



The Visual Computing Company

GPU Acceleration Benefits for Applied CAE

Axel Koehler, Senior Solutions Architect HPC, NVIDIA

HPC Advisory Council Meeting, April 2014, Lugano

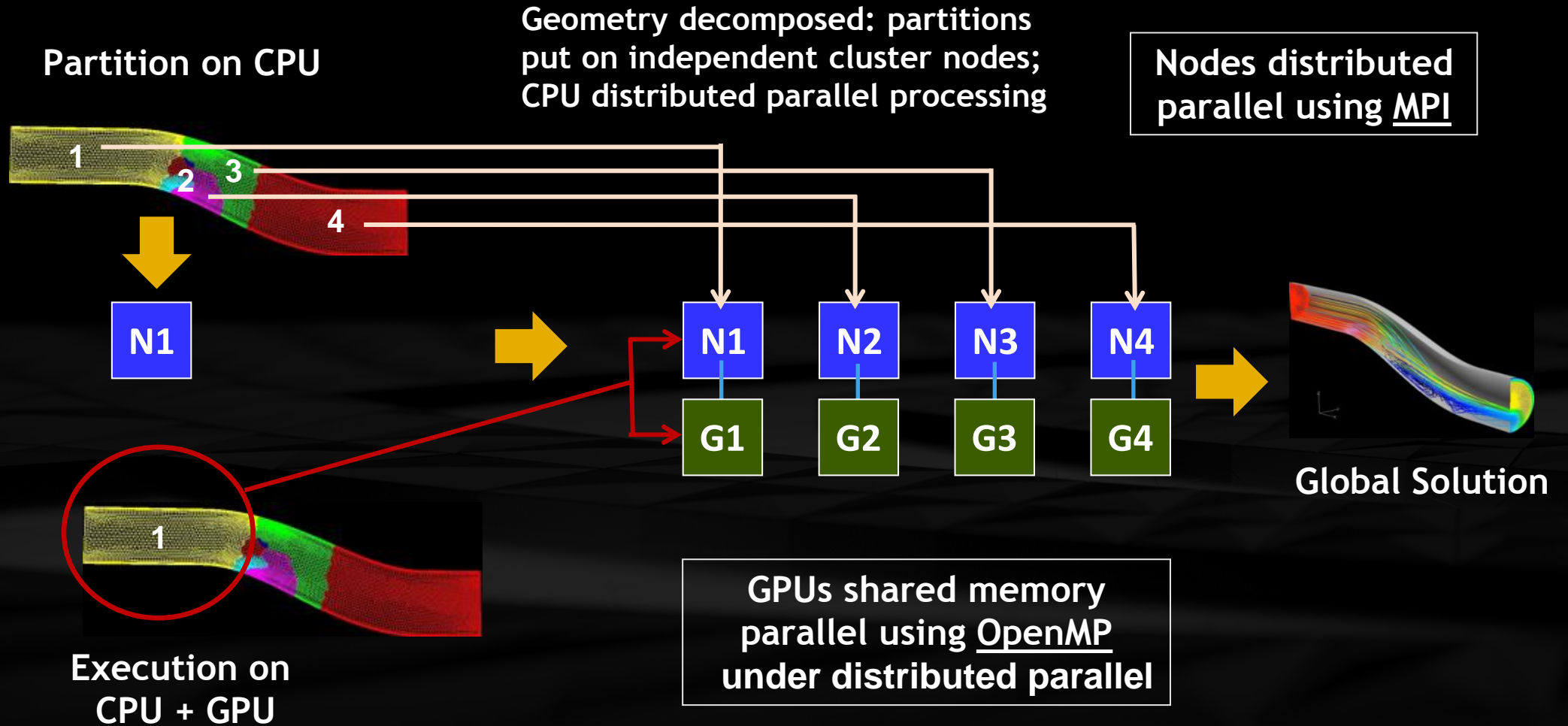
Outline

- General overview about GPU efforts in CAE
- Computational Structural Mechanics (CSM)
 - ANSYS Mechanical, SIMULIA Abaqus/Standard, MSC Nastran, MSC Marc
- Computational Fluid Dynamics (CFD)
 - ANSYS Fluent, OpenFOAM (FluiDyna, PARALUTION)
- Computational Electromagnetics (CEM)
 - CST Studio Suite
- Conclusion

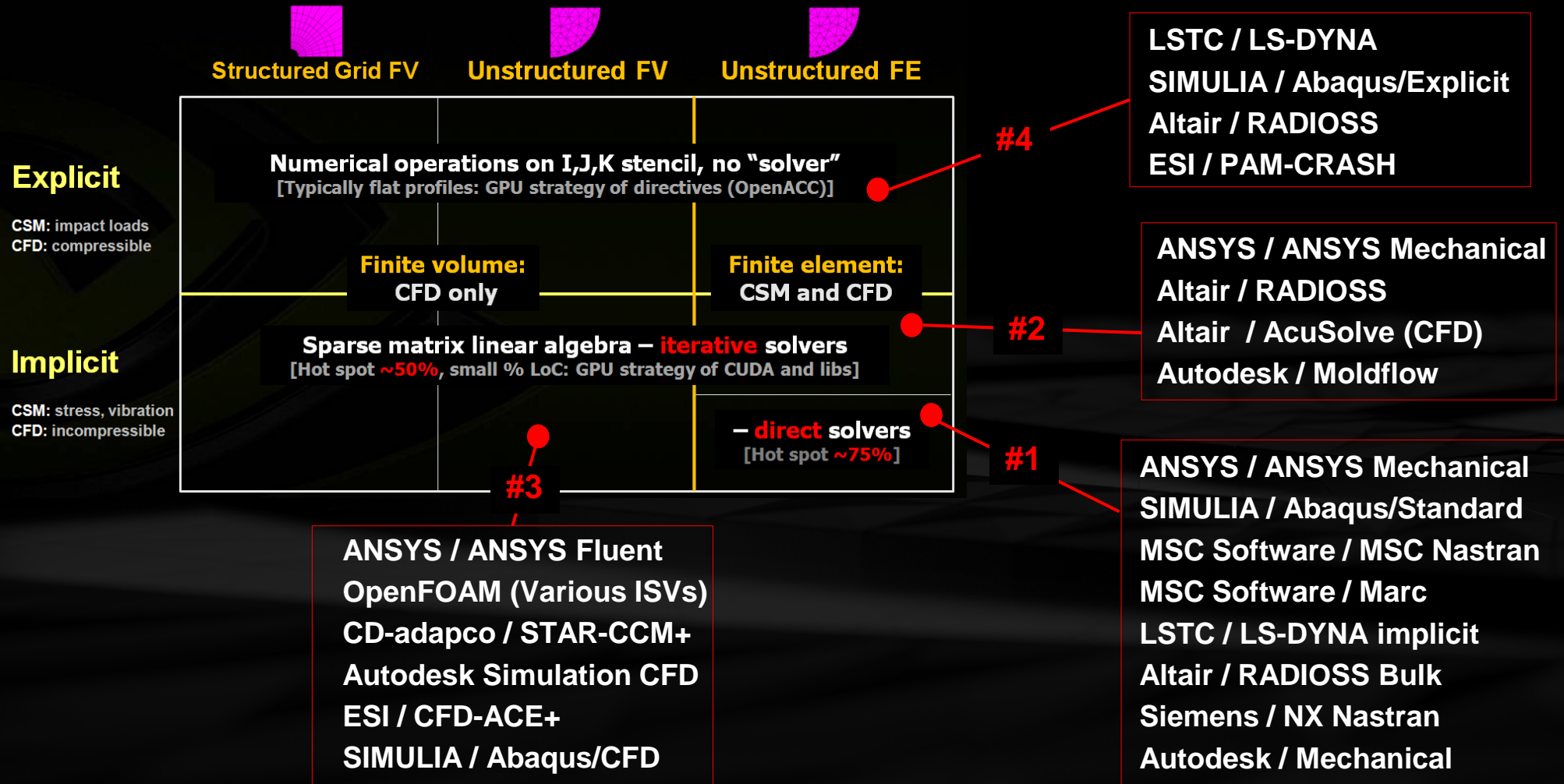
Status Summary of ISVs and GPU Computing

- Every primary ISV has products available on GPUs or undergoing evaluation
- The 4 largest ISVs all have products based on GPUs
 - #1 ANSYS, #2 DS SIMULIA, #3 MSC Software, and #4 Altair
- The top 4 out of 5 ISV applications are available on GPUs today
 - ANSYS Fluent, ANSYS Mechanical, SIMULIA Abaqus/Standard, MSC Nastran, (LS-DYNA implicit only)
- In addition several other ISV applications are already ported to GPUs
 - AcuSolve, OptiStruct (Altair), NX Nastran (Siemens), Permas (Intes), Fire (AVL), Moldflow(Autodesk), AMLS, FastFRS (CDH),
- Several new ISVs were founded with GPUs as a primary competitive strategy
 - Prometech, FluiDyna, Vratis, IMPETUS, Turbostream, ...
- Open source CFD OpenFOAM available on GPUs today with many options
 - Commercial options: FluiDyna, Vratis; Open source options: Cufflink, Symscape ofgpu, RAS, etc.

GPUs and Distributed Cluster Computing



CAE Priority for ISV Software Development on GPUs

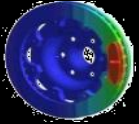

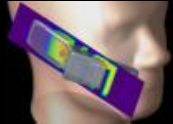


Computational Structural Mechanics

ANSYS Mechanical

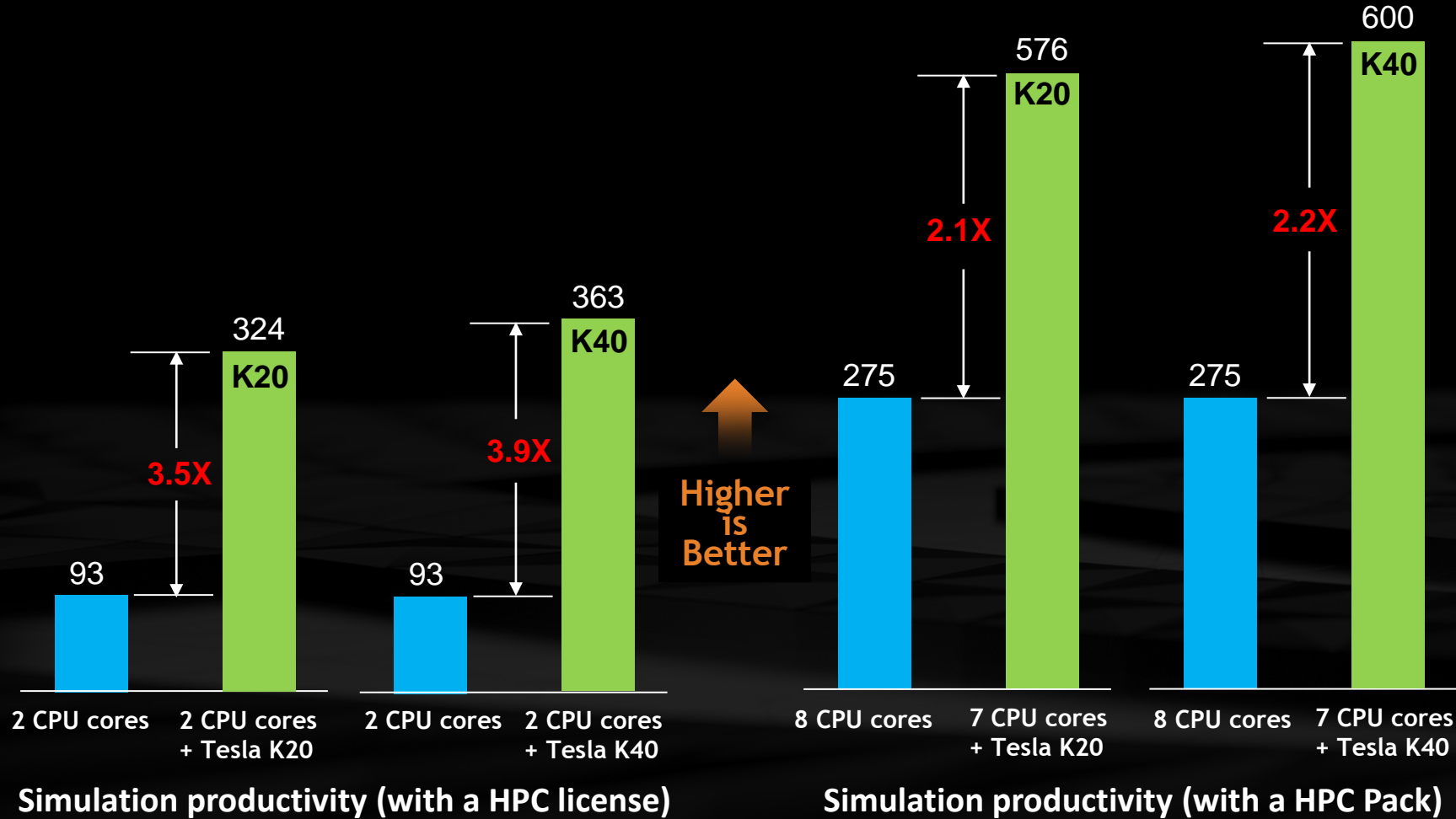
ANSYS and NVIDIA Collaboration Roadmap



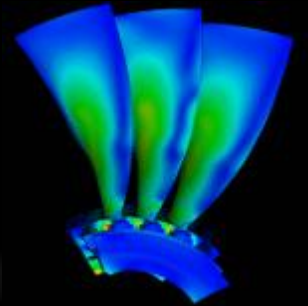
Release	ANSYS Mechanical 	ANSYS Fluent 	ANSYS EM 
13.0 Dec 2010	SMP, Single GPU, Sparse and PCG/JCG Solvers		ANSYS Nexxim
14.0 Dec 2011	+ Distributed ANSYS; + Multi-node Support	Radiation Heat Transfer (beta)	ANSYS Nexxim
14.5 Nov 2012	+ Multi-GPU Support; + Hybrid PCG; + Kepler GPU Support	+ Radiation HT; + GPU AMG Solver (beta), Single GPU	ANSYS Nexxim
15.0 Dec 2013	+ CUDA 5 Kepler Tuning	+ Multi-GPU AMG Solver; + CUDA 5 Kepler Tuning	ANSYS Nexxim ANSYS HFSS (Transient)

ANSYS Mechanical 15.0 on Tesla GPUs

ANSYS Mechanical jobs/day



V14sp-5 Model



- Turbine geometry
- 2,100,000 DOF
- SOLID187 FEs
- Static, nonlinear
- Distributed ANSYS 15.0
- Direct sparse solver

Considerations for ANSYS Mechanical on GPUs

- Problems with high solver workloads benefit the most from GPU
 - Characterized by both high DOF and high factorization requirements
 - Models with solid elements and have >500K DOF experience good speedups
- Better performance when run on DMP mode over SMP mode
- GPU and system memories both play important roles in performance
 - Sparse solver:
 - If the model exceeds 5M DOF, then either add another GPU with 5-6 GB of memory (Tesla K20 or K20X) or use a single GPU with 12 GB memory (eg. Tesla K40)
 - PCG/JCG solver:
 - Memory saving (MSAVE) option should be turned off for enabling GPUs
 - Models with lower Level of Difficulty value (Lev_Diff) are better suited for GPUs

Computational Structural Mechanics

Abaqus/Standard

SIMULIA and Abaqus GPU Release Progression



- Abaqus 6.11, June 2011
 - Direct sparse solver is accelerated on the GPU
 - Single GPU support; Fermi GPUs (Tesla 20-series, Quadro 6000)
- Abaqus 6.12, June 2012
 - Multi-GPU/node; multi-node DMP clusters
 - Flexibility to run jobs on specific GPUs
 - Fermi GPUs + Kepler Hotfix (since November 2012)
- Abaqus 6.13, June 2013
 - Un-symmetric sparse solver on GPU
 - Official Kepler support (Tesla K20/K20X/K40)

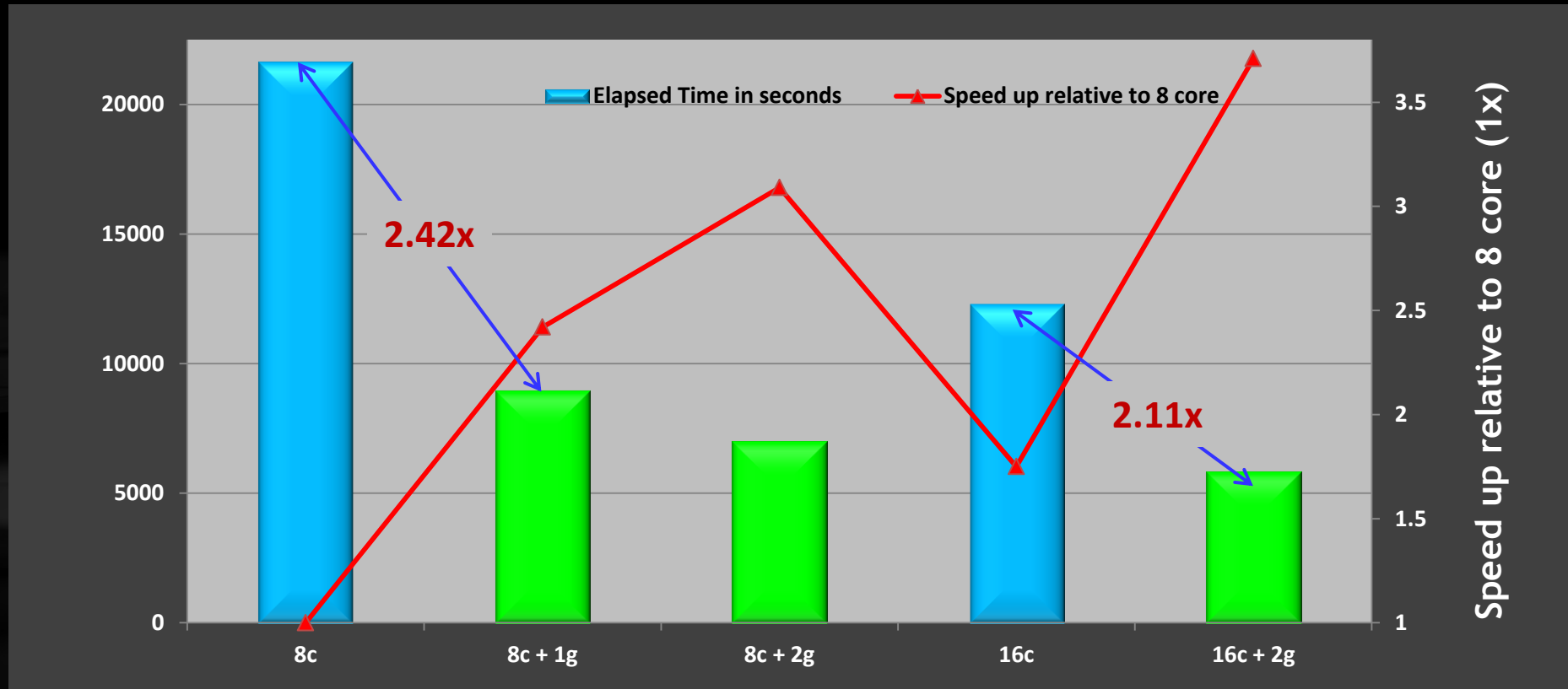
Rolls Royce: Abaqus 3.5x Speedup with 5M DOF



- 4.71M DOF (equations); ~77 TFLOPs
- Nonlinear Static (6 Steps)
- Direct Sparse solver, 100GB memory



Sandy Bridge + Tesla K20X Single Server



Server with 2x E5-2670, 2.6GHz CPUs, 128GB memory, 2x Tesla K20X, Linux RHEL 6.2, Abaqus/Standard 6.12-2

Rolls Royce: Abaqus Speedups on an HPC Cluster

Sandy Bridge + Tesla K20X for 4 x Servers



- 4.71M DOF (equations); ~77 TFLOPs
- Nonlinear Static (6 Steps)
- Direct Sparse solver, 100GB memory



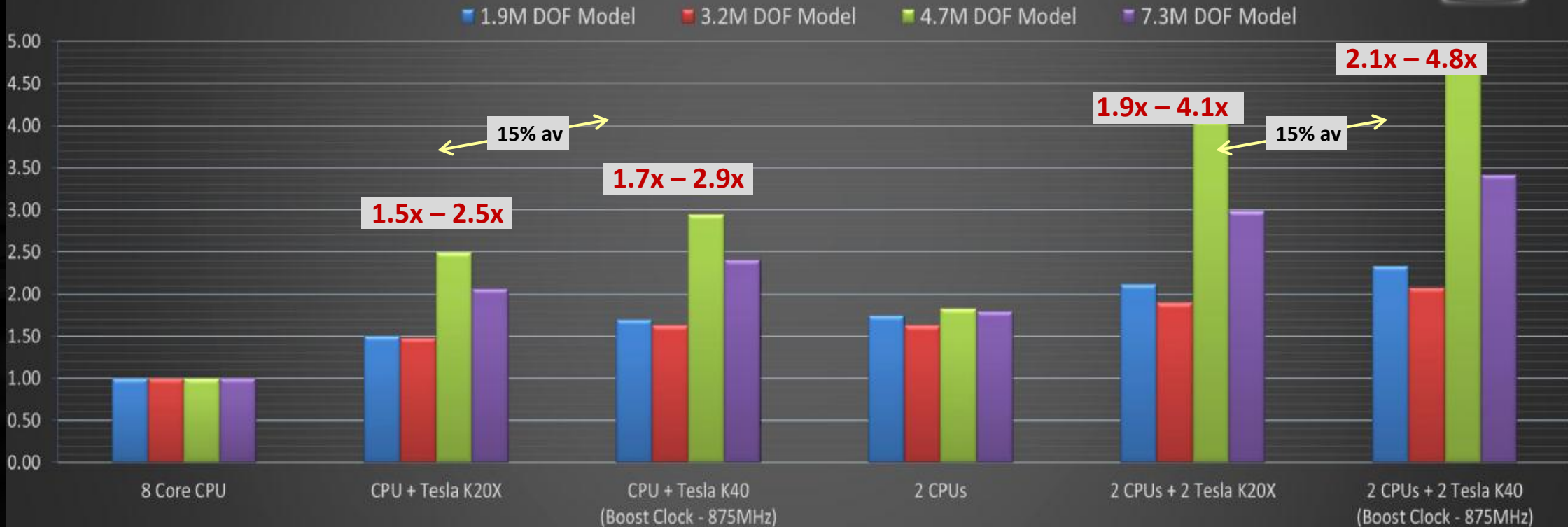
Servers with 2x E5-2670, 2.6GHz CPUs, 128GB memory, 2x Tesla K20X, Linux RHEL 6.2, Abaqus/Standard 6.12-2

Abaqus/Standard ~15% Gain from K20X to K40

ABAQUS STANDARD GPU ACCELERATION

Relative Performance

- Abaqus 6.13 is used on a SuperMicro node with Sandy Bridge CPUs 3.1GHz (8 core), Tesla GPUs

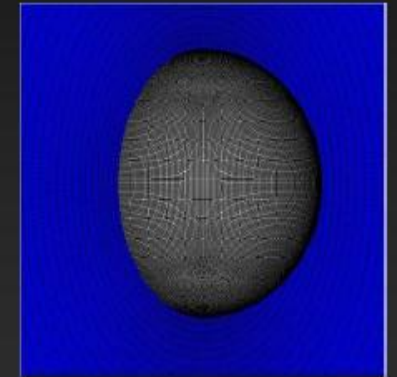
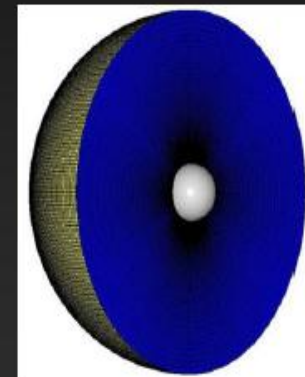


Abaqus 6.13-DEV Scaling on Tesla GPU Cluster



Structural-acoustics coupling

- 1.3M equations; 68 TFLOPs
- Forced frequency response analysis
- Sparse solver (**un-symmetric**)



Abaqus Licensing in a node and across a cluster

Cores	Tokens	GPU	Tokens	GPU	Tokens
1	5	1	6	2	7
2	6	1	7	2	8
3	7	1	8	2	9
4	8	1	9	2	10
5	9	1	10	2	11
6	10	1	11	2	12
7	11	1	12	2	12
8 (1 CPU)	12	1	12	2	13
9	12	1	13	2	13
10	13	1	13	2	14
11	13	1	14	2	14
12	14	1	14	2	15
13	14	1	15	2	15
14	15	1	15	2	16
15	15	1	16	2	16
16 (2 CPUs)	16	1	16	2	16

2 nodes: 2x 16 cores + 2x 2 GPUs

32 cores: 21 tokens

32 cores + 4 GPUs: 22 tokens

3 nodes: 3x 16 cores + 3x 2 GPUs

48 cores: 25 tokens

48 cores + 6 GPUs: 26 tokens

Abaqus 6.12 Power consumption in a node



Abaqus Model: 13 TFLOPs

- 2.15M DoF (equations)
- Nonlinear Static (6 Steps)
- Direct Sparse solver
- ~24GB of memory to reduce IO

Total Estimated Power Consumption

- 8c: 2.14 KWh
- 16c: 1.44 KWh
- 8c + 1g: 1.25 KWh
- 16c + 2g: 1.07 KWh

Elapsed Time (hh:mm:ss)

Computational Structural Mechanics

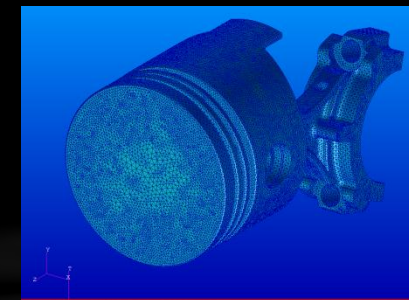
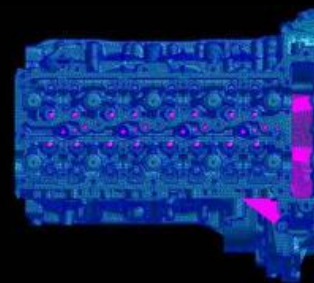
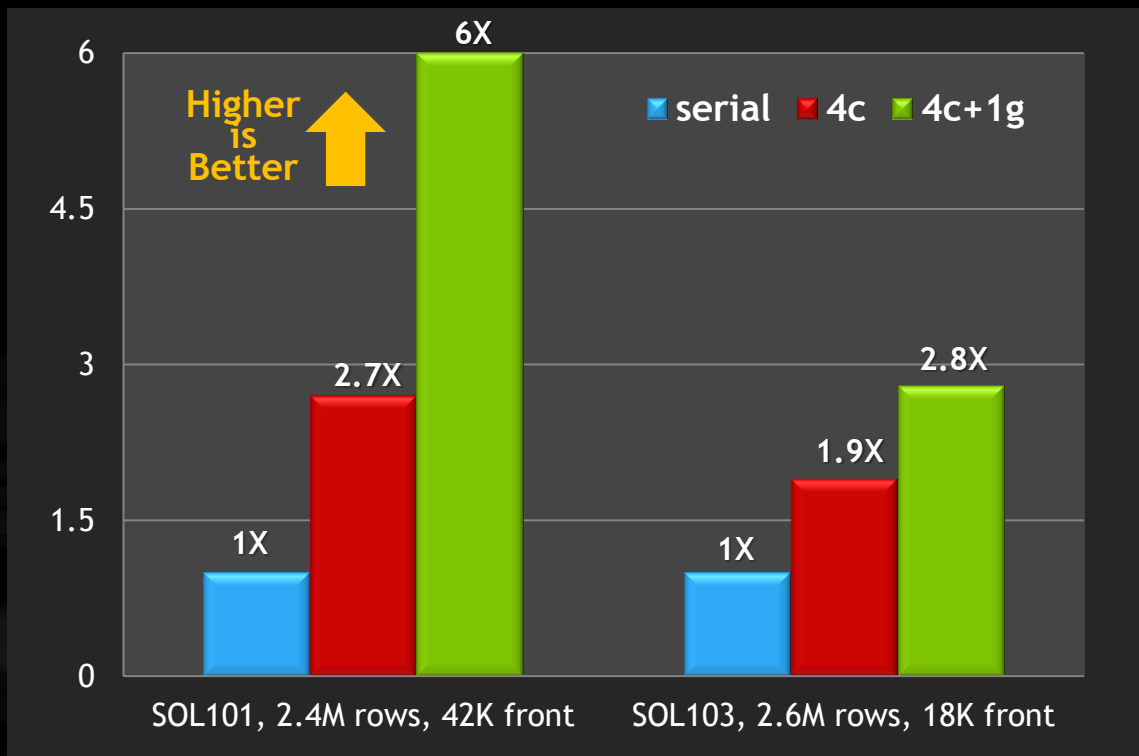
MSC Nastran

MSC Nastran 2013

- Nastran direct equation solver is GPU accelerated
 - Sparse direct factorization (MSCLDL, MSCLU)
 - Real, Complex, Symmetric, Un-symmetric
 - Handles very large fronts with minimal use of pinned host memory
 - Lowest granularity GPU implementation of a sparse direct solver; solves unlimited sparse matrix sizes
 - Impacts several solution sequences:
 - High impact (SOL101, SOL108), Mid (SOL103), Low (SOL111, SOL400)
- Support of multi-GPU and for Linux and Windows
 - With DMP > 1, multiple fronts are factorized concurrently on multiple GPUs; 1 GPU per matrix domain
 - NVIDIA GPUs include Tesla 20-series, Tesla K20/K20X, Tesla K40, Quadro 6000
 - CUDA 5 and later

MSC Nastran 2013

SMP + GPU acceleration of SOL101 and SOL103



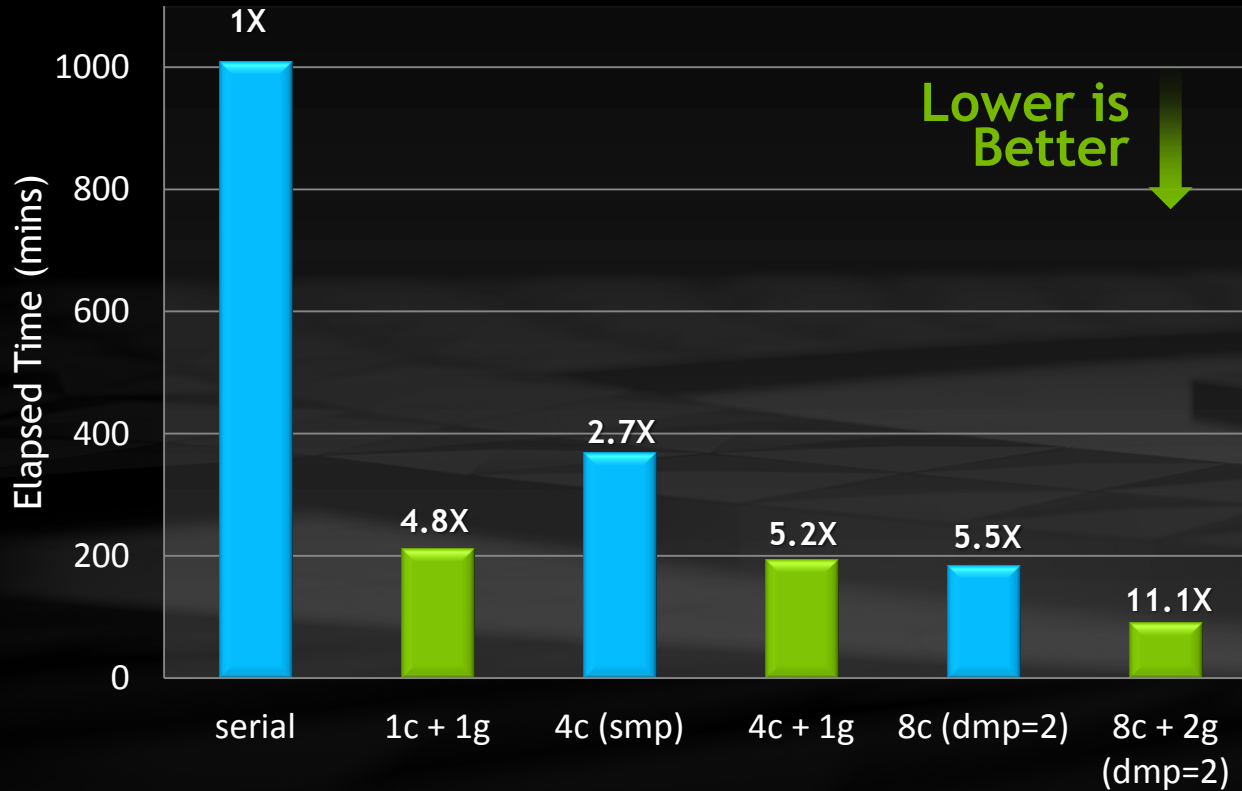
Lanczos solver (SOL 103)

- Sparse matrix factorization
- Iterate on a block of vectors (solve)
- Orthogonalization of vectors

Server node: Sandy Bridge E5-2670 (2.6GHz), Tesla K20X GPU, 128 GB memory

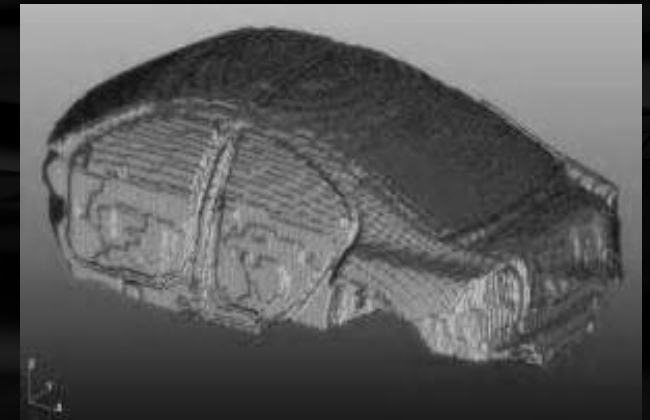
MSC Nastran 2013

Coupled Structural-Acoustics simulation with SOL108



Europe Auto OEM

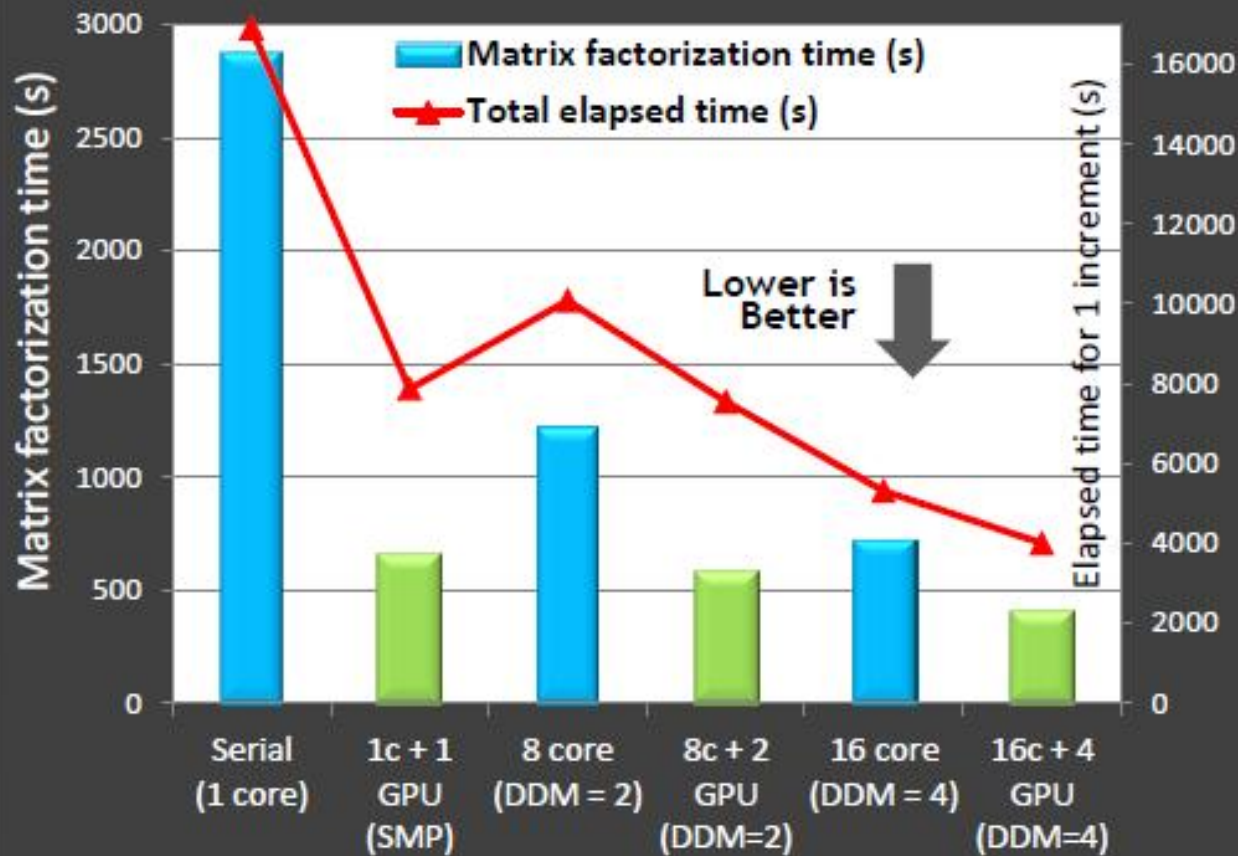
- 710K nodes, 3.83M elements
- 100 frequency increments (FREQ1)
- Direct Sparse solver



Server node: Sandy Bridge 2.6GHz, 2x 8 core, Tesla 2x K20X GPU, 128GB memory

Computational Structural Mechanics

MSC MARC





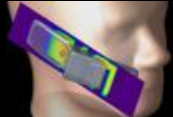
2.5 Million Elements
 10 Million DOF
 Nonlinear Bolt Tightening
 12 increments, 48 cycles

Computational Fluid Dynamics

ANSYS Fluent

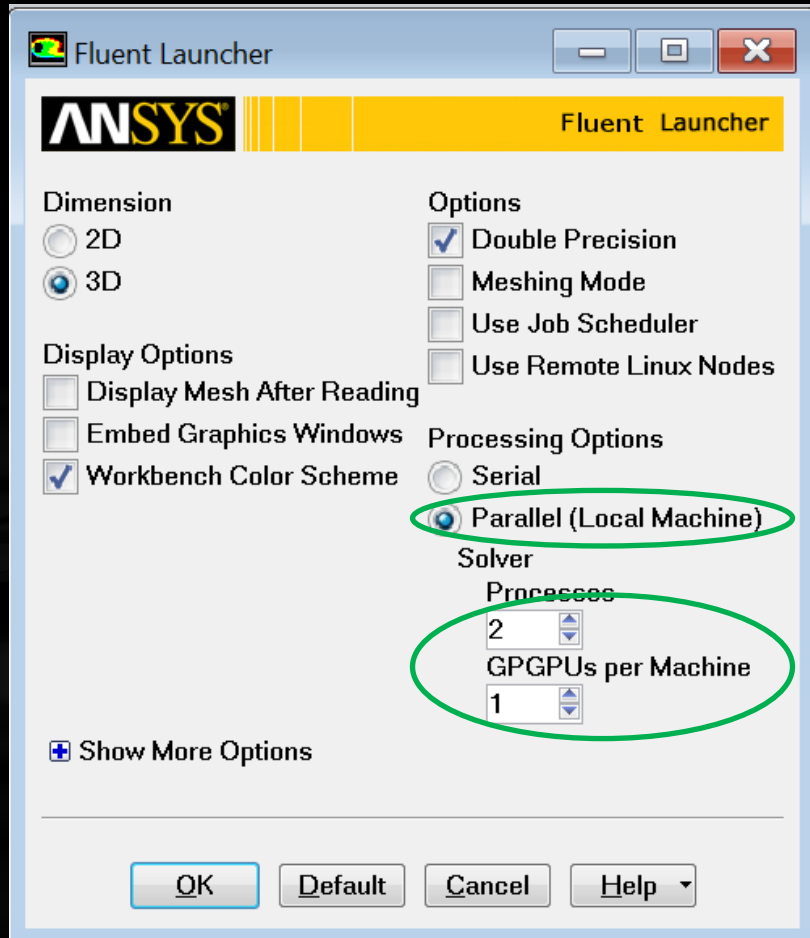
ANSYS and NVIDIA Collaboration Roadmap



Release	ANSYS Mechanical 	ANSYS Fluent 	ANSYS EM 
13.0 Dec 2010	SMP, Single GPU, Sparse and PCG/JCG Solvers		ANSYS Nexxim
14.0 Dec 2011	+ Distributed ANSYS; + Multi-node Support	Radiation Heat Transfer (beta)	ANSYS Nexxim
14.5 Nov 2012	+ Multi-GPU Support; + Hybrid PCG; + Kepler GPU Support	+ Radiation HT; + GPU AMG Solver (beta), Single GPU	ANSYS Nexxim
15.0 Dec 2013	+ CUDA 5 Kepler Tuning	+ Multi-GPU AMG Solver; + CUDA 5 Kepler Tuning	ANSYS Nexxim ANSYS HFSS (Transient)

How to Enable NVIDIA GPUs in ANSYS Fluent

Windows:



Linux:

```
fluent 3ddp -g -ssh -t2 -gpgpu=1 -i journal.jou
```

Cluster specification:

$nprocs$ = Total number of fluent processes

M = Number of machines

$ngpgpus$ = Number of GPUs per machine

Requirement 1

$$nprocs \bmod M = 0$$

Same number of solver processes on each machine

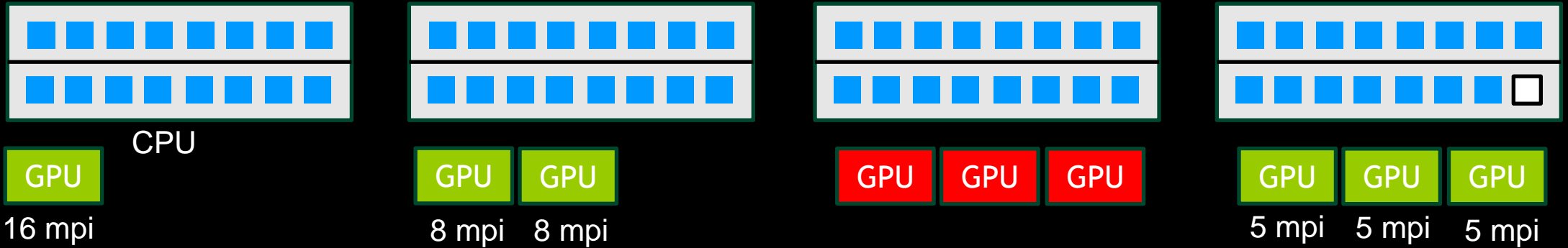
Requirement 2

$$\left(\frac{nproc}{M}\right) \bmod ngpgpus = 0$$

No. of processes should be an integer multiple of GPUs

Cluster Specification Examples

Single-node configurations:



Multi-node configurations:

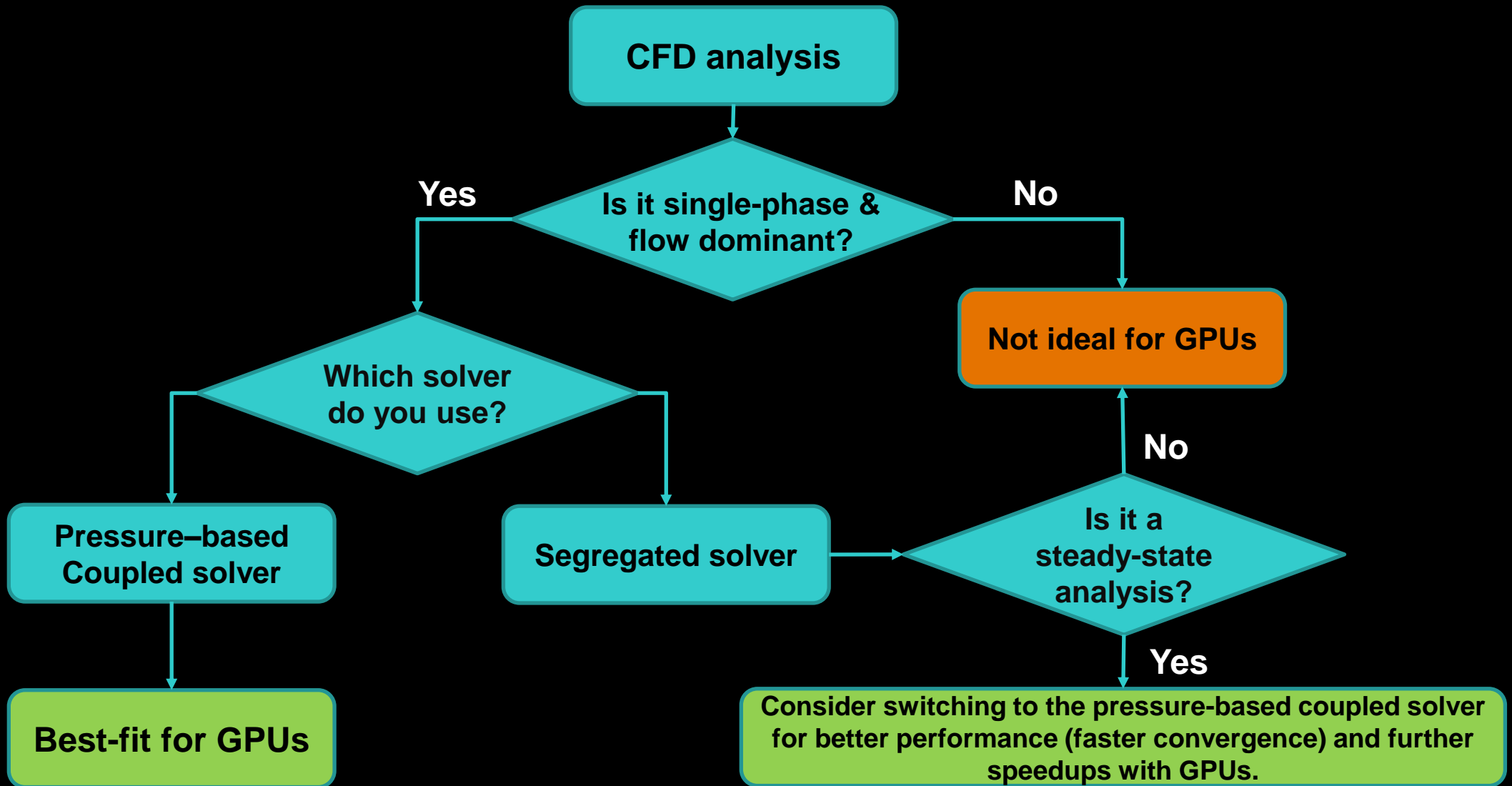


Note: The problem must fit in the GPU memory for the solution to proceed

Considerations for ANSYS Fluent on GPUs

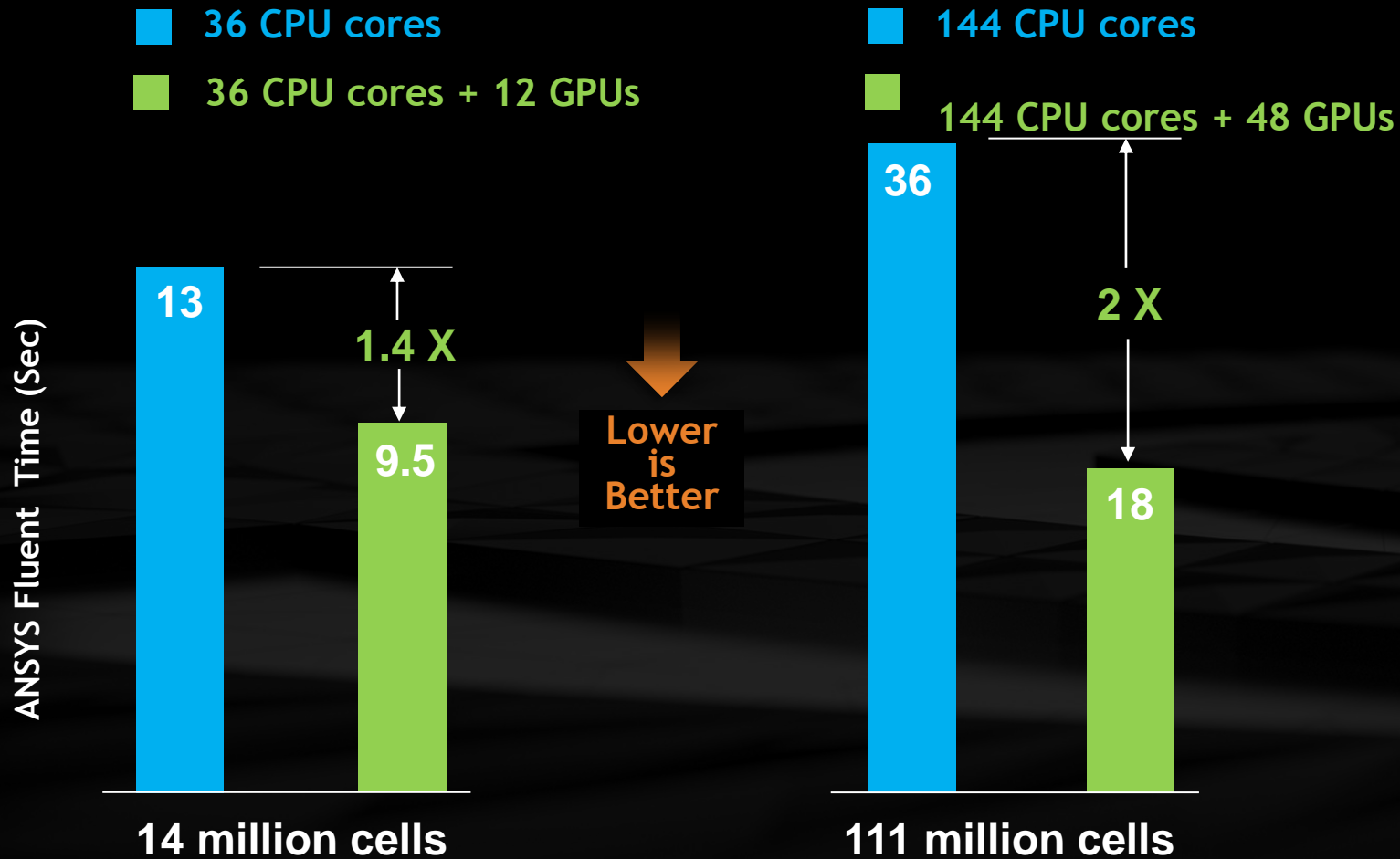
- GPUs accelerate the AMG solver of the CFD analysis
 - Fine meshes and low-dissipation problems have high %AMG
 - Coupled solution scheme spends 65% on average in AMG
- In many cases, pressure-based coupled solvers offer faster convergence compared to segregated solvers (problem-dependent)
- The system matrix must fit in the GPU memory
 - For coupled PBNS, each 1 MM cells need about 4 GB of GPU memory
 - High-memory GPUs such as Tesla K40 or Quadro K6000 are ideal
- Better performance with use of lower CPU core counts
 - A ratio of 4 CPU cores to 1 GPU is recommended

NVIDIA-GPU Solution fit for ANSYS Fluent

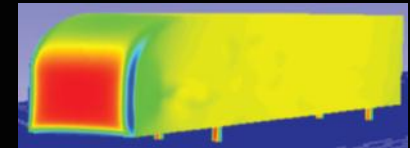


ANSYS Fluent GPU Performance for Large Cases

Better speed-ups on larger and harder-to-solve problems



Truck Body Model

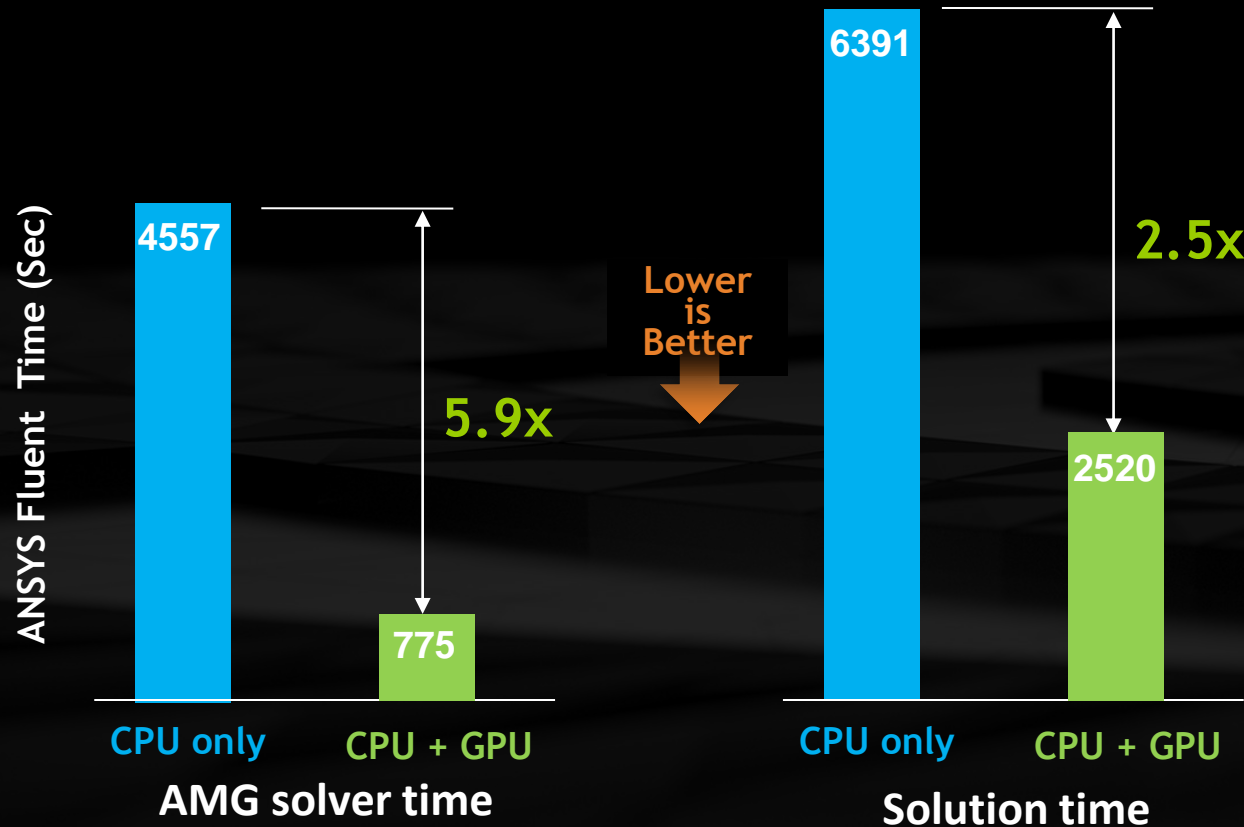


- External aerodynamics
- Steady, k-ε turbulence
- Double-precision solver
- CPU: Intel Xeon E5-2667; 12 cores per node
- GPU: Tesla K40, 4 per node

NOTE: Reported times are Fluent solution time in second per iteration

GPU Acceleration of Water Jacket Analysis

ANSYS Fluent 15.0 performance on pressure-based coupled Solver



Water jacket model

- Unsteady RANS model
- Fluid: water
- Internal flow
- CPU: Intel Xeon E5-2680; 8 cores
- GPU: 2 X Tesla K40

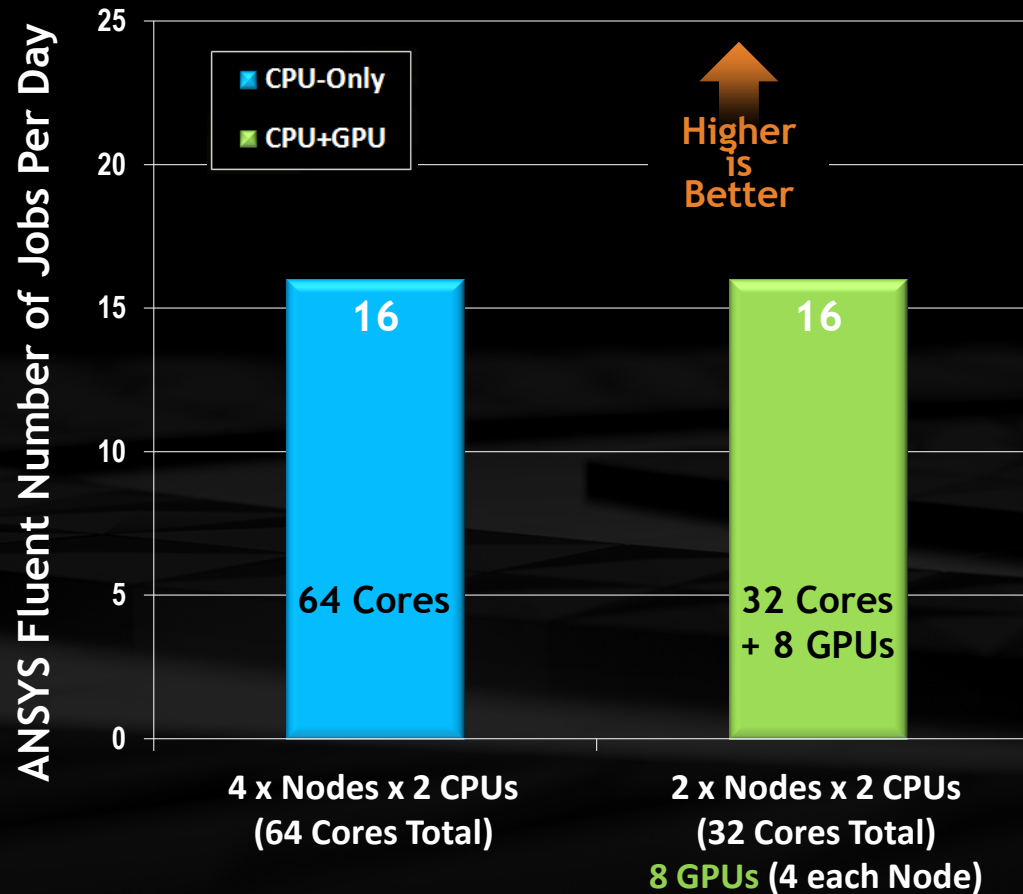
NOTE: Times for 20 time steps

ANSYS Fluent GPU Study on Productivity Gains

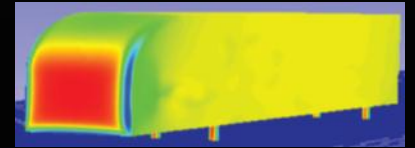
ANSYS Fluent 15.0 Preview 3 Performance - Results by NVIDIA, Sep 2013

ANSYS

- Same solution times:
64 cores vs.
32 cores + 8 GPUs
- Frees up 32 CPUs
and HPC licenses for
additional job(s)
- Approximate 56%
increase in overall
productivity for 25%
increase in cost



Truck Body Model

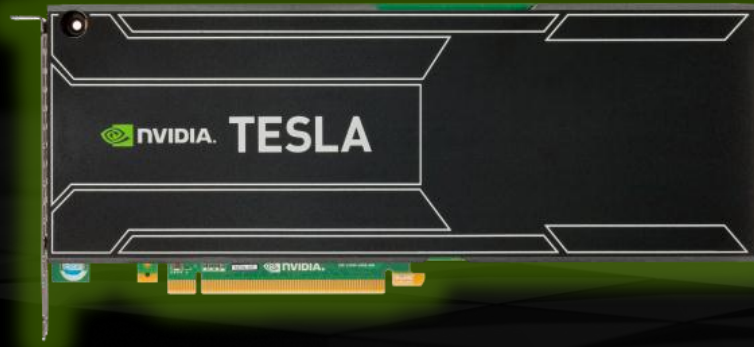


- 14 M Mixed cells
- Steady, $k-\epsilon$ turbulence
- Coupled PBNS, DP
- Total solution times
- CPU: AMG F-cycle
- GPU: FGMRES with AMG Preconditioner

NOTE: All results fully converged

ANSYS 15.0 New HPC Licenses for GPUs

Treats each GPU socket as a CPU core, which significantly increases simulation productivity of your HPC licenses



Needs 1 HPC task to enable a GPU

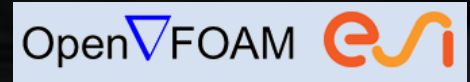
All ANSYS HPC products unlock GPUs in 15.0, including HPC, HPC Pack, HPC Workgroup, and HPC Enterprise products.

Computational Fluid Dynamics

OpenFOAM

NVIDIA GPU Strategy for OpenFOAM

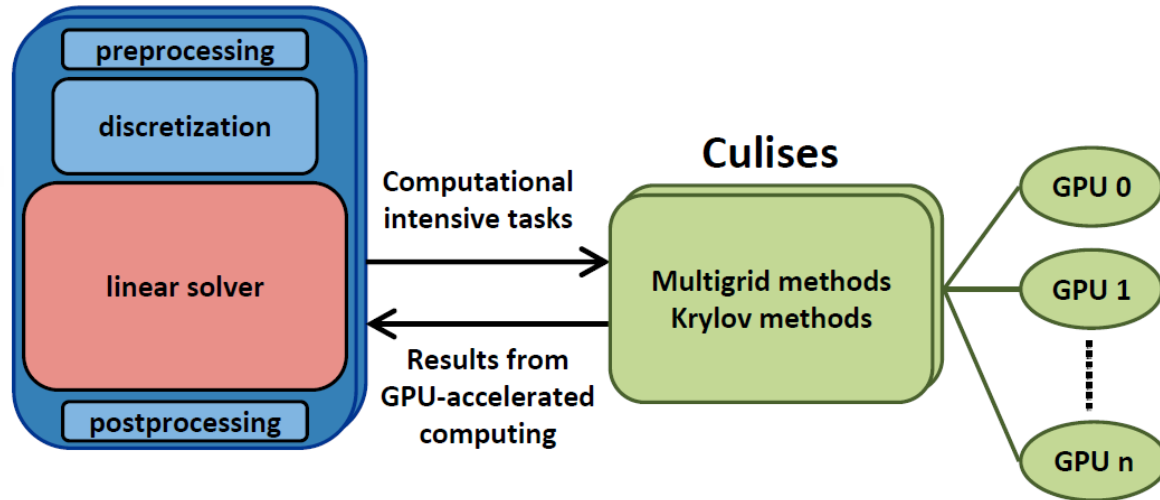
- Provide technical support for GPU solver developments
 - FluiDyna (implementation of NVIDIA's AMG), Vratiss and PARALUTION
 - AMG development by Russian Academy of Science ISP (A. Monakov)
 - Cufflink development by WUSTL now Engys North America (D. Combest)
- Invest in strategic alliances with OpenFOAM developers
 - ESI and OpenCFD Foundation (H. Weller, M. Salari)
 - Wikki and OpenFOAM-extend community (H. Jasak)
- Conduct performance studies and customer evaluations
 - Collaborations: developers, customers, OEMs (Dell, SGI, HP, etc.)



Library Culises

Concept and Features

Simulation tool e.g.
OpenFOAM[®]



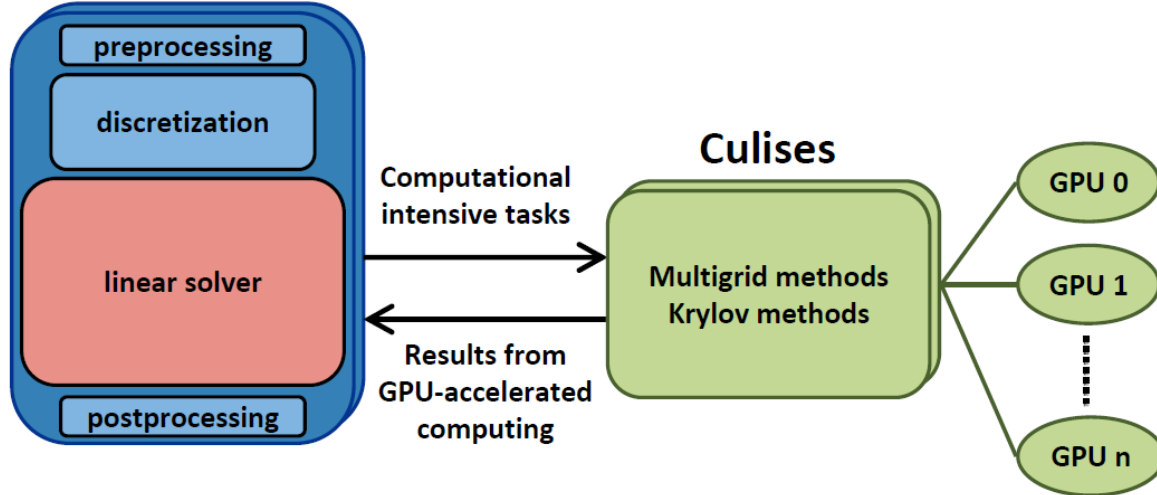
Culises = **C**uda **L**ibrary for **S**olving Linear **E**quation **S**ystems

See also www.culises.com

- State-of-the-art solvers for solution of linear systems
 - Multi-GPU and multi-node capable
 - Single precision or double precision available
- Krylov subspace methods
 - CG, BiCGStab, GMRES for symmetric /non-symmetric matrices
 - Preconditioning options
 - Jacobi (Diagonal)
 - Incomplete Cholesky (IC)
 - Incomplete LU (ILU)
 - Algebraic Multigrid (AMG), see below
- Stand-alone Multigrid method
 - Algebraic aggregation and classical coarsening
 - Multitude of smoothers (Jacobi, Gauss-Seidel, ILU etc.)
- Flexible interfaces for arbitrary applications e.g.: established coupling with OpenFOAM[®]

Summary hybrid approach

Simulation tool e.g.
OpenFOAM[®]



Advantage:

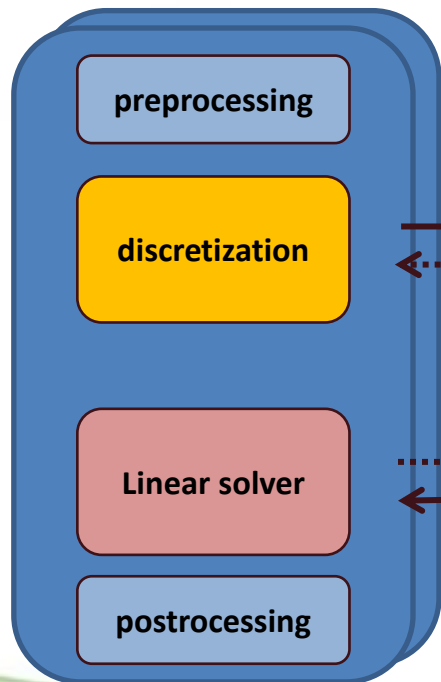
- **Universally applicable** (coupled to simulation tool of choice)
- Full availability of existing flow models
- Easy/no validation needed
- Unsteady approach better for hybrid due to large linear solver times

Disadvantages:

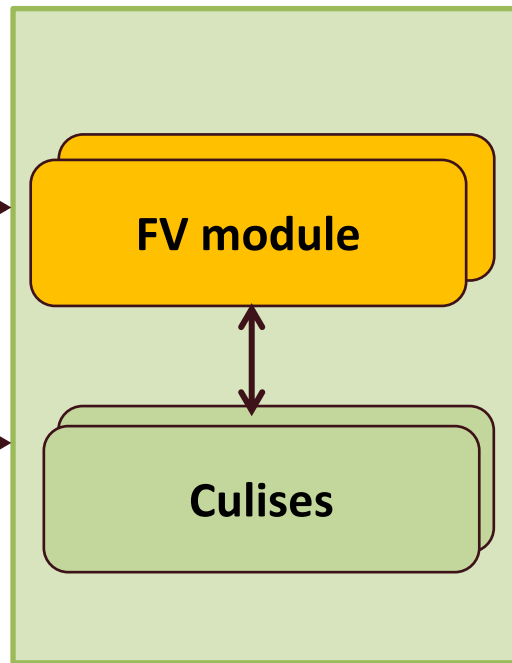
- Hybrid CPU-GPU produces overhead
- In case that solution of linear system not dominant
→ **Application speedup can be limited**

an extension of the hybrid approach

CPU flow solver
e.g. OpenFOAM[®]



aeroFluidX
GPU implementation

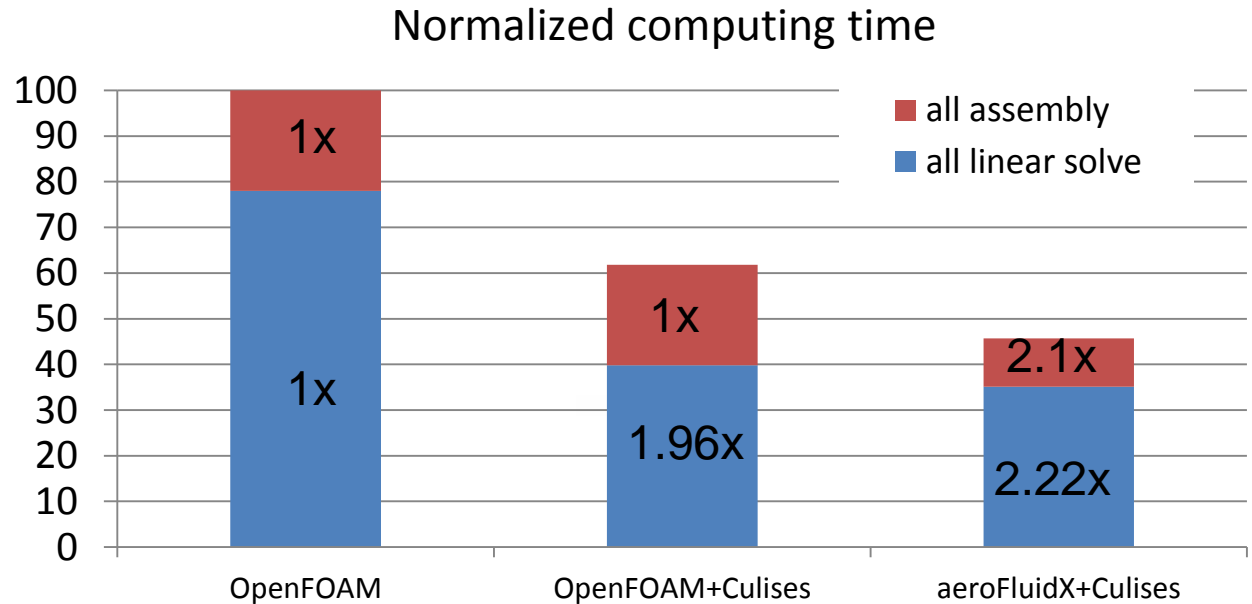


- Porting discretization of equations to GPU
 - ➔ discretization module (Finite Volume) running on GPU
 - ➔ Possibility of direct coupling to Culises
 - ➔ Zero overhead from CPU-GPU-CPU memory transfer and matrix format conversion
 - ➔ Solution of momentum equations also beneficial
- OpenFOAM[®] environment supported
 - ➔ Enables plug-in solution for OpenFOAM[®] customers
 - ➔ But communication with other input/output file formats possible

aeroFluidX

Cavity flow

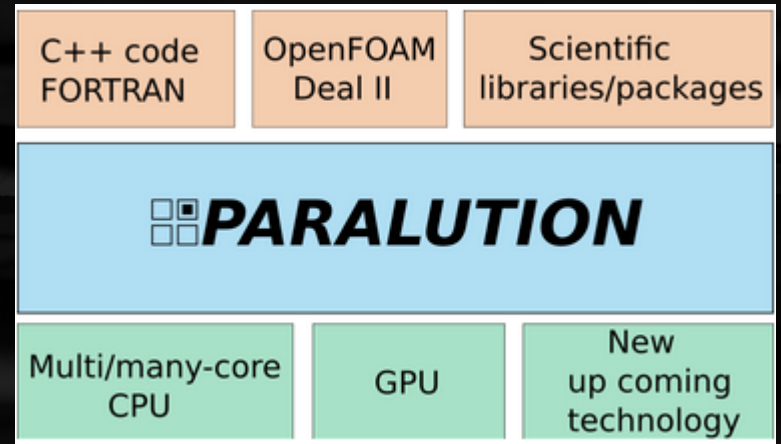
- CPU: Intel E5-2650 (all 8 cores)
GPU: Nvidia K40
- 4M grid cells (unstructured)
- Running 100 SIMPLE steps with:
 - OpenFOAM® (OF)
 - pressure: GAMG
 - Velocity: Gauss-Seidel
 - OpenFOAM® (OFC)
 - Pressure: **Culises AMGPCG (2.4x)**
 - Velocity: Gauss-Seidel
 - aeroFluidX (AFXC)
 - Pressure: **Culises AMGPCG**
 - Velocity: **Culises Jacobi**
- Total speedup:
 - OF (1x)
 - OFC 1.62x
 - AFXC 2.20x



all assembly = assembly of all linear systems (pressure and velocity)
all linear solve = solution of all linear systems (pressure and velocity)

PARALUTION

- C++ Library to perform various sparse iterative solvers and preconditioner
 - Contains Krylov subspace solvers (CR, CG, BiCGStab, GMRES, IDR), Multigrid (GMG, AMG), Deflated PCG,
- Multi/many-core CPU and GPU support
- Allows seamless integration with other scientific software
- PARALUTION Library is Open Source released under GPL v3



www.paralution.com

PARALUTION OpenFOAM plugin

CFD Problem

OpenFOAM

- ▶ Incompressible NS
- ▶ 6.8M 3D Cavity (190x190x190)
- ▶ icoFoam
- ▶ $\Delta t = 0.25e^{-5}$
- ▶ Pressure solver only

Hardware configuration

- ▶ Intel Haswell (i7 4770K)
- ▶ NVIDIA K40 (ECC)

CFD Problem

	GAMG vs PCG-AMG*		
	# iter	Time [s]	Speed-up
OpenFOAM	11.3	14.4	1.0 ×
OpenFOAM MPI	16.3	8.8	1.6 ×
PARALUTION CPU	7.0	8.7	1.6 ×
PARALUTION GPU	7.0	3.1	4.6 ×

*Preliminary results (still under development)

Total application speed-up 2×
(OpenFOAM MPI/GPU PARALUTION)

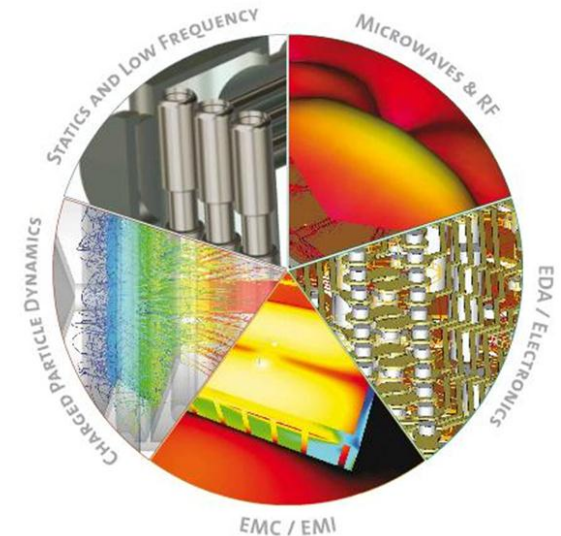
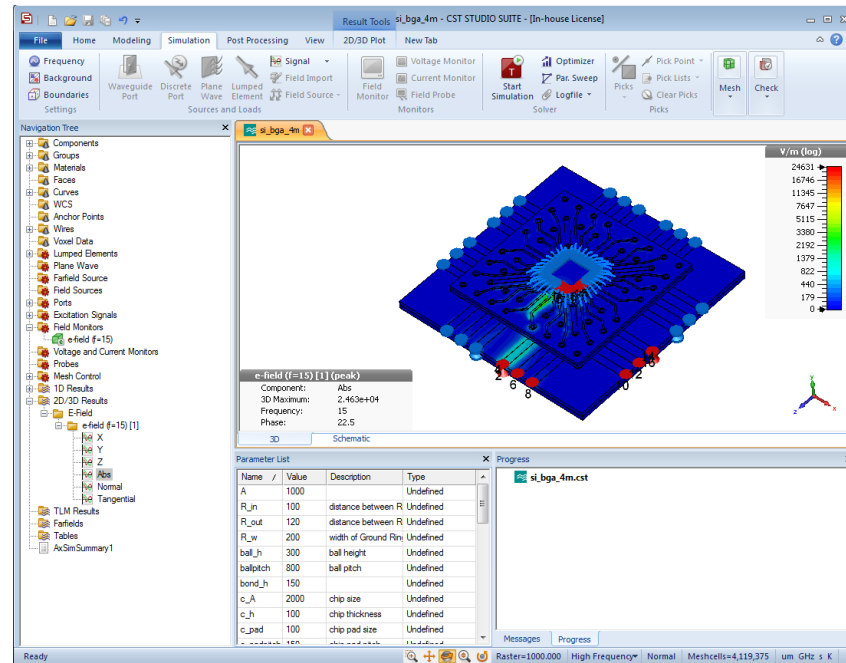
OpenFOAM Plugin will be released soon

Computational Electromagnetics

CST Studio Suite

CST - Company and Product Overview

- CST AG is one of the two largest suppliers of 3D EM simulation software.
- CST STUDIO SUITE is an integrated solution for 3D EM simulations it includes a parametric modeler, more than 20 solvers, and integrated post-processing. Currently, three solvers support GPU Computing.



New GPU Cards - Quadro K6000/Tesla K40

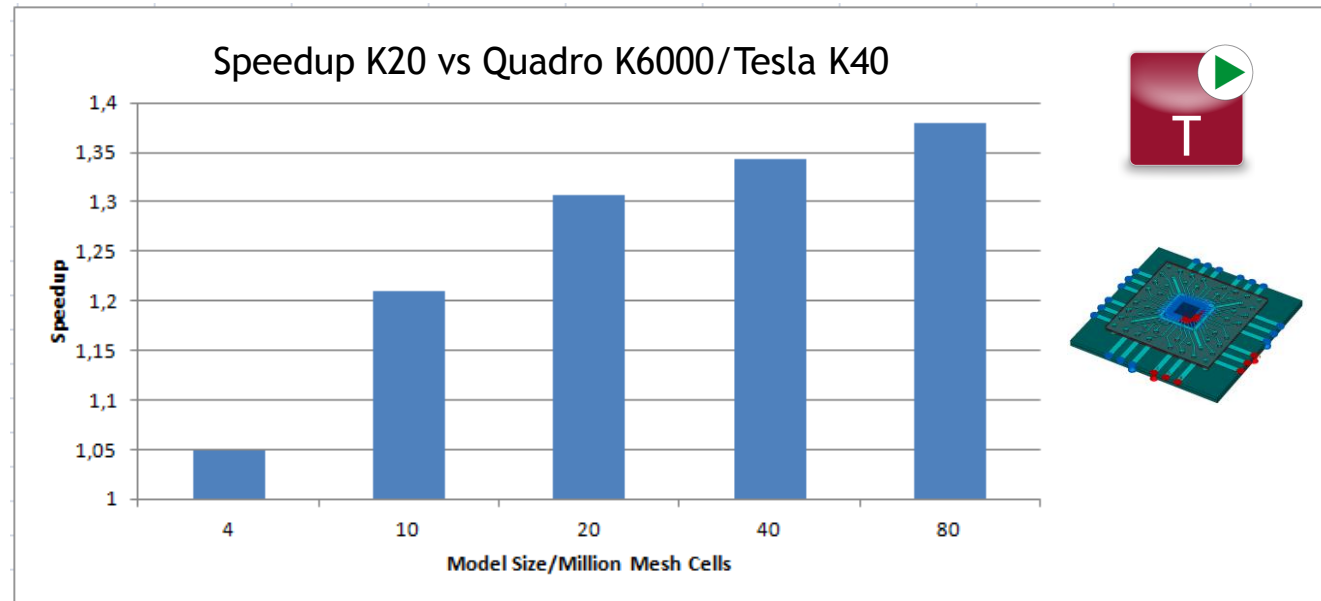
The Quadro K6000 is the new high-end graphics adapter of the Kepler series whereas the Tesla K40 card is the new high-end computing device. CST STUDIO SUITE 2013 supports both cards for GPU computing with service pack 5.



SPECIFICATIONS

GPU Memory	12 GB GDDR5
Memory Interface	384-bit
Memory Bandwidth	288 GB/s
CUDA Cores	2880
System Interface	PCI Express 3.0 x16
Max Power Consumption	225 W

www.nvidia.com

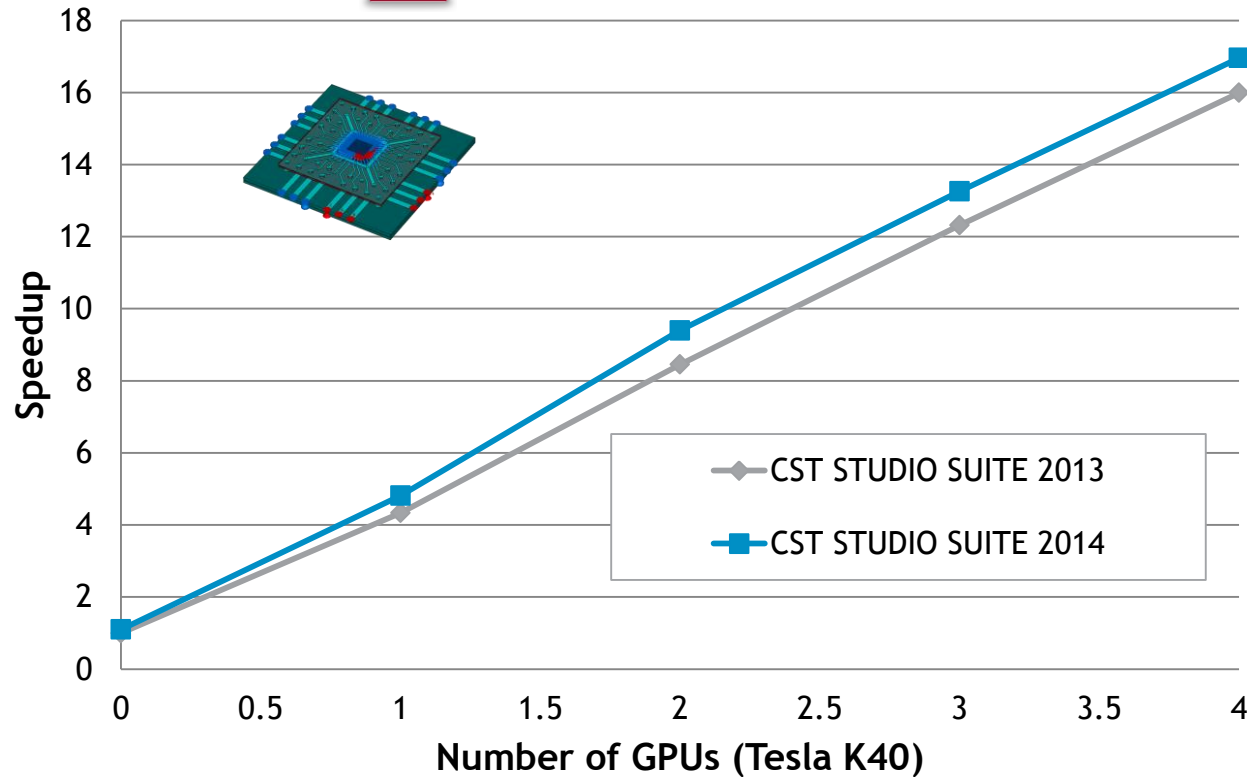


- The Quadro K6000/Tesla K40 card is about 30..35% faster than the K20 card.
- 12 GB onboard RAM allows for larger model size.

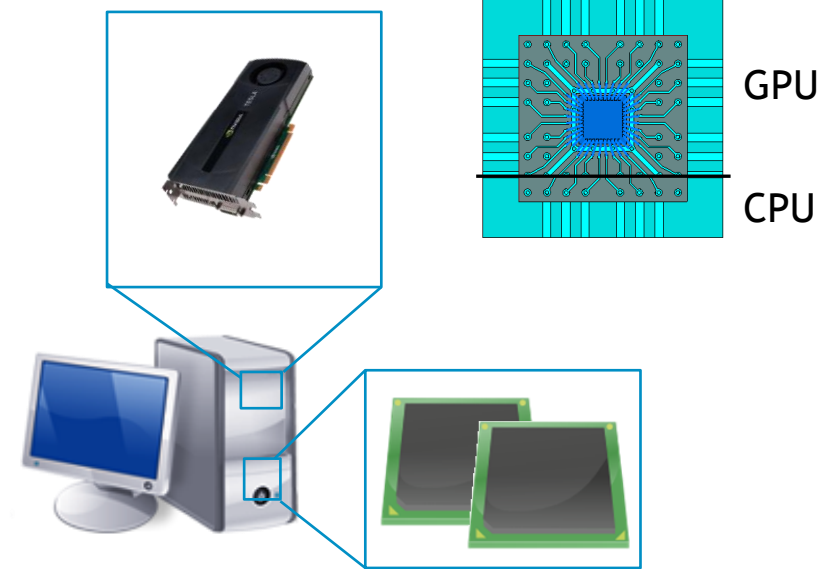
GPU Computing Performance



Speedup of Solver Loop



GPU computing **performance has been improved** for CST STUDIO SUITE 2014 as CPU and GPU resources are used in parallel.



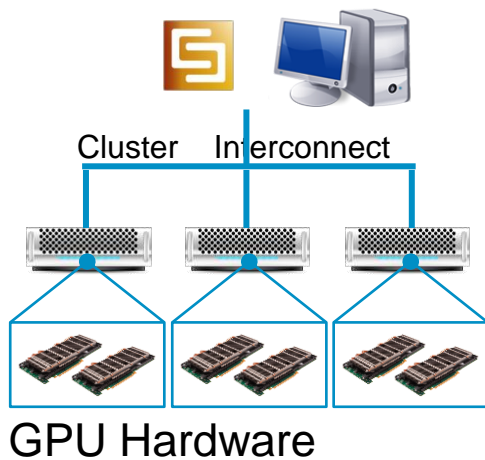
Benchmark performed on system equipped with dual Xeon E5-2630 v2 (Ivy Bridge EP) processors, and four Tesla K40 cards. Model has 80 million mesh cells.

MPI Computing – Performance

CST STUDIO SUITE® offers native support for high speed/low latency networks

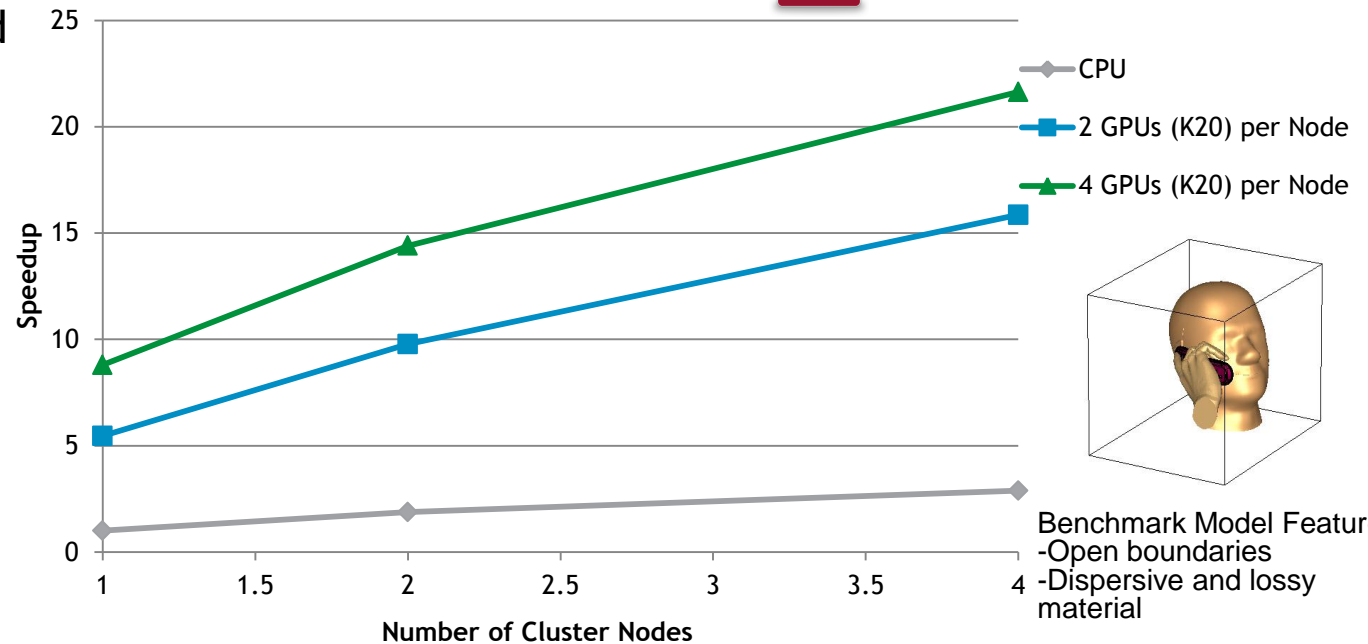
MPI Cluster System

CST STUDIO SUITE Frontend



Note: A GPU accelerated cluster system requires high-speed network in order to perform well!

Speedup of Solver Loop



Base model size is 80 million cells. Problem size is scaled up linearly with the number of cluster nodes (i.e., weak scaling). Hardware: dual Xeon E5-2650 processors, 128GB RAM per node (1600MHz), Infiniband QDR interconnect (40Gb/s).

Conclusion

- GPUs provide significant performance acceleration for solver intensive large jobs
 - Shorter product engineering cycles (Faster Time-to-Market) with improved product quality
 - Cut down energy consumption in the CAE process
 - Better Total Cost of Ownership (TCO)
- GPUs for 2nd level parallelism, preserves costly MPI investment
- GPU acceleration contributing to growth in emerging CAE
 - New ISV developments in particle based CFD (LBM, SPH, etc.)
 - Rapid growth for range of CEM applications and GPU adoption
- Simulations recently considered intractable are now possible
 - Large Eddy Simulation (LES) with a high degree of arithmetic intensity
 - Parameter optimization with highly increased number of jobs



The Visual Computing Company

Axel Koehler
akoehler@nvidia.com

NVIDIA, the NVIDIA logo, GeForce, Quadro, Tegra, Tesla, GeForce Experience, GRID, GTX, Kepler, ShadowPlay, GameStream, SHIELD, and The Way It's Meant To Be Played are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Other company and product names may be trademarks of the respective companies with which they are associated.

© 2014 NVIDIA Corporation. All rights reserved.