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
## Feature Progression of NVIDIA GPU Architectures

<table>
<thead>
<tr>
<th></th>
<th>H100 (2023)</th>
<th>A100 (2020)</th>
<th>V100 (2017)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peak FP64 TF/s</td>
<td>26 / 34 <strong>High order dense ops</strong> 3x</td>
<td>9.7</td>
<td>1.3x</td>
</tr>
<tr>
<td>Peak FP64 TC TF/s</td>
<td>51 / 67</td>
<td>19.5</td>
<td></td>
</tr>
<tr>
<td>Peak FP32 TFlop/s</td>
<td>51 / 67 3x</td>
<td>19.5</td>
<td></td>
</tr>
<tr>
<td>Peak TF32 TC TF/s</td>
<td>756 / 989</td>
<td>156</td>
<td></td>
</tr>
<tr>
<td>Peak FP16 TFlop/s</td>
<td>1513 / 1979 6.4x</td>
<td>312 2.6x</td>
<td>120</td>
</tr>
<tr>
<td>Memory BW (GB/s)</td>
<td><strong>2000 / 3350 1.6x</strong></td>
<td>1555 / 2038 2.3x</td>
<td>900</td>
</tr>
<tr>
<td>Memory Capacity (GB)</td>
<td>96</td>
<td>40 / 80 2.5x</td>
<td>16 / 32</td>
</tr>
<tr>
<td>Interconnect</td>
<td><strong>NVLink: Up to 900 GB/s</strong> 2.0x</td>
<td><strong>NVLink: Up to 600 GB/s</strong> 2.0x</td>
<td><strong>NVLink: Up to 300 GB/s</strong> PCIe: 32 GB/s</td>
</tr>
<tr>
<td>Max Power (W)</td>
<td><strong>300 – 700</strong></td>
<td>400</td>
<td>250 - 300</td>
</tr>
</tbody>
</table>
NVIDIA Next-gen GPU H100 and Arm CPU “Grace”

Breakthrough Designs for Large-Scale HPC and AI Applications

Grace Arm + H100 (Hopper)

Grace Arm-only Node

Available Q3 2023

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

- Physics-ML Models:
  - Turbulence
  - Wall
  - Collision
  - Coarse Graining

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

<table>
<thead>
<tr>
<th>Org</th>
<th>Software</th>
<th>Method</th>
<th>GPU Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>NASA</td>
<td>FUN3D</td>
<td>CUDA</td>
<td>Full application, RG chemistry</td>
</tr>
<tr>
<td>NASA</td>
<td>OVERFLOW</td>
<td>OpenACC</td>
<td>Central diff schemes, Euler flux, smoothing</td>
</tr>
<tr>
<td>DoD NRL</td>
<td>JENRE</td>
<td>CUDA</td>
<td>Full application, high-order FE LES</td>
</tr>
<tr>
<td>DOE ANL</td>
<td>NekRS</td>
<td>OCCA (CUDA)</td>
<td>Full application, high-order spectral element</td>
</tr>
<tr>
<td>OpenFOAM®</td>
<td>OpenFOAM</td>
<td>CUDA Lib</td>
<td>Linear solver only using AmgX lib</td>
</tr>
<tr>
<td>Boeing</td>
<td>BCFD</td>
<td>OpenACC</td>
<td>Full application, 2nd order FV RANS</td>
</tr>
<tr>
<td>GE</td>
<td>GENESIS</td>
<td>CUDA</td>
<td>Full application, high-order LES</td>
</tr>
<tr>
<td>ANSYS</td>
<td>Fluent</td>
<td>CUDA</td>
<td>Full application, core features</td>
</tr>
<tr>
<td>Siemens</td>
<td>STAR-CCM+</td>
<td>CUDA</td>
<td>Full application, core features</td>
</tr>
<tr>
<td>Altair</td>
<td>ultraFluidX</td>
<td>CUDA</td>
<td>Full application LBM, core features</td>
</tr>
<tr>
<td>Cadence</td>
<td>CharLES</td>
<td>CUDA</td>
<td>Full application, core features</td>
</tr>
</tbody>
</table>
Computational Investigation of the Effects of Chemistry on Mars Retropulsion Environments

Jan-Renee Carlson
Bill Jones
Ashley Korzun
Gabriel Nastac
Eric Nielsen
Aaron Walden
Li Wang
NASA Langley Research Center
Pat Moran
Tim Sandstrom
NASA Ames Research Center
Paul Kolano
Inu Teq, LLC

Alexander Kuhn
Justin Luijens
Jörg Mensmann
Marc Nienhaus
Dragos Tatules
Rajko Yasei-Schoeffel
NVIDIA Corp.

Christopher Stone
National Institute of Aerospace
Mohammad Zubair
Old Dominion University

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

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

OpenFOAM GPU Strong Scaling on ORNL Summit

Strong scaling on Summit

Lid driven cavity problem:
- 8M cells (on node cpu wall time per time-step = 12.4625s)
- 64M cells (on node cpu wall time per time-step = 156.7556s)
- 216M cells (on node cpu wall time per time-step = 749.4231s)

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

- **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:
  - “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

**Graph: 434s vs 225W vs 703s vs 399W vs 417W**

- Grace (72c) vs Sapphire Rapids (52c) vs Genoa (96c)

**NOTE:** All results single socket processor
Example: Hybrid CPU/GPU moving solver Fidelity CharLES
Grace-Hopper vs X86+A100

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
Trek Bicycle Apply CFD Workflow on NVIDIA GPUs

Up to 6x performance gain across product development

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 $2.20 dollars.

Estimated cost savings:

- 58% for 29 GPU nodes
- 52% for 50 GPU nodes
- 47% for 68 GPU nodes
- 70% for 53 GPU nodes

Estimated price on the cloud* [

<table>
<thead>
<tr>
<th>Nodes</th>
<th>Price Reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 GPU node 8x V100</td>
<td>- 58%</td>
</tr>
<tr>
<td>2 GPU nodes 8x V100</td>
<td>- 52%</td>
</tr>
<tr>
<td>3 GPU nodes 8x V100</td>
<td>- 47%</td>
</tr>
<tr>
<td>4 GPU nodes 8x A100</td>
<td>- 70%</td>
</tr>
</tbody>
</table>

*Cost saving may vary. CPU cost used to match GPU runtime. Prices for GPU and CPU compute resources estimated from popular cloud vendors: 8x NVIDIA V100 32GB HGX $23/hr; 8x NVIDIA A100 80GB DGX $33/hr. Compute-optimized CPU instance dual-socket Xeon Gold nodes (60 cores per node) $2.10/hr


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/

https://www.ansys.com/blog/unleashing-the-full-power-of-gpus-for-ansys-fluent
AWS Performance-Cost Study for Cloud CFD

GPUs More Favorable Performance-Cost Profile

- Similar conclusions on previous work with Zenotech. 1 GPU node (i.e. x8 A100) = ~2000 CPU cores or 1 GPU = ~250-300 CPU cores.
- GPU’s (p4d.24xlarge, p3.24xlarge and g4dn.16xlarge Amazon EC2 instances i.e Nvidia A100/v100 and T4’s) deliver faster and cheaper results.

Source: 18th OpenFOAM Workshop, 11-14 July 2033 – Dr. Neil Ashton, AWS
Learning with the flow: GE study on Summit could lead to cleaner, greener jet flights

May 19, 2023

Simulation 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


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

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

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

“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