Overview of GPU Suitability and Progress of CFD Applications

NASA Ames Applied Modeling & Simulation (AMS) Seminar – 21 Apr 2015

Stan Posey; sposey@nvidia.com; NVIDIA, Santa Clara, CA, USA

Agenda: GPU Suitability and Progress of CFD

- NVIDIA HPC Introduction
- CFD Suitability for GPUs
- CFD Progress and Directions

NVIDIA - Core Technologies and Products

Company Revenue of ~$5B USD; ~8,800 Employees; HPC Growing > 35% CAGR

GPU

GeForce®
Quadro®, Tesla®

Mobile

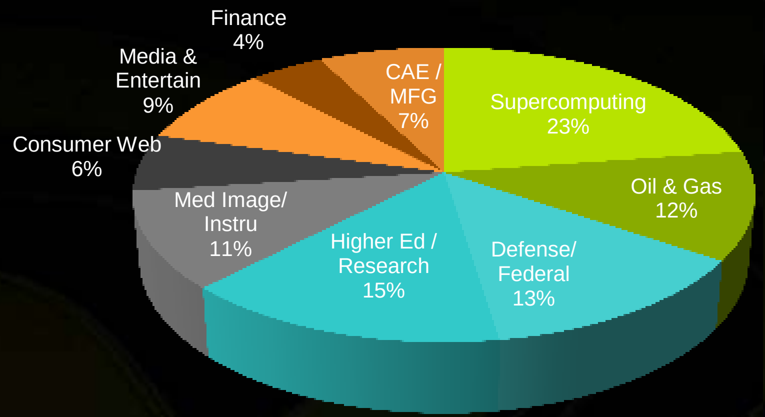
Tegra®

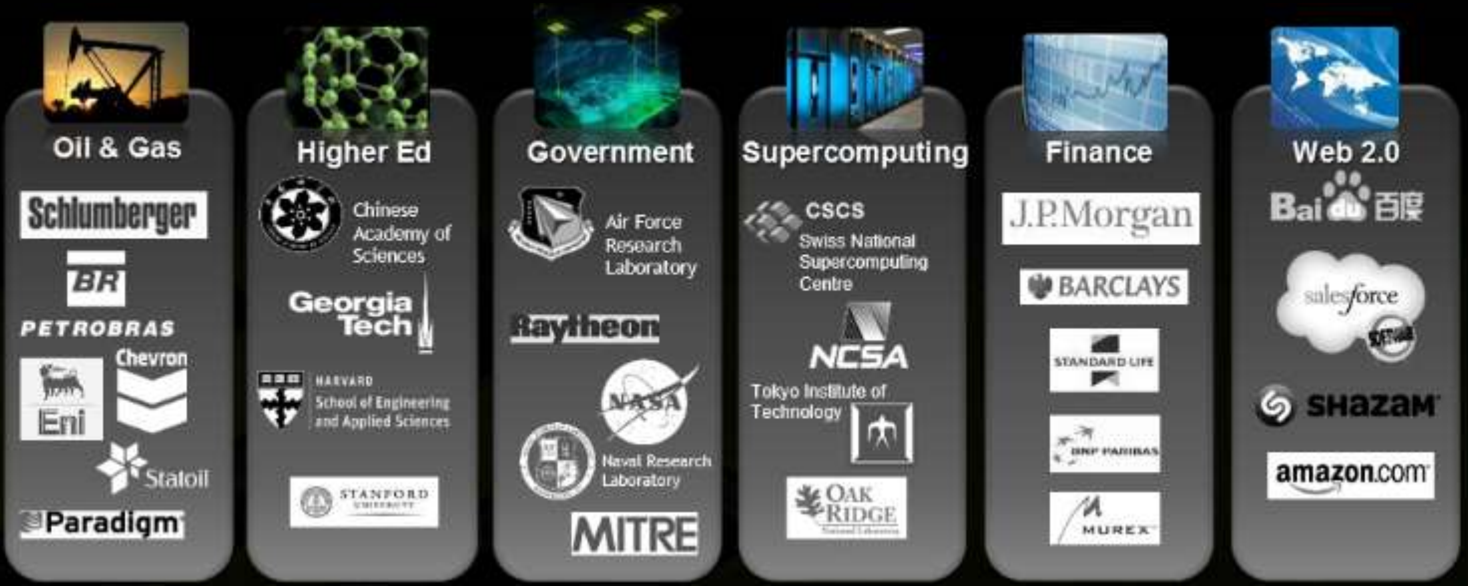
Cloud

GRID

GPUs Mainstream Across Diverse HPC Domains

FY14 Segments





World's Top 3 Servers are GPU-Accelerated

Tesla GPU Progression During Recent Years

	2012 (Fermi) **M2075**	2014 (Kepler) **K20X**	2014 (Kepler) **K40**	2014 (Kepler) **K80**	**Kepler / Fermi**
Peak SP Peak SGEMM	1.03 TF	3.93 TF 2.95 TF	4.29 TF 3.22 TF	8.74 TF	**4x**
Peak DP Peak DGEMM	.515 TF	1.31 TF 1.22 TF	1.43 TF 1.33 TF	2.90 TF	**3x**
Memory size	6 GB	6 GB	**12 GB**	**24 GB (12 each)**	**2x**
Mem BW (ECC off)	150 GB/s	250 GB/s	**288 GB/s**	**480 GB/s (240 each)**	**2x**
Memory Clock		2.6 GHz	3.0 GHz	3.0 GHz	
PCIe Gen	Gen 2	Gen 2	Gen 3	Gen 3	**2x**
# of Cores	448	2688	2880	4992 (2496 each)	**5x**
Board Power	**235W**	**235W**	**235W**	**300W**	**0% – 28%**

Note: Tesla K80 specifications are shown as aggregate of two GPUs on a single board

GPU Motivation (II): Energy Efficient HPC

Top500 Rank	TFLOPS/s	Site
1	33,862.7	National Super Computer Centre Guangzhou
2	17,590.0	Oak Ridge National Lab **#1 USA**
3	17,173.2	DOE, United States
4	10,510.0	RIKEN Advanced Institute for Computational Science
5	8,586.6	Argonne National Lab
6	6,271.0	Swiss National Supercomputing Centre (CSCS) **#1 Europe**
7	5,168.1	University of Texas
8	5,008.9	Forschungszentrum Juelich
9	4,293.3	DOE, United States
10	3,143.5	Government

Green500 Rank	MFLOPS/W	Site
1	4,389.82	GSIC Center, Tokyo Tech **KFC**
2	3,631.70	Cambridge University
3	3,517.84	University of Tsukuba
4	3,459.46	SURFsara
5	3,185.91	Swiss National Supercomputing (CSCS)
6	3,131.06	ROMEO HPC Center
7	3,019.72	CSIRO
8	2,951.95	GSIC Center, Tokyo Tech **2.5**
9	2,813.14	Eni
10	2,629.10	(Financial Institution)
16	2,495.12	Mississippi State (top non-NVIDIA) **Intel Phi**
59	1,226.60	ICHEC (top X86 cluster)

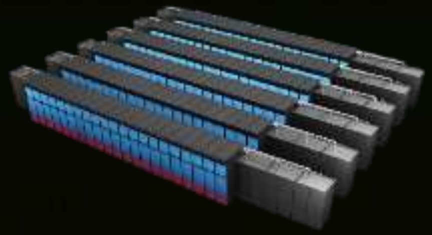
GPU Motivation (III): Advanced CFD Trends

- Higher fidelity models within manageable compute and energy costs

RANS → URANS → LES → ? DNS

- Increase in non-deterministic ensembles to manage/quantify uncertainty

Number of Jobs > 10x



- HOMs for improved resolution of transitional and vortex-heavy flows

Accelerator technology identified as a cost-effective and practical approach to future computational challenges

NVIDIA Strategy for GPU-Accelerated HPC

Strategic Alliances

- Business and technical alliances with COTS vendors
- Investment in long-term collaboration for solver-level libraries
- Development of collaborations with academic research community:
  - Examples in CFD: Imperial College—Vincent, Oxford—Giles, Wyoming—Mavriplis, GMU—Lohner, UFRJ—Coutinho, TiTech—Aoki, GWU—Barba, SU—Jameson, others

Software Development

- Libraries cuSPARSE, cuBLAS; OpenACC with PGI (acquisition) and Cray
- NVIDIA linear solver toolkit with emphasis on AMG for industry CFD

Applications Support

- Application engineering support for COTS vendors and customers
  - **Implicit Schemes:** Integration of libraries and solver toolkit
  - **Explicit Schemes:** Stencil libraries, OpenACC for directives-based

Agenda: GPU Suitability and Progress of CFD

- NVIDIA HPC Introduction
- **CFD Suitability for GPUs**
- CFD Progress and Directions

Programming Strategies for GPU Acceleration

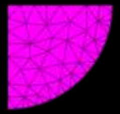
Applications

GPU Libraries	OpenACC Directives	Programming Languages
Provides Fast "Drop-In" Acceleration	GPU-acceleration in Standard Language (Fortran, C, C++)	Maximum Flexibility with GPU Architecture and Software Features

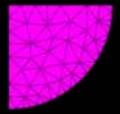
Increasing Development Effort →

CFD Characteristics and GPU Suitability

Structured Grid FV



Unstructured FV



Unstructured FE

Explicit

Usually Compressible

Numerical operations on I,J,K stencil, no "solver"
[Flat profiles: Typical GPU strategy is directives (OpenACC)]

Finite Volume

Finite Element:

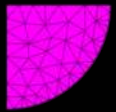
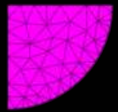
Implicit

Usually Incompressible

Sparse matrix linear algebra – iterative solvers
[Hot spot ~50%, few LoC: Typical GPU strategy is CUDA and libs]

Select GPU Implementations for CFD Practice

NVIDIA

	 Structured Grid FV	 Unstructured FV	 Unstructured FE
Explicit Usually Compressible	SJTU RANS (COMAC) HiPSTAR (University of Southampton) Turbostream (Turbostream) TACOMA (GE)	HYDRA (Rolls-Royce) Flare (BAE SYSTEMS)	SD++ (Stanford University) PyFR (Imperial College London) Hyperflux (cfms, Hyperflux) JENRE, Propel
	Finite Volume		Finite Element:
Implicit Usually Incompressible		ANSYS Fluent (ANSYS) Culises for OpenFOAM (FluiDyna, OpenFOAM)	AcuSolve (Altair) Moldflow (Autodesk)

What is Meant by "CFD Practice"

- These are not demonstrators, rather meaningful developments towards production use CFD
- Proven performance on large-scale engineering simulations
- Long-term maintenance and software engineering considerations
- Co-design efforts between CFD scientists and computer scientists
- In most cases, contributions from NVIDIA devtech engineering

Select GPU Implementations for CFD Practice

NVIDIA

	Structured Grid FV	Unstructured FV	Unstructured FE
Explicit Usually Compressible	SJTU RANS HiPSTAR Turbostream TACOMA	HYDRA Flare	SD++ PyFR Hyperflux JENRE, Propel
Implicit Usually Incompressible		ANSYS Fluent Culises for OpenFOAM	AcuSolve Moldflow

Finite Volume

Finite Element:

Structured grid explicit generally best GPU fit

Select GPU Implementations for CFD Practice

	Structured Grid FV	Unstructured FV	Unstructured FE
Explicit Usually Compressible	SJTU RANS HiPSTAR Turbostream TACOMA	HYDRA Flare	SD++ PyFR Hyperflux JENRE, Propel
Implicit Usually Incompressible		ANSYS Fluent Culises for OpenFOAM	AcuSolve Moldflow

Finite Volume

Finite Element:

Unstructured grid usually with renumbering/coloring

Select GPU Implementations for CFD Practice

NVIDIA

	Structured Grid FV	Unstructured FV	Unstructured FE
Explicit Usually Compressible	COMAC — SJTU RANS University of Southampton — HiPSTAR Turbostream — Turbostream GE — TACOMA	Rolls-Royce — HYDRA BAE SYSTEMS — Flare	Stanford University — SD++ Imperial College London — PyFR cfms, Hyperflux — Hyperflux JENRE, Propel
Implicit Usually Incompressible		ANSYS — ANSYS Fluent FluiDyna, OpenFOAM — Culises for OpenFOAM	Altair — AcuSolve Autodesk — Moldflow

Finite Volume

Finite Element:

COTS CFD mostly apply solver library (AmgX)

Select GPU Implementations (Summary)

Organization	Location	Software	GPU Approach
COMAC/SJTU	China	SJTU RANS	Fortran and CUDA
U Southhampton	UK	HiPSTAR	Fortran and OpenACC
Turbostream	UK	Turbostream	Fortran, python templates s-to-s to CUDA
GE GRC	US	TACOMA	Fortran and OpenACC
Rolls Royce	UK	HYDRA	Fortran, python DSL s-to-s to CUDA-F
BAE Systems	UK	Flare	C++ and s-to-s templates to CUDA
Stanford U	US	SD++	C++ and CUDA
PyFR	UK	PyFR	Python s-to-s to CUDA (C for CPU)
CFMS	UK	Hyperflux	Python s-to-s to CUDA (C for CPU)
JENRE, Propel	US	USA	C++ and Thrust templates for CUDA
ANSYS Fluent	US	Implicit FEA	C++ and AmgX library, OpenACC
FluiDyna	DE	Culises/OpenFOAM	C++ (OpenFOAM), AmgX library, CUDA
Altair	US	AcuSolve	Fortran and CUDA
Autodesk	US	Moldflow	Fortran and AmgX library

Select GPU Developments at Various Stages

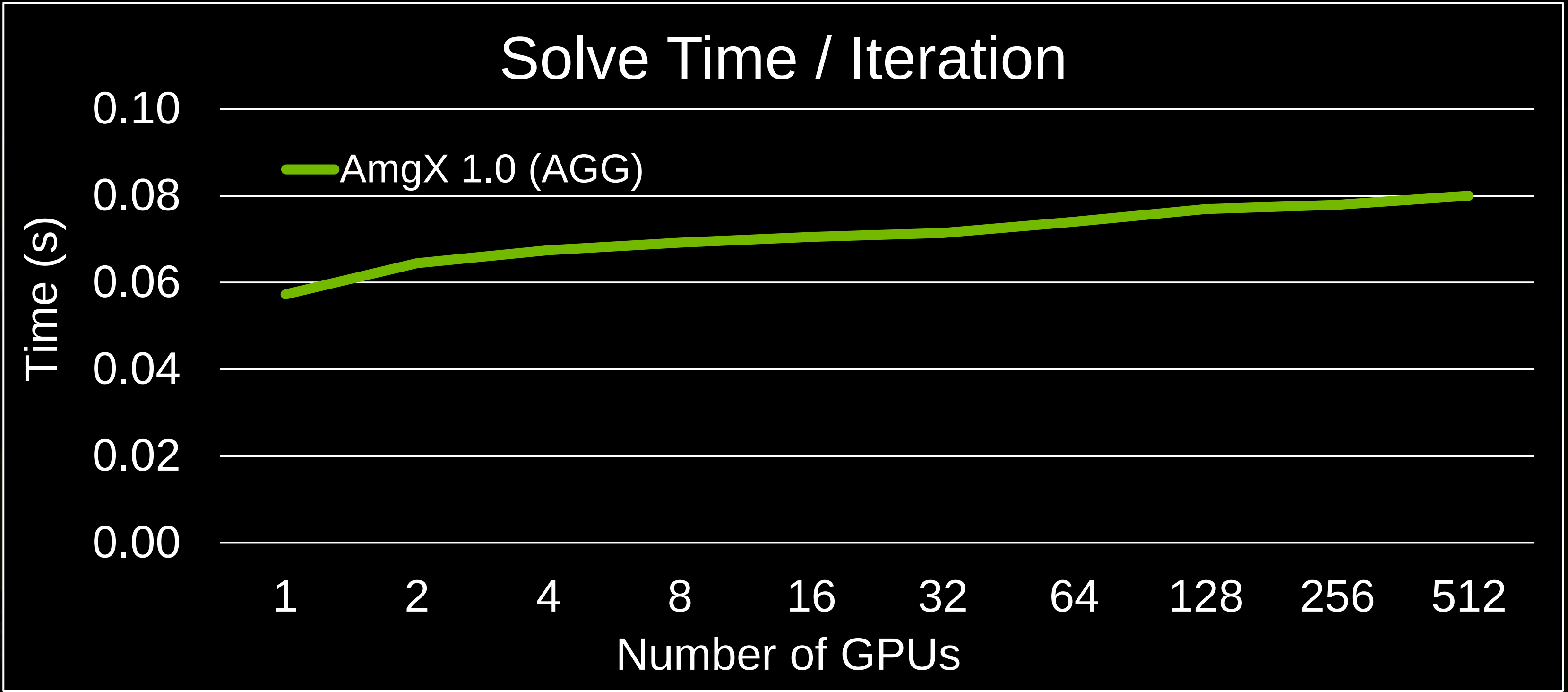
Organization	Location	Software	GPU Approach
U Wyoming /Mavriplis	US	*Not specific*	CU++ object oriented templates
GMU / Lohner	US	FEFLO	Python Fortran-to-CUDA translator
SpaceX	US	*Not specific*	C++ and CUDA
CPFD	US	BARRACUDA	Fortran and CUDA
GWU / Barba	US	*Not specific*	C++, python, pyCUDA
UTC Research	US	*Combustion*	Fortran and CUDA
Convergent Science / LLNL	US	CONVERGE	C++ and CUDA, cuSOLVE (NVIDIA)
Craft Tech	US	CRAFT, CRUNCH	Fortran and CUDA, OpenACC
Bombardier	CA	*Not specific*	C++ and CUDA
DLR	DE	TAU	Fortran and CUDA
ONERA	FR	elsA	Fortran and CUDA
Vratis	PL	Speed-IT (OFOAM)	C++ and CUDA
NUMECA	BE	Fine/Turbo	Fortran and OpenACC
Prometech	JP	Particleworks	C++ and CUDA
TiTech / Aoki	JP	*Not specific*	C++ and CUDA
JAXA	JP	UPACS	Fortran and OpenACC
KISTI / Park	KR	KFLOW	Fortran and CUDA, OpenACC
VSSC	IN	PARAS3D	Fortran and CUDA

NVIDIA AmgX for Iterative Implicit Methods

- Scalable linear solver library for Ax = b iterative methods
- No CUDA experience required, C API: links with Fortran, C, C++
- Reads common matrix formats (CSR, COO, MM)
- Interoperates easily with MPI, OpenMP, and hybrid parallel
- Single and double precision; Supported on Linux, Win64
- Multigrid; Krylov: GMRES, PCG, BiCGStab; Preconditioned variants
- Classic Iterative: Block-Jacobi, Gauss-Seidel, ILU’s; Multi-coloring
- Flexibility: All methods as solvers, preconditioners, or smoothers
- Download AmgX library: http://developer.nvidia.com/amgx

NVIDIA AmgX Weak Scaling on Titan 512 GPUs

Use of 512 nodes on ORNL TITAN System



- **Poisson matrix with ~8.2B rows solved in under 13 sec (200e3 Poisson matrix per GPU)**
- **ORNL TITAN: NVIDIA K20X one per node; CPU 16 core AMD Opteron 6274 @2.2GHz**

Agenda: GPU Suitability and Progress of CFD

- NVIDIA HPC Introduction
- CFD Suitability for GPUs
- **CFD Progress and Directions**

Progress Summary for GPU-Parallel CFD

- GPU progress in CFD research continues to expand
  - Growth from arithmetic intensity in high-order methods
  - Breakthroughs with Hyper-Q feature (Kepler), GPUDirect, etc.
- Strong GPU investments by commercial (COTS) vendors
  - Breakthroughs with AmgX linear solvers and preconditioners
  - Preservation of costly MPI investment: GPU 2nd-level parallelism
- Success in end-user developed CFD with OpenACC
  - Most benefits currently with legacy Fortran, C++ emerging
- GPUs behind fast growth in particle-based commercial CFD
  - New commercial developments in LBM, SPH, DEM, etc.

OpenACC Acceleration of TACOMA at GE GRC

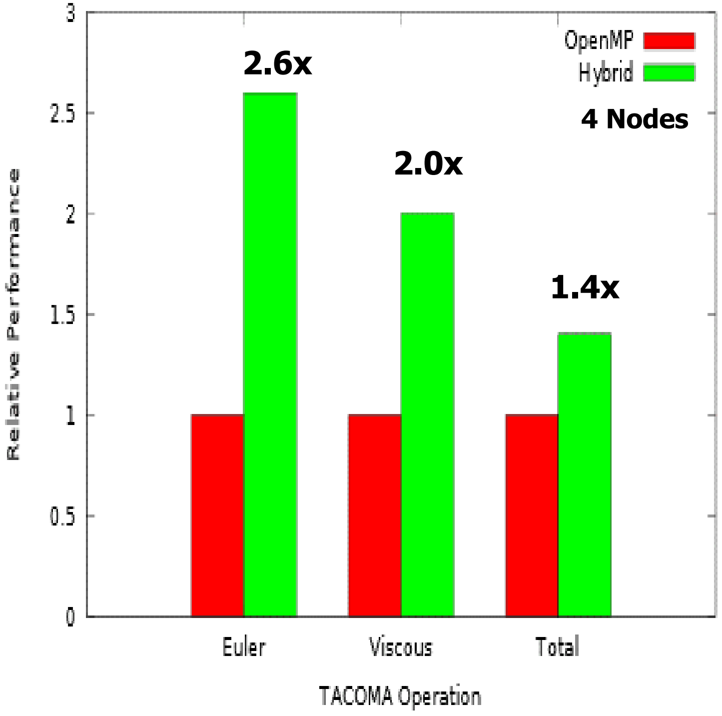
Tri-Hybrid Computational Fluid Dynamics on DOE's Cray XK7, Titan.

Aaron Vose[†], Brian Mitchell[*], and John Levesque[‡].

Cray User Group, May 2014.

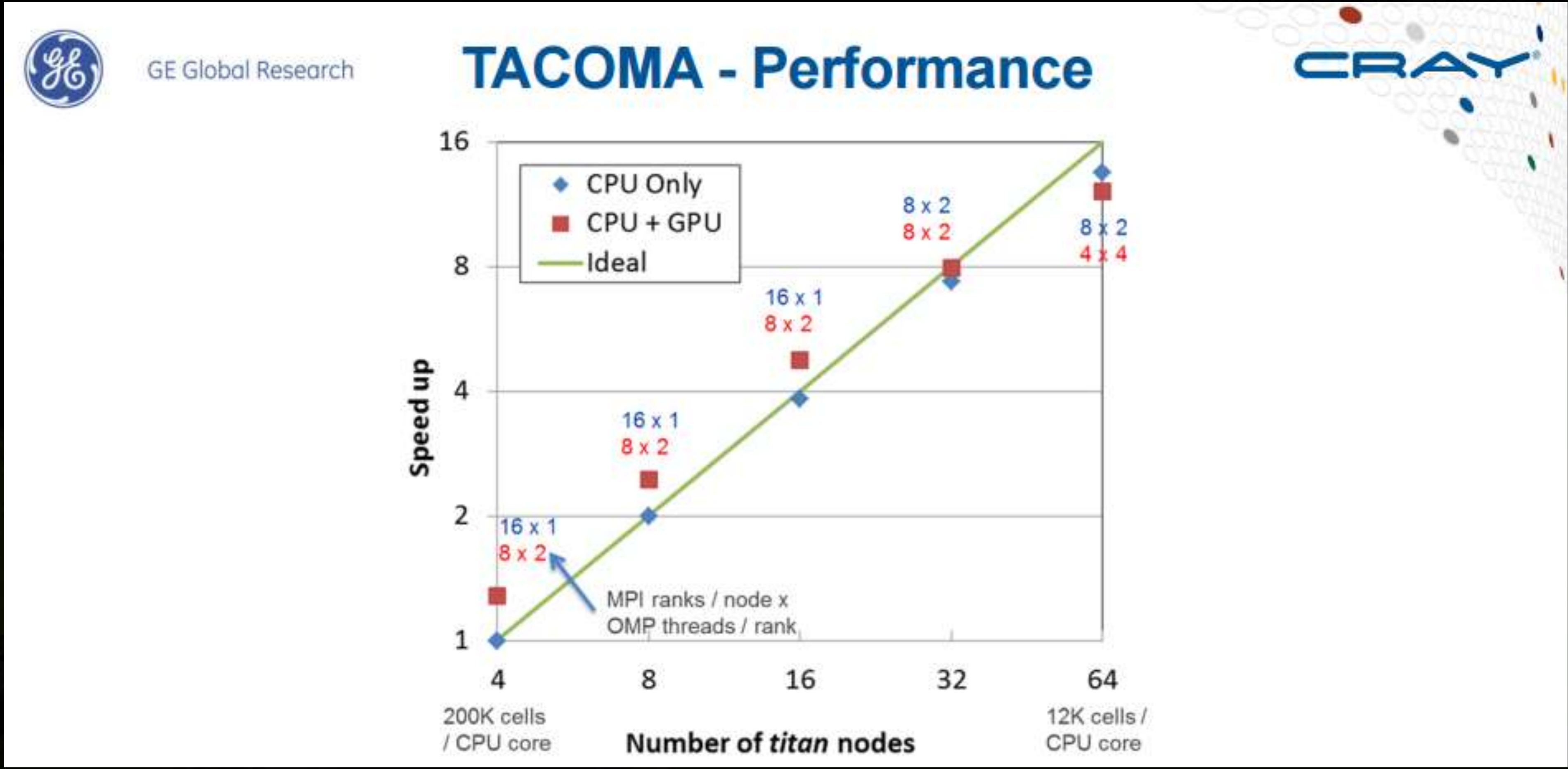
GE Global Research: *mitchellb@ge.com — Cray Inc.: †avose@cray.com, ‡levesque@cray.com.

Abstract — *A tri-hybrid port of General Electric's in-house, 3D, Computational Fluid Dynamics (CFD) code TACOMA is created utilizing MPI, OpenMP, and OpenACC technologies. This new port targets improved performance on NVidia Kepler accelerator GPUs, such as those installed in the world's second largest supercomputer, Titan, the Department of Energy's 27 petaFLOP Cray XK7 located at Oak Ridge National Laboratory. We demonstrate a 1.4x speed improvement on Titan when the GPU accelerators are enabled. We highlight key optimizations and techniques used to achieve these results. These optimizations enable larger and more accurate simulations than were previously possible with TACOMA, which not only improves GE's ability to create higher performing turbomachinery blade rows, but also provides "lessons learned" which can be applied to the process of optimizing other codes to take advantage of tri-hybrid technology with MPI, OpenMP, and OpenACC.*



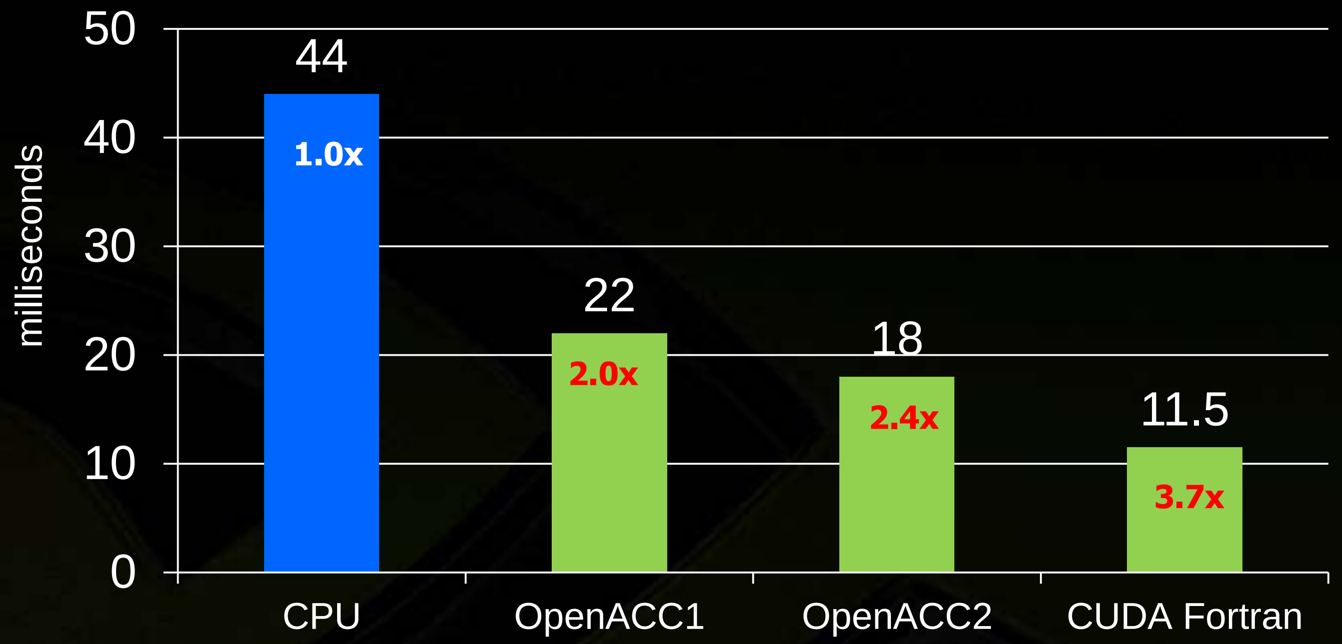
Source: https://cug.org/proceedings/cug2014_proceedings/includes/files/pap113.pdf
http://on-demand-gtc.gputechconf.com/gtc-quicklink/e7FnYI

OpenACC Acceleration of TACOMA at GE GRC



Source: https://cug.org/proceedings/cug2014_proceedings/includes/files/pap113.pdf
http://on-demand-gtc.gputechconf.com/gtc-quicklink/e7FnYI

NASA FUN3D and 5-Point Stencil Kernel on GPUs



FUN3D Fully Unstructured Navier-Stokes

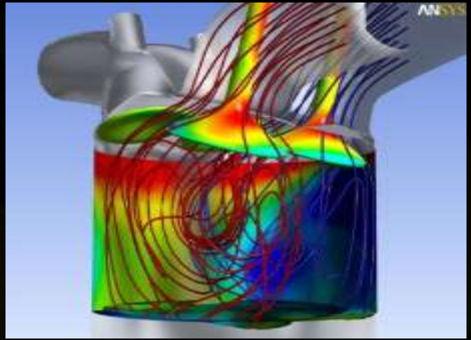
www.fun3d.larc.nasa.gov

- **CPU**: E5-2690 @ 3Ghz, 10 cores
- **GPU**: K40c, boost clocks, ECC off
- **Case**: DPW-Wing, 1M cells
- 1 call of point_solve5 over all colors
- No data transfers in GPU results
- **1 CPU core**: 300ms
- **10 CPU cores**: 44ms (6.8x on 10)

- OpenACC1 = Unchanged code: 2.0x
- OpenACC2 = Modified code : 2.4x (same modified code runs 50% slower on CPUs)
- CUDA Fortran = Highly optimized CUDA version: 3.7x
- Compiler options (e.g. how memory is accessed) have huge influence on OpenACC results
- Possible compromise: CUDA for few hotspots, OpenACC for the rest
- Demonstrated good interoperability: CUDA can use buffers managed with OpenACC data clauses

ANSYS Fluent

ANSYS Fluent Profile for Coupled PBNS Solver



Non-linear iterations

Assemble Linear System of Equations

Solve Linear System of Equations: $Ax = b$

Converged ?

No

Yes

Stop

Accelerate this first

Runtime:

~ 35%

~ 65%

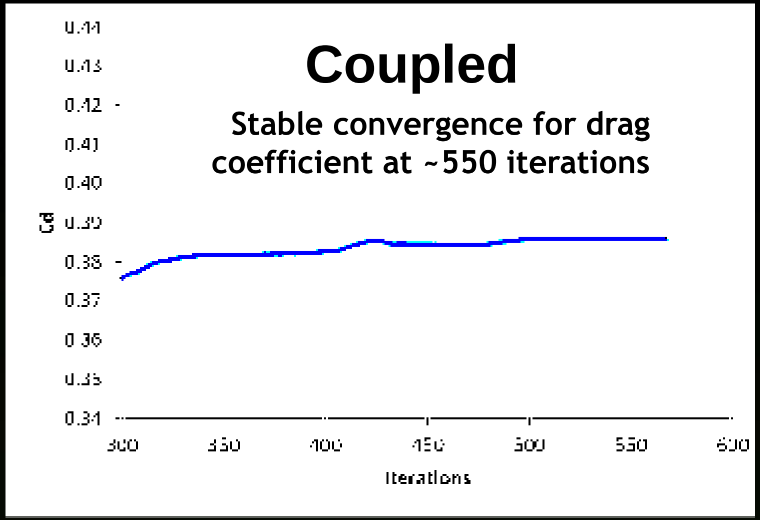
ANSYS Fluent Convergence Behavior

Coupled vs segregated solver

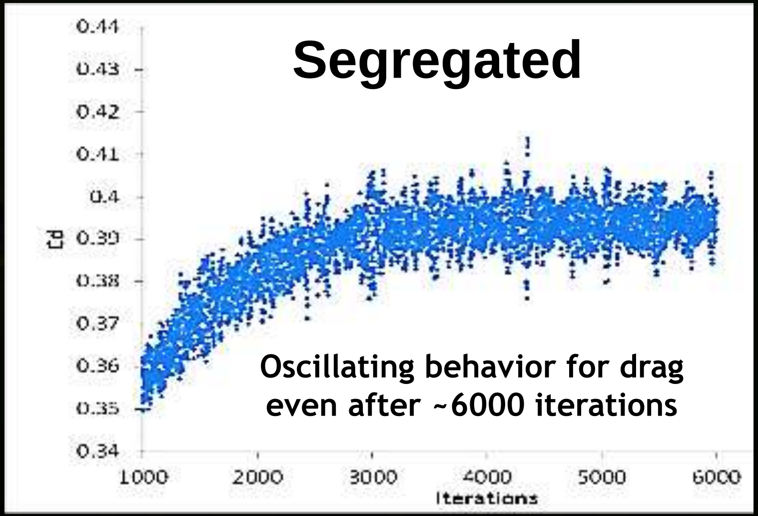
Coupling Momentum and Continuity Increases CFD Robustness

FLUENT technology introduces a pressure-based coupled solver to reduce computation time for low-speed compressible and incompressible flow applications.

By Franklyn J. Kelecy, Applications Specialist, ANSYS, Inc.

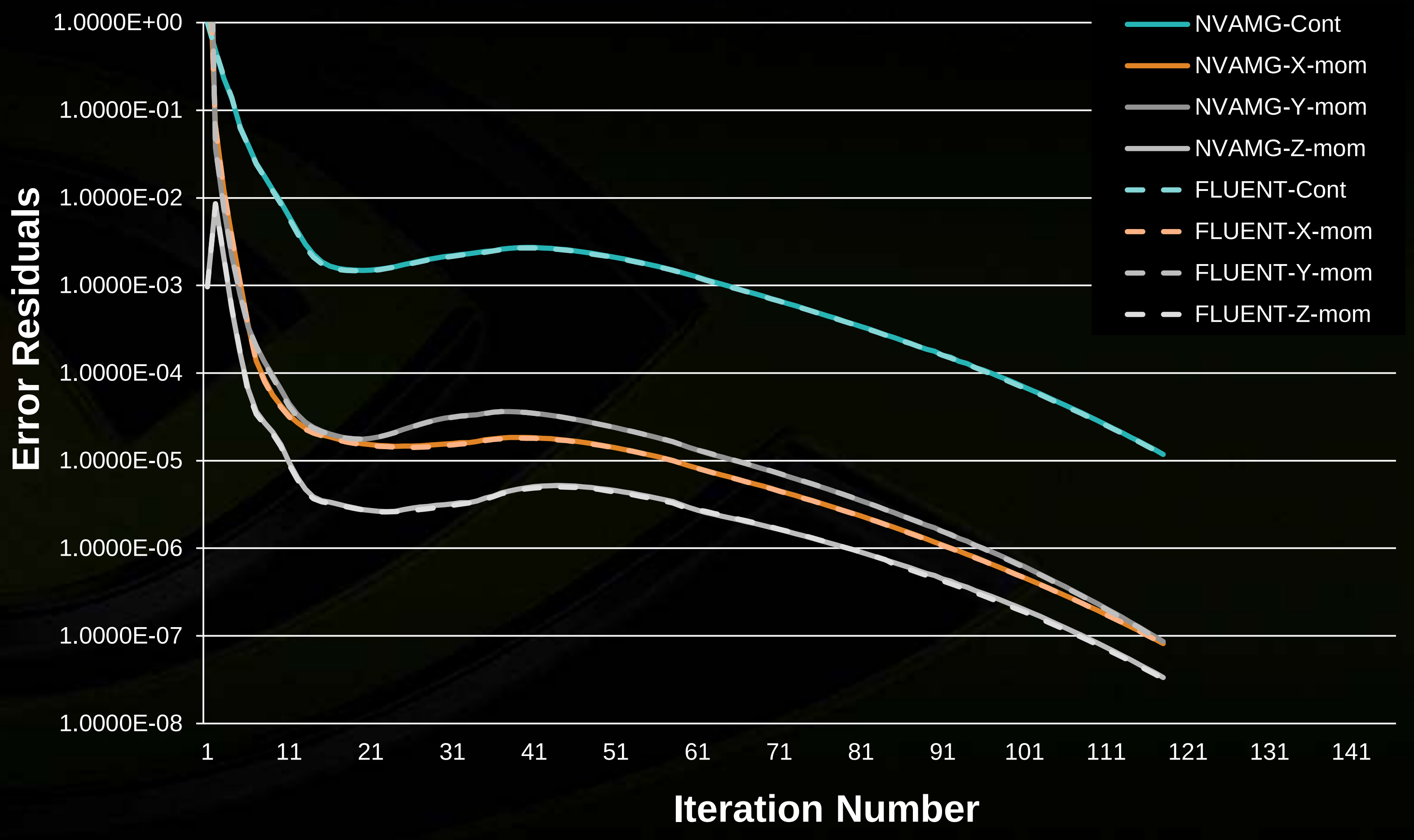


TRUCK BODY MODEL
(14 million cells)



ANSYS Fluent 14.5 GPU Solver Convergence

Preview of ANSYS Fluent Convergence Behavior Matched CPU



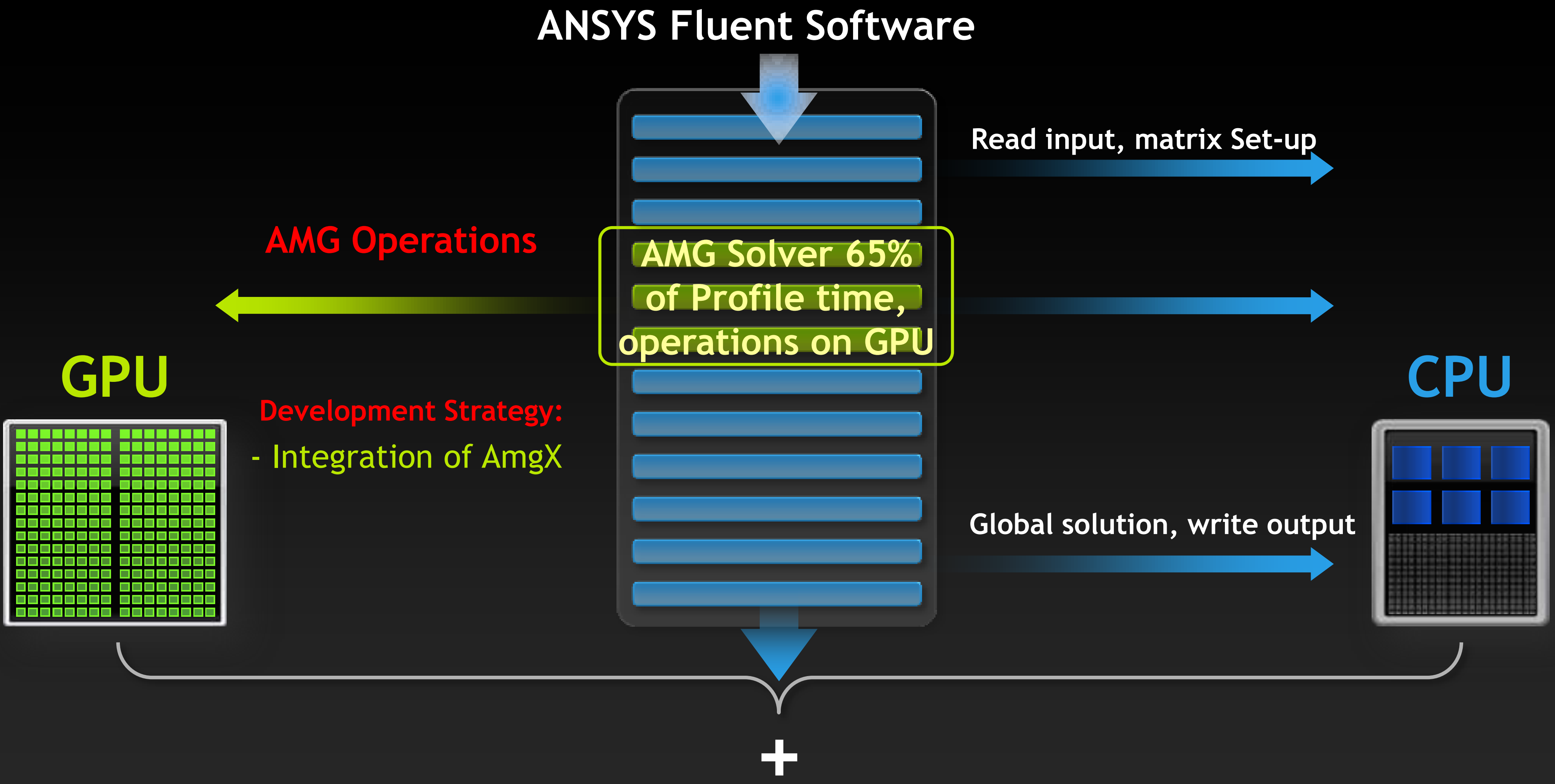
Numerical Results

Mar 2012: Test for convergence at each iteration matches precise Fluent behavior

Model FL5S1:

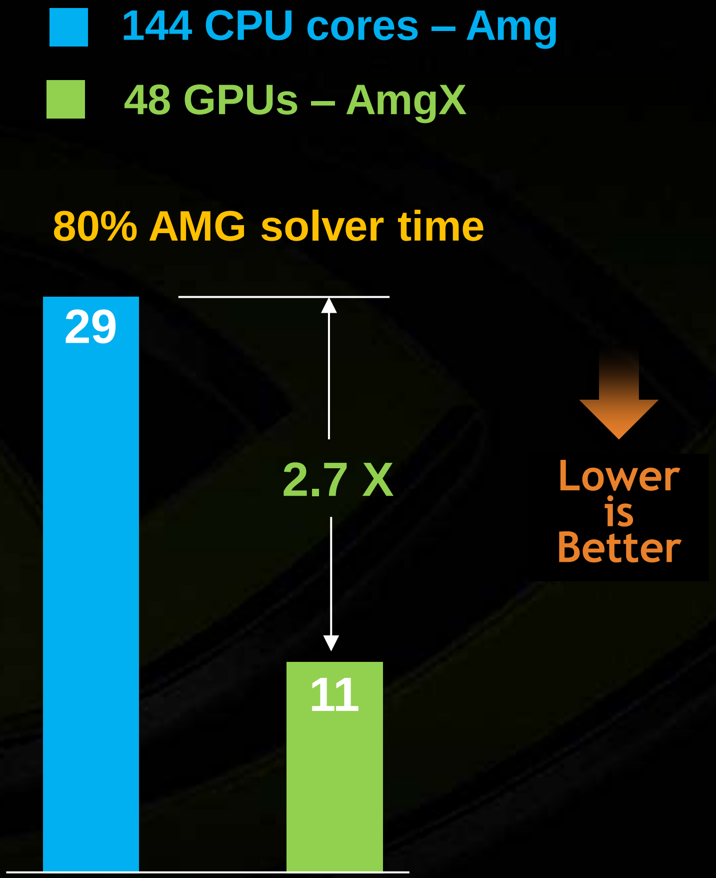
- Incompressible
- Flow in a Bend
- 32K Hex Cells
- Coupled Solver

ANSYS Fluent and NVIDIA AmgX Solver Library

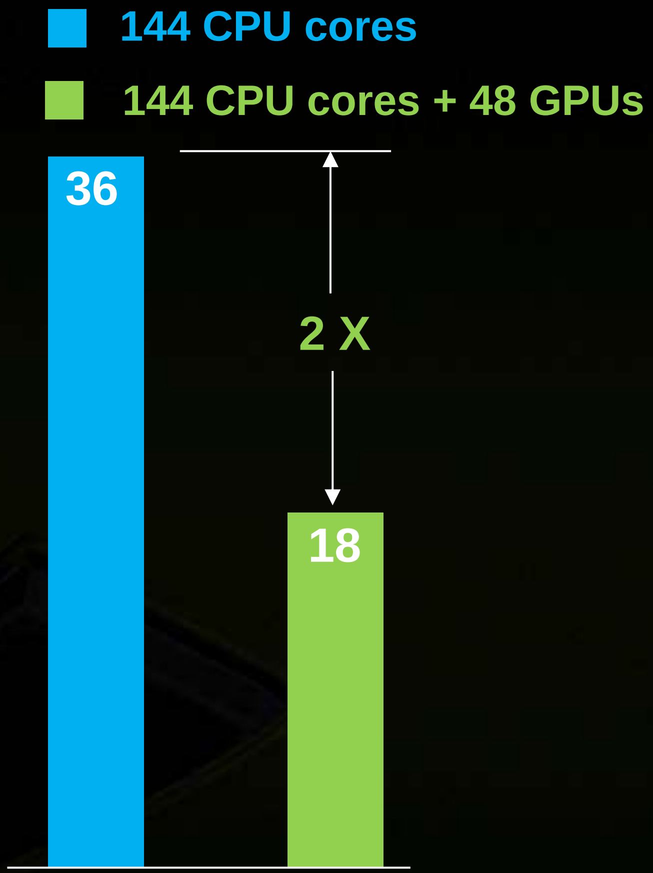


ANSYS Fluent 15 Performance for 111M Cells

ANSYS Fluent 15.0 Performance – Results by NVIDIA, Dec 2013

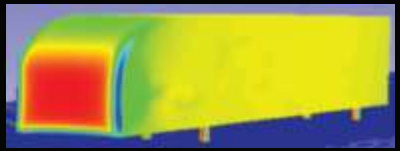


AMG solver time per iteration (secs)



Fluent solution time per iteration (secs)

Truck Body Model

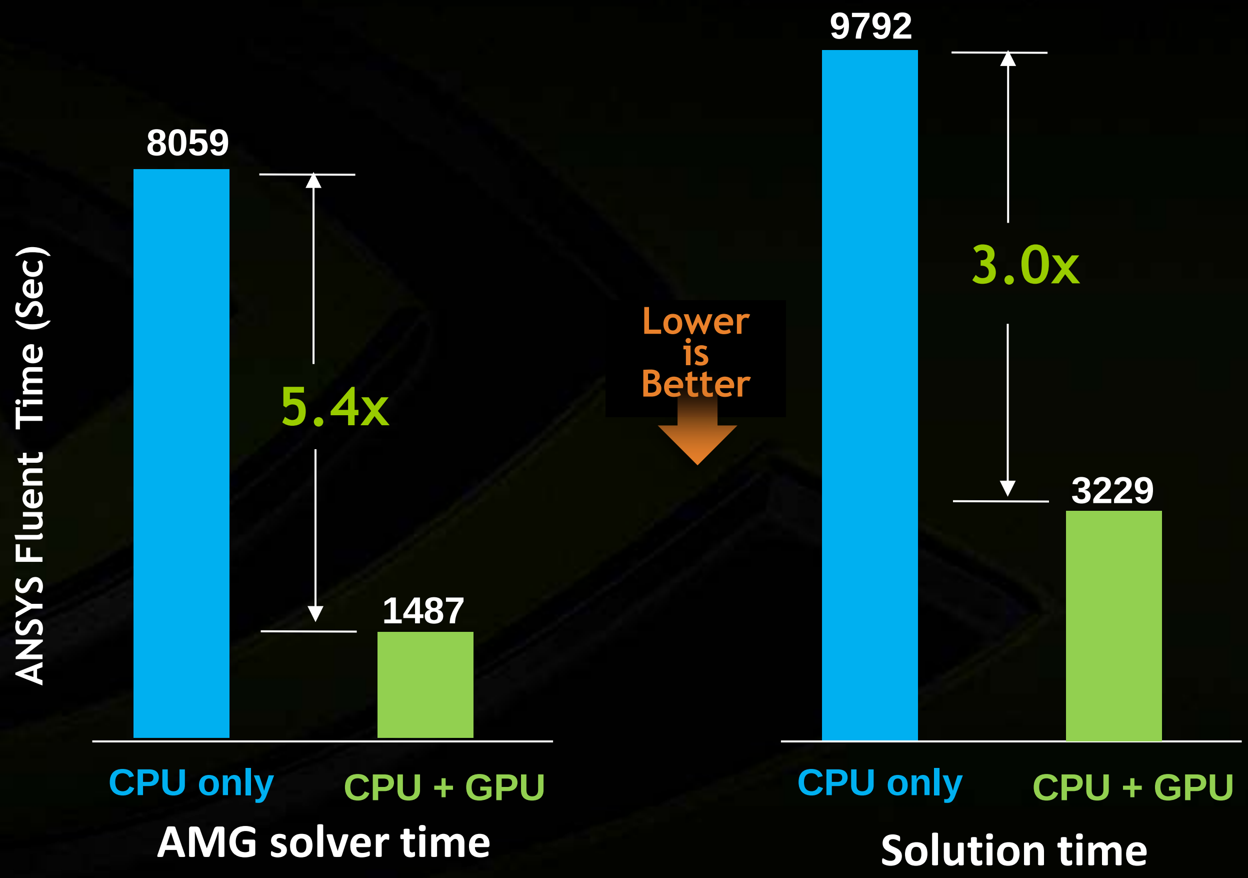


- 111M mixed cells
- External aerodynamics
- Steady, k-ε turbulence
- Double-precision solver
- CPU: Intel Xeon E5-2667; 12 cores per node
- GPU: Tesla K40, 4 per node

NOTE: AmgX is a linear solver toolkit from NVIDIA, used by ANSYS

ANSYS Fluent 16 Performance for 14M Cells

ANSYS Fluent 16.0 Performance – Results by NVIDIA, Dec 2014



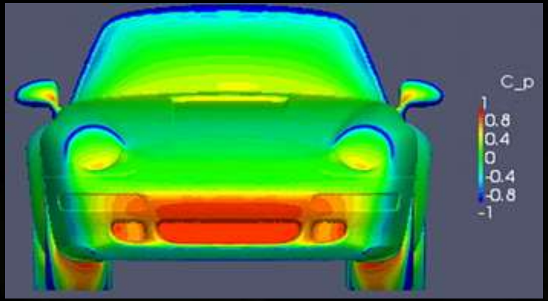
Truck Body Model

- Steady RANS model
- External flow, 14M cells
- CPU: Intel Xeon E5-2697v2 @ 2.7GHz; 48 cores (2 nodes)
- GPU: 4 X Tesla K80 (2 per node)

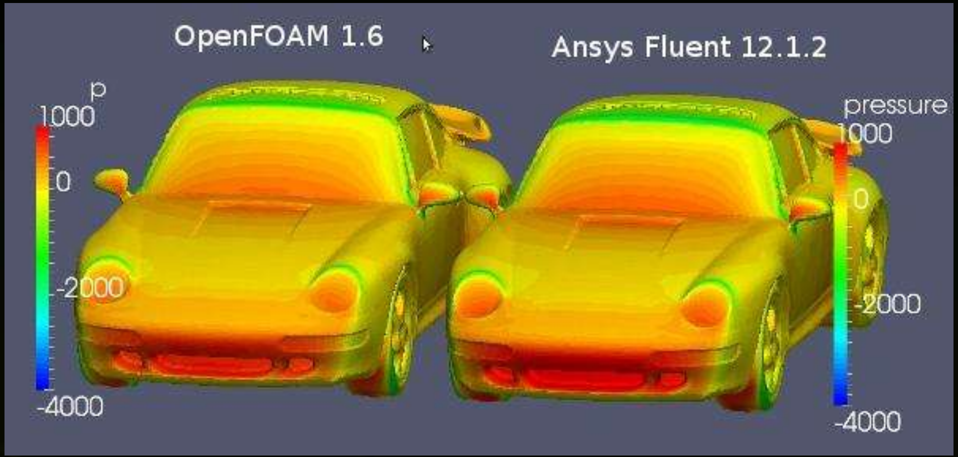
NOTE: Time until convergence

OpenFOAM

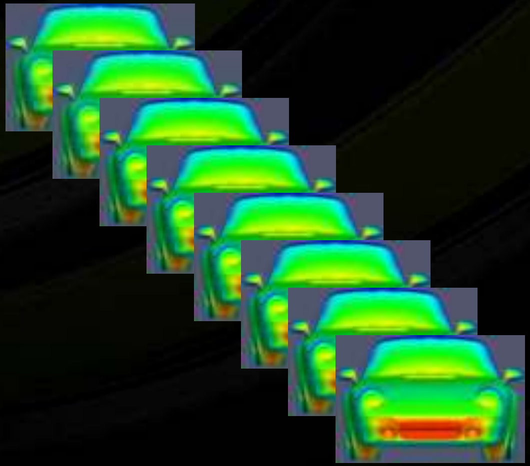
Typical OpenFOAM Use: Parameter Optimization



#1: Develop validated CFD model in ANSYS Fluent or other commercial CFD software in production



#2: Develop CFD model in OpenFOAM, validate against commercial CFD model



#3: Conduct parameter sweeps with OpenFOAM (procedure considered by many users to be cost-prohibitive using commercial CFD license models)

Culises: CFD Solver Library for OpenFOAM

www.fluidyna.de

Culises Easy-to-Use AMG-PCG Solver:

#1. Download and license from http://www.FluiDyna.de

#2. Automatic installation with FluiDyna-provided script

#3. Activate Culises and GPUs with 2 edits to config-file

config-file CPU-only

solvers {
p
solver PCG
preconditioner DIC
tolerance 1e-6
...
}

config-file CPU+GPU

solvers {
p
solver ~~PCG~~ PCGGPU
preconditioner AMG
tolerance 1e-6
...
}

FluiDyna: TU Munich Spin-Off from 2006

Culises provides a linear solver library

Culises requires only two edits to control file of OpenFOAM

Multi-GPU ready

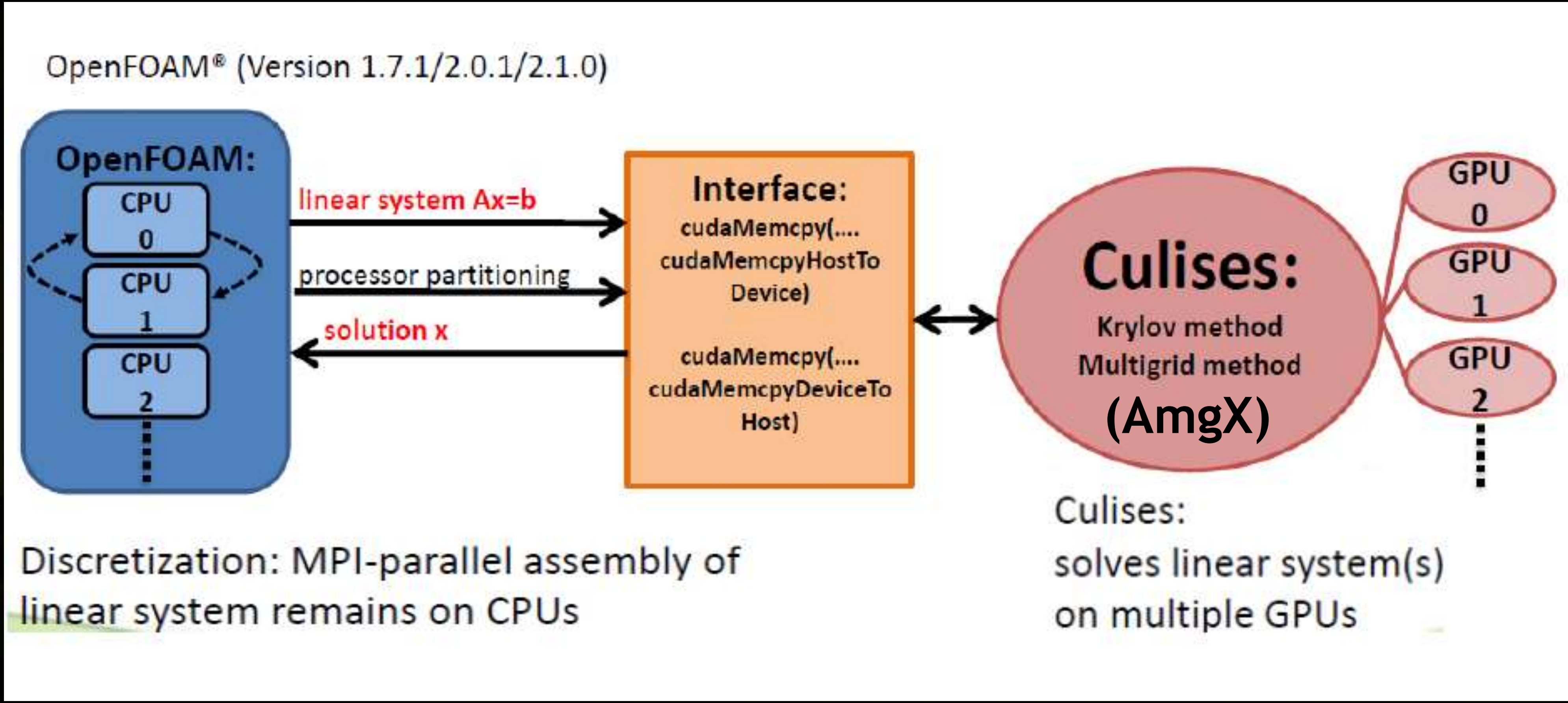
Contact FluiDyna for license details

www.fluidyna.de

Culises (with AmgX) Coupling to OpenFOAM

FluiDyna
www.fluidyna.de

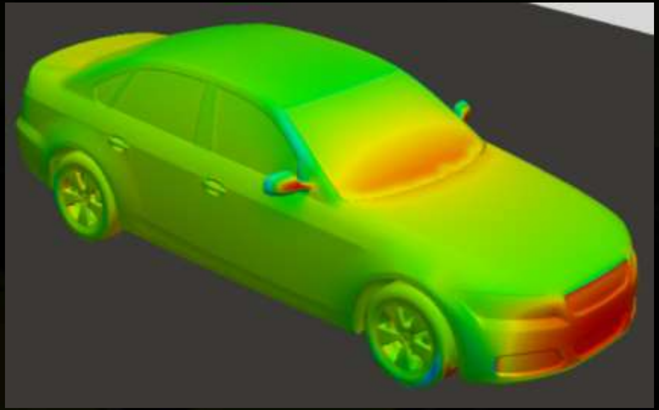
Culises Coupling is User-Transparent:



FluiDyna Culises: CFD Solver for OpenFOAM

Culises: A Library for Accelerated CFD on Hybrid GPU-CPU Systems
Dr. Bjoern Landmann, FluiDyna
developer.download.nvidia.com/GTC/PDF/GTC2012/PresentationPDF/S0293-GTC2012-Culises-Hybrid-GPU.pdf

www.fluidyna.de



Solver speedup of 7x for 2 CPU + 4 GPU

- 36M Cells (mixed type)
- GAMG on CPU
- AMGPCG on GPU

DrivAer: Joint Car Body Shape by BMW and Audi
http://www.aer.mw.tum.de/en/research-groups/automotive/drivaer

Mesh Size - CPUs	9M - 2 CPU	18M - 2 CPU	36M - 2 CPU
GPUs	+1 GPU	+2 GPUs	+4 GPUs
Culises	2.5x	4.2x	6.9x
OpenFOAM	1.36x	1.52x	1.67x

ANSYS Fluent Investigation of DrivAer

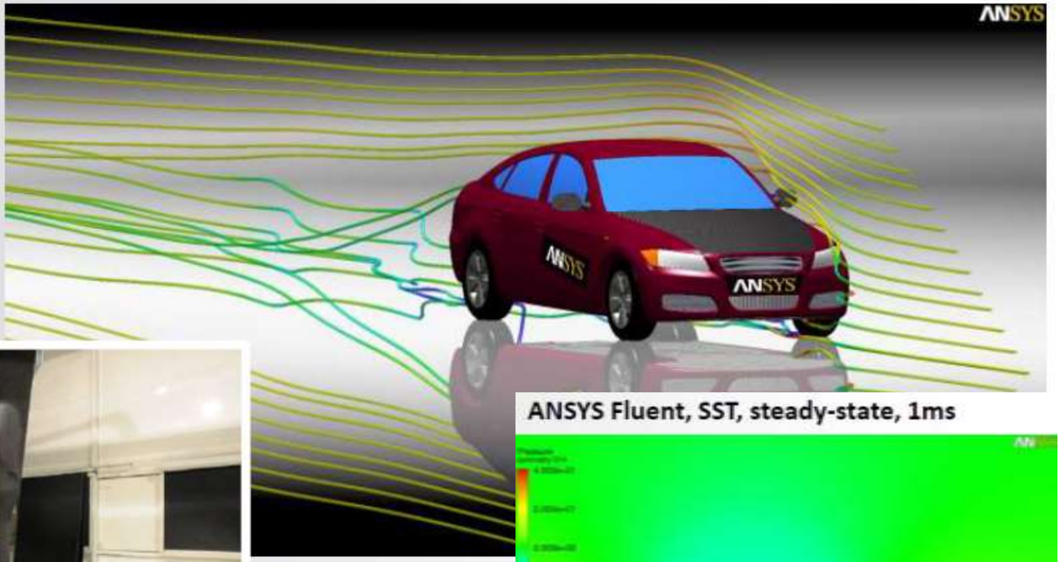
NVIDIA

DrivAer Validation Project: RANS & SRS Modeling

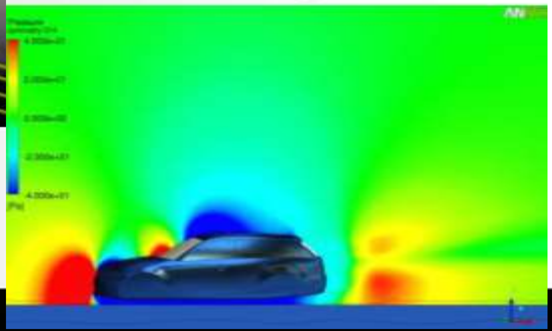
2012 Automotive Simulation World Congress

http://www.aer.mw.tum.de

Courtesy by TU Munich, Inst. of Aerodynamics



ANSYS Fluent, SST, steady-state, 1ms



Source: ANSYS Automotive Simulation World Congress, 30 Oct 2012 – Detroit, MI

"Overview of Turbulence Modeling"

By Dr. Paul Galpin, ANSYS, Inc.

Available ANSYS models

	Mesh 1	Mesh 2	Mesh 2 Full Domain
Elements	17,493,930	56,568,437	113,136,874
Nodes	6,778,624	18,992,636	37,901,816
Max. Aspect Ratio	30,092.5	27,493.5	27,493.5
Min. Grid Angle [degree]	12.8513	8.9184	8.9184
Avg. Y+ (Avg. over car-surface)	0.958305	0.968906	0.952016

Particle-Based CFD for GPUs

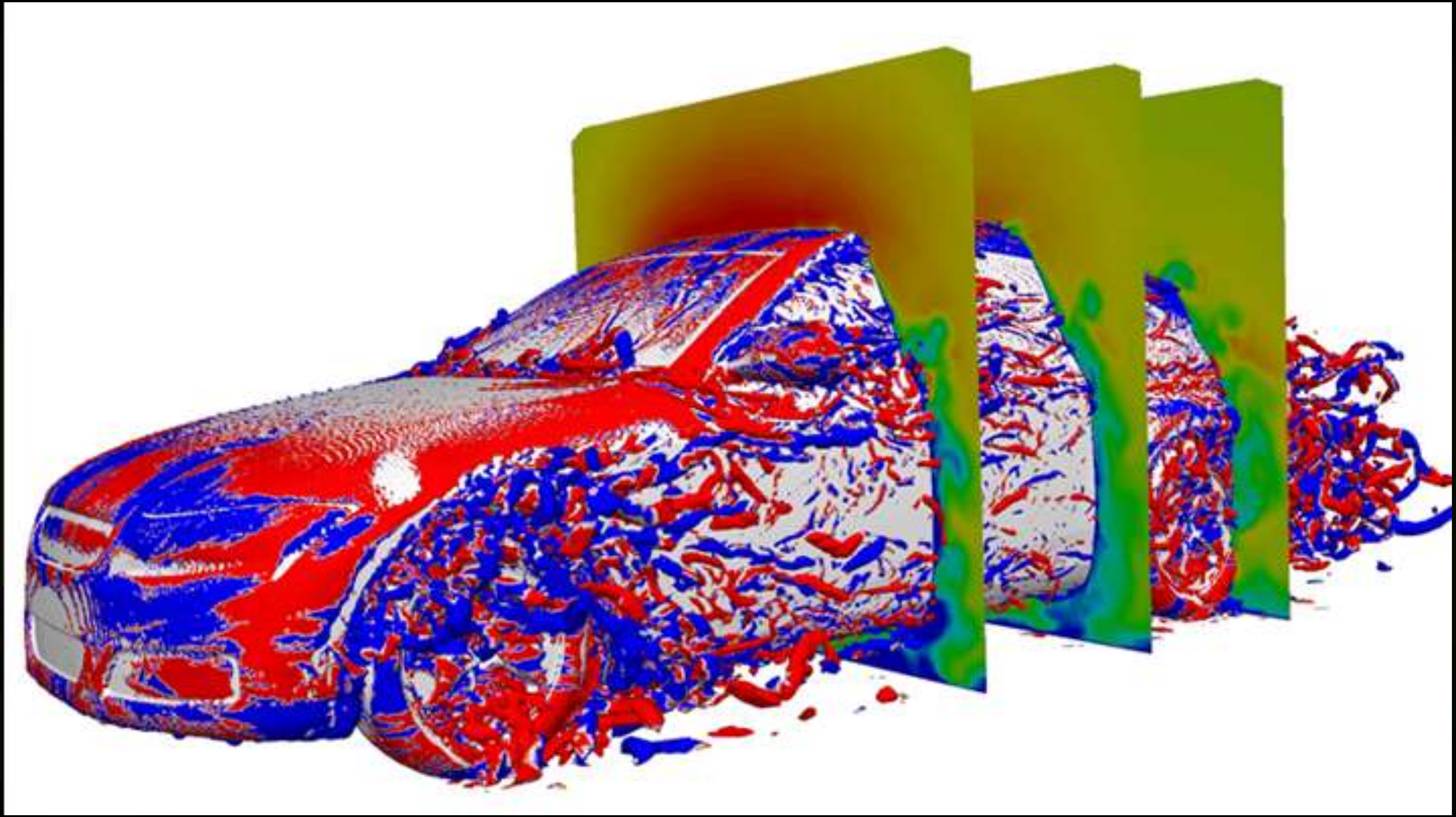
Particle-Based Commercial CFD Software Growing

ISV	Software	Application	Method	GPU Status
Exa	PowerFLOW	Aerodynamics	LBM	Evaluation
FluiDyna	LBultra	Aerodynamics	LBM	Available v2.0
Next Limit Technologies	XFlow	Aerodynamics	LBM	Evaluation
Autodesk	Project Falcon	Aerodynamics	LBM	Evaluation
Prometech	Particleworks	Multiphase/FS	MPS (~SPH)	Available v3.5
cpfd Computational Particle Fluid Dynamics	BARRACUDA	Multiphase/FS	MP-PIC	In development
DEM Solutions	EDEM	Discrete phase	DEM	In development
ANSYS	ANSYS Fluent–DDPM	Multiphase/FS	DEM	In development
CD-adapco	STAR-CCM+	Multiphase/FS	DEM	Evaluation
IMPETUS	AFEA	High impact	SPH	Available v2.0
esi	ESI	High impact	SPH, ALE	In development
LSTC	LSTC	High impact	SPH, ALE	Evaluation
Altair	Altair	High impact	SPH, ALE	Evaluation

FluiDyna Lattice Boltzmann Solver ultraFluidX

http://www.fluidyna.com/content/ultrafluidx

www.fluidyna.de



Spin-Off in 2006 from TU Munich

CFD solver using Lattice Boltzmann method (LBM)

Demonstrated 25x speedup single GPU

Multi-GPU ready

Contact FluiDyna for license details

TiTech Aoki Lab LBM Solution of External Flows

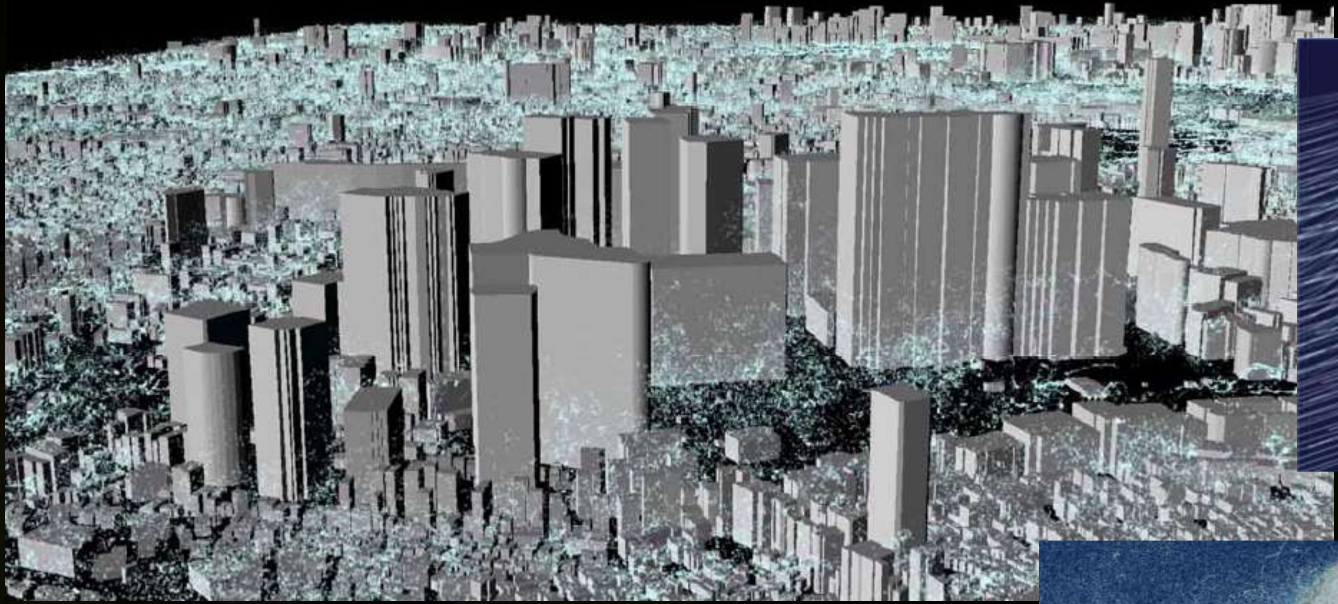
A Peta-scale LES (Large-Eddy Simulation) for Turbulent Flows Based on Lattice Boltzmann Method, Prof. Dr. Takayuki Aoki

http://registration.gputechconf.com/quicklink/8Is4ClC

GPU TECHNOLOGY CONFERENCE

AOKI Lab.

www.sim.gsic.titech.ac.jp



Aoki CFD solver using Lattice Boltzmann method (LBM) with Large Eddy Simulation (LES)

Summary: GPU Suitability and Progress of CFD

- **NVIDIA observes strong CFD community interest in GPU acceleration**
  - New technologies in 2016: Pascal, NVLink, more CPU platform choices
  - NVIDIA business and engineering collaborations in many CFD projects
  - Investments in OpenACC: PGI release of 15.3; Continued Cray collaborations

- **GPU progress for several CFD applications – we examined a few of these**
  - NVIDIA AmgX linear solver library for iterative implicit solvers
  - OpenACC for legacy Fortran-based CFD
  - Novel use of DSLs, templates, Thrust, source-to-source translation

- **Check for updates on continued collaboration with NASA (and SGI)**
  - Further developments for FUN3D undergoing review at NASA LaRC
  - Collaborations with NASA GSFC ongoing with climate model and other

Thank you and Questions?

Stan Posey; sposey@nvidia.com; NVIDIA, Santa Clara, CA, USA