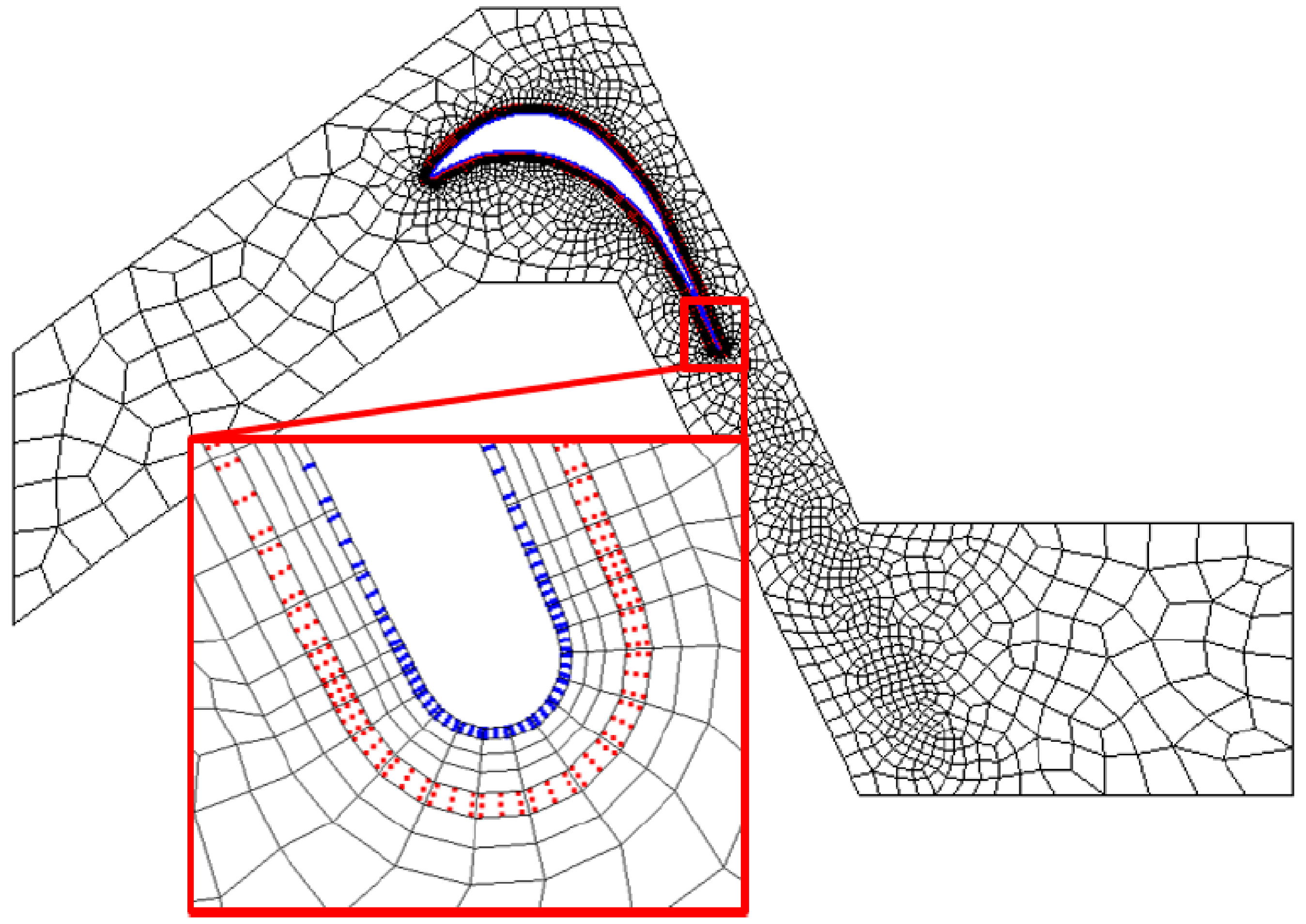
May 19, 2023

GE Research

OAK RIDGE National Laboratory

Simulations of turbulence performed on Oak Ridge National Laboratory's Summit supercomputer by GE and ORNL researchers could lead to better aircraft designs, environmentally cleaner flights and savings of as much as $400 million per year. Credit: Getty Images

R. Bhaskaran, R. Kannan, B. Barr and S. Priebe, **"Science-Guided Machine Learning for Wall-Modeled Large Eddy Simulation,"** *2021 IEEE Intl Conference on Big Data*, Orlando, FL, USA, 2021, pp. 1809-1816, doi: 10.1109/BigData52589.2021.9671436.



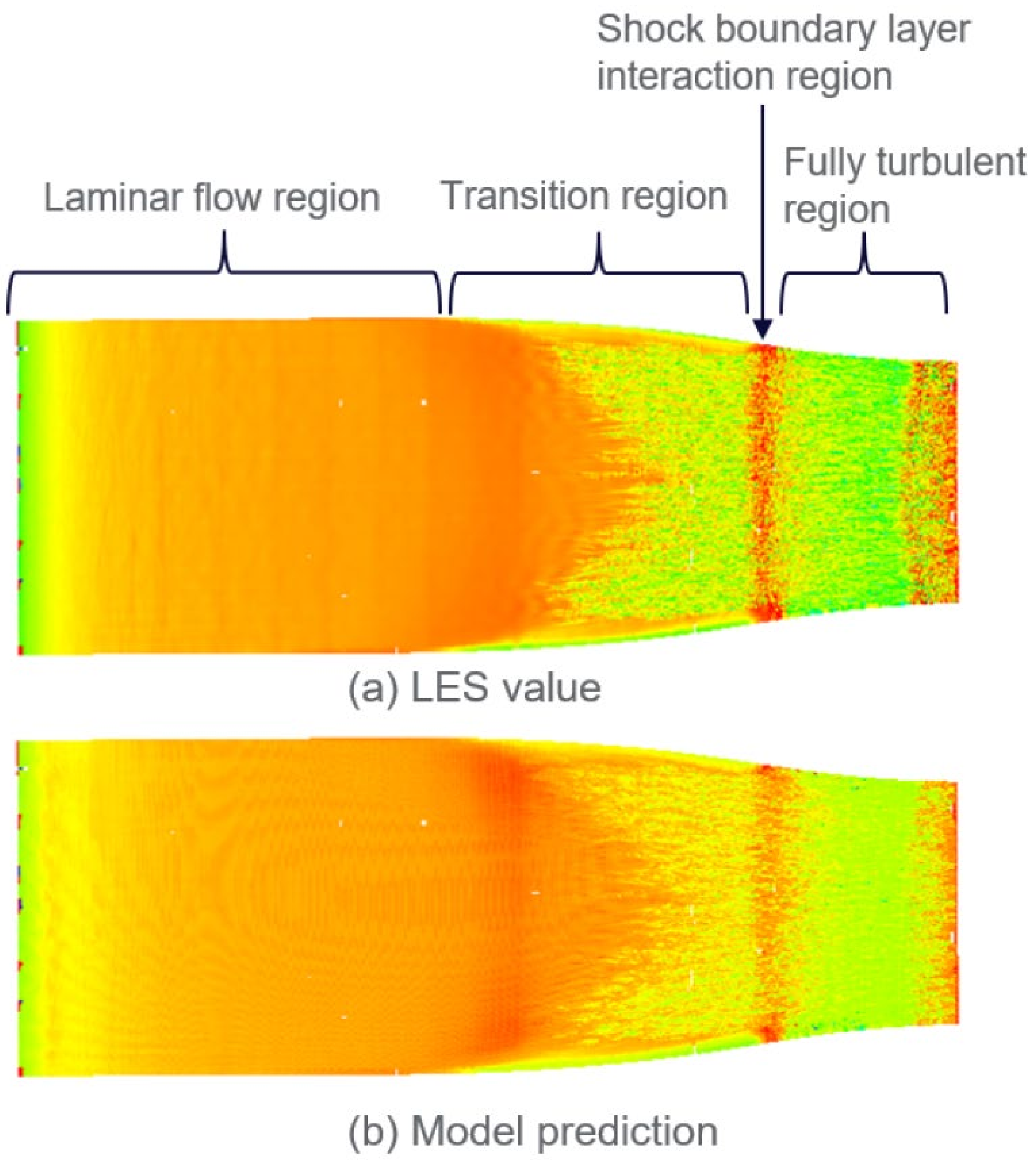
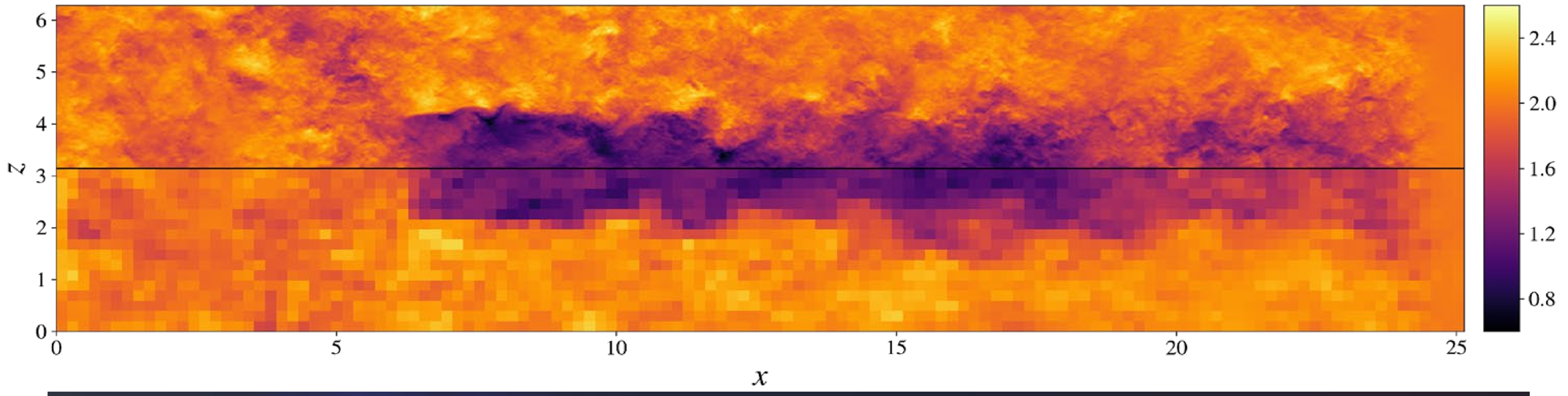


Fig. 7. Comparison of contours of streamwise component of wall shear. (a) LES value, (b) ML model prediction. Local boundary layer characteristics, including the laminar-turbulent transition and SBLI effects are well captured by the model.

Siemens Gamesa Wind Farm AI-Based Application

AI to maximize wind energy lay-out and production using wake optimization

Super resolution of low-fidelity results computed by conventional LES solver



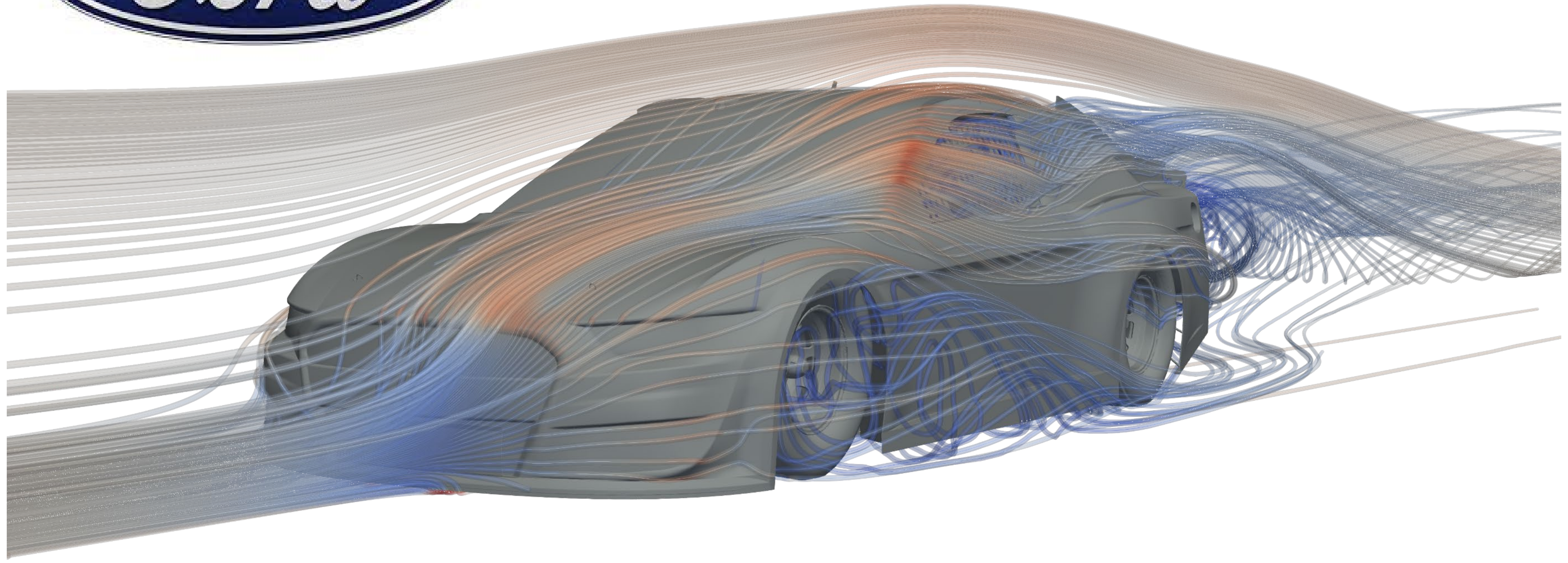
NVIDIA. SIEMENS Gamesa RENEWABLE ENERGY

Modulus and Omniverse

~4,000X speedup for high-fidelity inference simulation

https://www.youtube.com/watch?v=mQuvYQmdbtw

AI-Based Applications in Automotive Aerodynamics



Augmenting CFD, Ford developed a virtual wind tunnel neural network on NVIDIA DGX to simulate race care aerodynamic performance for various configurations.

Simulations were 99% accurate and completed in a few hours vs. 3-4 days, enabling the Ford team with vehicle adjustments before every race.

Nissan applied convolutional neural networks (CNNs) trained on NVIDIA DGX systems to obtain flow fields in 20 seconds vs. 1 day for conventional CFD simulation.

Designers can interactively revise vehicle shapes and know aerodynamic for rapid concept evalutions.

"Now Using NVIDIA DGX A100, we were able to train models that can quickly estimate vehicle flow fields in seconds instead of at least a day. This has enabled our designers to interactively revise a vehicle's shape and speed up time to market," said Tetsuro Ueda, Expert Leader, AI and Data Science, Nissan Motor

Thank You and Q&A

sposey@nvidia.com