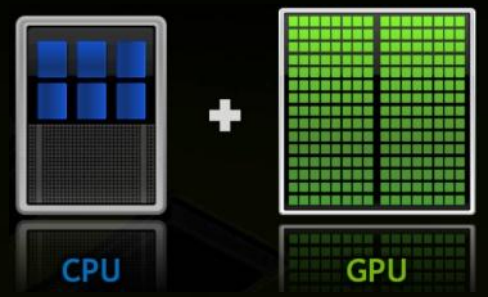
Performance Benefits of NVIDIA GPUs for Abaqus

Stan Posey and Srinivas Kodiyalam
CAE Industry and ISV Development
NVIDIA, Santa Clara, CA, USA
sposey@nvidia.com; skodiyalam@nvidia.com

NVIDIA and HPC Evolution of GPUs

- Public, based in Santa Clara, CA | ~$4B revenue | ~5,500 employees
- Founded in 1999 with primary business in semiconductor industry
  - Products for graphics in workstations, notebooks, mobile devices, etc.
  - Began R&D of GPUs for HPC in 2004, released first Tesla and CUDA in 2007
- Development of GPUs as a co-processing accelerator for x86 CPUs

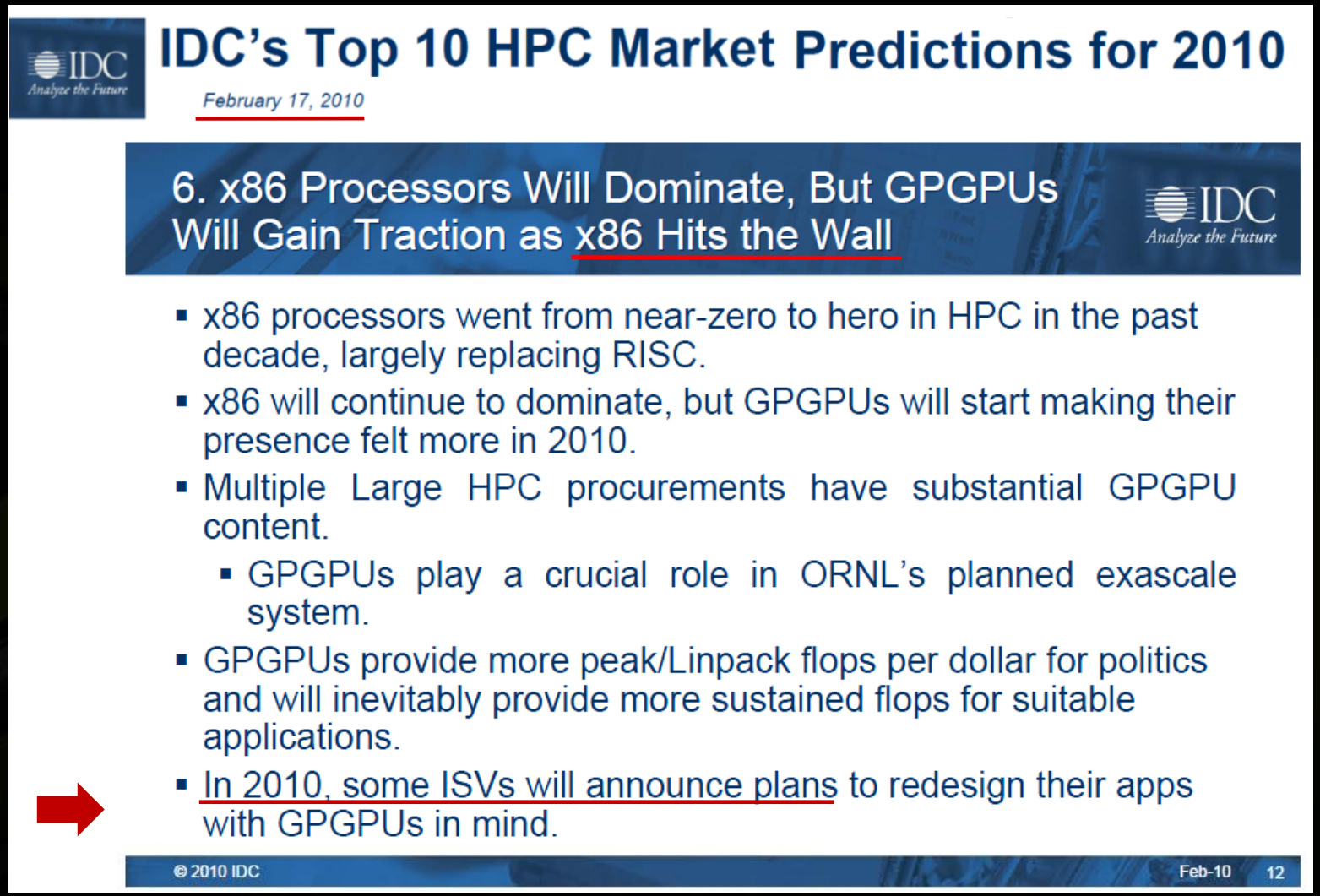


HPC Evolution of GPUs

- 2004: Began strategic investments in GPU as HPC co-processor
- 2006: G80 first GPU with built-in compute features, 128 cores; CUDA SDK Beta
- 2007: Tesla 8-series based on G80, 128 cores – CUDA 1.0, 1.1
- 2008: Tesla 10-series based on GT 200, 240 cores – CUDA 2.0, 2.3
- 2009: Tesla 20-series, code named "Fermi" up to 512 cores – CUDA SDK 3.0, 3.2

3 Years With
3 Generations

Motivation for CPU Acceleration with GPUs



IDC Analyze the Future

IDC's Top 10 HPC Market Predictions for 2010

February 17, 2010

6. x86 Processors Will Dominate, But GPGPUs Will Gain Traction as x86 Hits the Wall

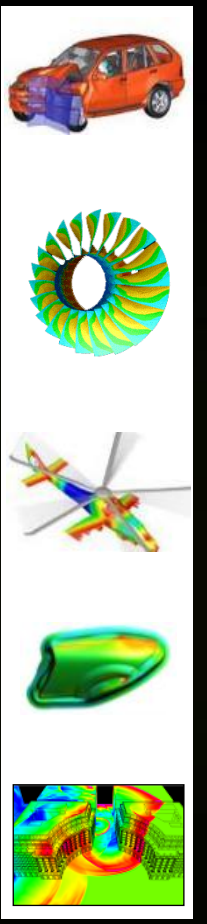
IDC Analyze the Future

- x86 processors went from near-zero to hero in HPC in the past decade, largely replacing RISC.
- x86 will continue to dominate, but GPGPUs will start making their presence felt more in 2010.
- Multiple Large HPC procurements have substantial GPGPU content.
  - GPGPUs play a crucial role in ORNL's planned exascale system.
- GPGPUs provide more peak/Linpack flops per dollar for politics and will inevitably provide more sustained flops for suitable applications.
- In 2010, some ISVs will announce plans to redesign their apps with GPGPUs in mind.

© 2010 IDC

Overview of CAE Software Progress for GPUs

- GPUs are an Emerging HPC Technology for ISVs
  - Industry Leading ISV Software is GPU-Enabled Today
- Initial GPU Performance Gains are Encouraging
  - Just the beginning of more performance and more applications
- NVIDIA Investments in ISV Developments
  - Joint technical collaborations at all major CAE ISVs



Value of GPU Acceleration for Abaqus

- More simulations increases the number of design candidates
  - More design candidates means improved quality in designs
- Faster simulations shorten engineering cycles
  - Shorter engineering cycles means faster time-to-market
- Simulations once considered intractable now possible
  - Next generation of simulation challenges become practical

Large Abaqus simulations completed in less time, permits more studies of design candidates and provides better predictions

SIMULIA and NVIDIA Collaboration Focus

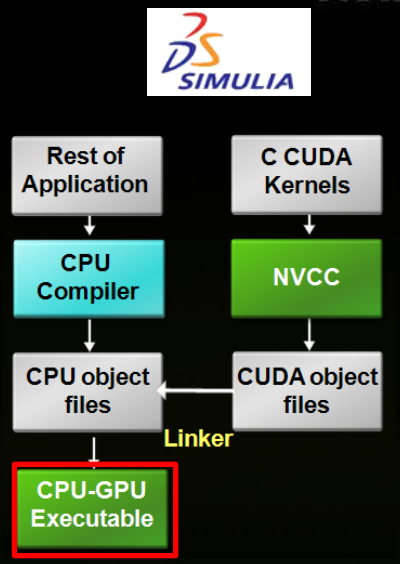
- **Abaqus/Standard:** GPU acceleration of the direct solver for Linux and Windows – **Full support announced for 6.11 release in May 2011**
- **Abaqus/Explicit:** GPU R&D evaluation for future release
- **Abaqus/CFD:** GPU evaluation of implicit iterative solvers
- **Abaqus/CAE:** Development of graphics acceleration for pre- and post-processing with OpenGL and use of Quadro GPUs

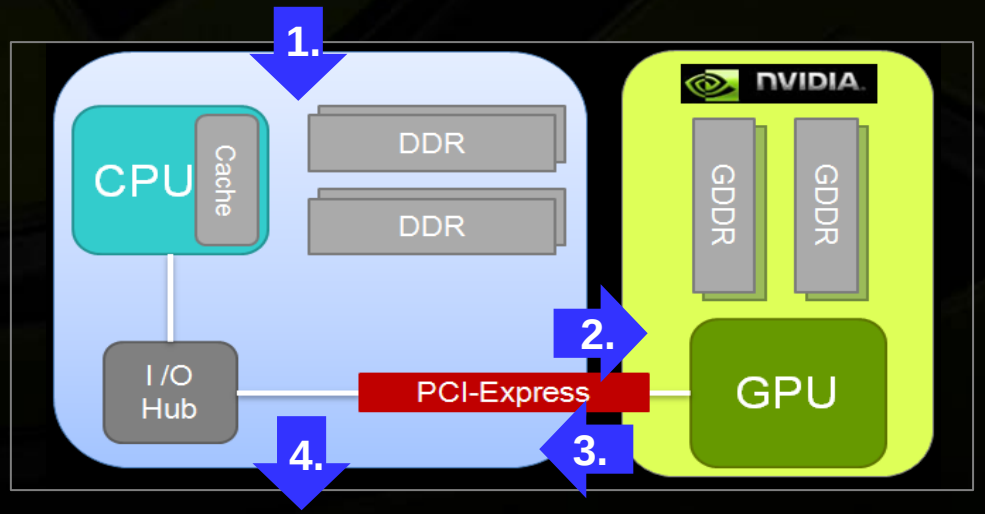
Details of Abaqus/Standard CUDA Release

- **SIMULIA announced support for CUDA 3.2 and GPUs in 6.11**
  - NVIDIA GPUs include Tesla C2050, C2070 and Quadro 6000, 5000
- **Only the direct equation solver is accelerated on the GPU**
  - More of Abaqus will be moved to GPUs in progressive stages
- **Initial release single GPU only (with model size limits) and for both Linux and Windows – multi-GPU release in the future**
  - Model size limits for single GPU will depend on model, estimates of 8M DOF limit for 6GB GPU memory are current guidance for 6.11
- **Basic system recommendations**
  - Large system memory (~32GB) to avoid scratch I/O of system matrix
  - Configuration of single GPU attached to single CPU socket (4-6 cores)

Basics of GPU Computing with Abaqus/Standard

- GPUs are an attached accelerator to an x86 CPU
  - GPUs cannot operate without an x86 CPU present
- Abaqus/Standard GPU acceleration is user-transparent
  - Jobs launch and complete without additional user steps
- Schematic of a CPU with an attached GPU accelerator:
  - CPU begins/ends job, GPU manages heavy computations





Schematic of an x86 CPU with a GPU accelerator

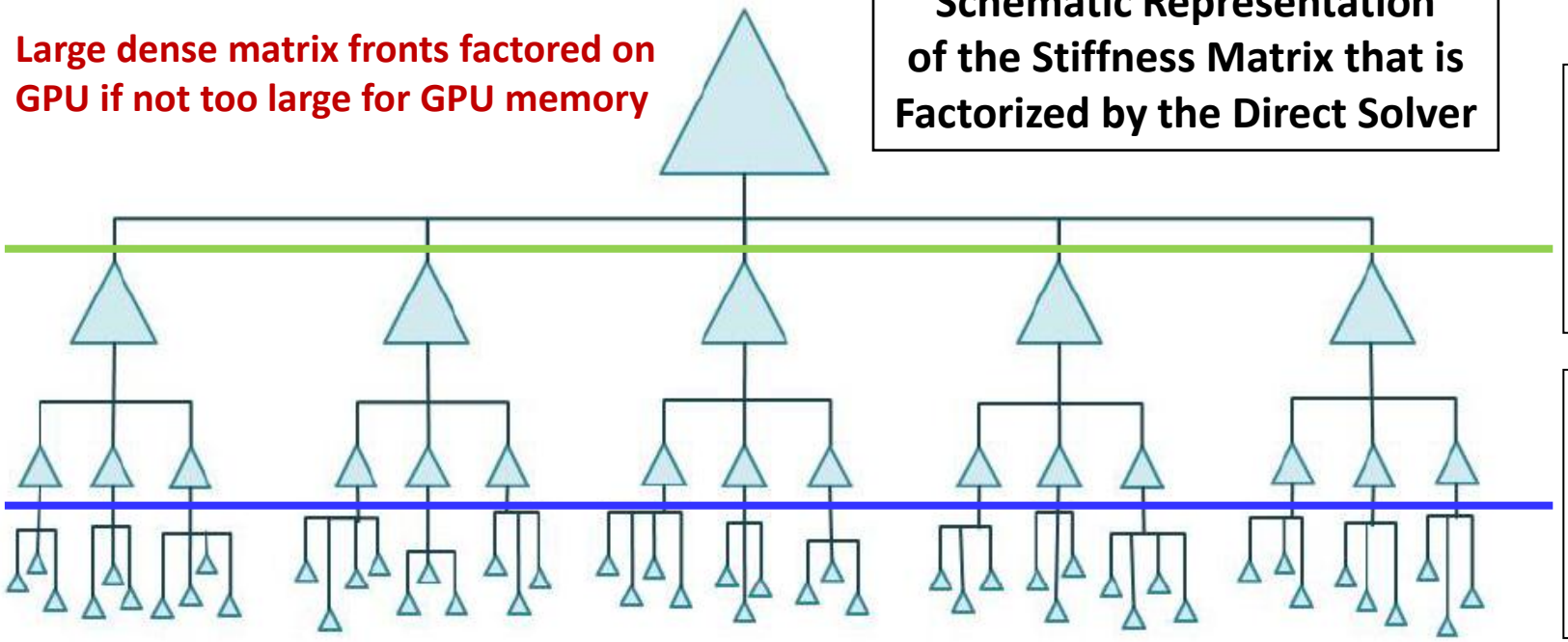
1. Abaqus job launched on CPU
2. Solver operations sent to GPU
3. GPU sends results back to CPU
4. Abaqus job completes on CPU

Basics of Abaqus/Standard Implementation

Abaqus/Standard – deployment of multi-frontal direct sparse solver

Large dense matrix fronts factored on GPU if not too large for GPU memory

Schematic Representation of the Stiffness Matrix that is Factorized by the Direct Solver



Upper threshold: Fronts too large for single GPU memory need multiple GPUs

Lower threshold: Fronts too small to overcome PCIe data transfer costs stay on CPU cores

Small dense matrix fronts factored in parallel on CPU – more cores means higher performance

Abaqus Performance Studies for Model S4b

SIMULIA: Dell T5500 Workstation Configuration

- 2 x Xeon X5530 QC 2. 4 GHz CPUs (Nehalem)
- 48 GB memory
- NVIDIA Tesla C2050 with 3 GB memory
- RHEL5.4, MKL 10.25, NVIDIA CUDA 3.1 – 256.40
- *Study conducted at SIMULIA HQ in Providence, RI*

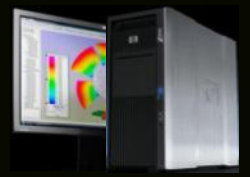
Dell T5500 Workstation

NVIDIA Tesla C2070 GPU

NVIDIA: HP Z800 Workstation Configuration

- 2 x Xeon X5570 QC 2.93 GHz CPUs (Nehalem)
- 48 GB memory – 12 x 4GB 1333 MHz DIMMs
- NVIDIA Tesla C2070 (2) with 6 GB memory each
- RHEL5.4, NVIDIA CUDA 3.2
- *Study conducted at NVIDIA by Performance Lab*

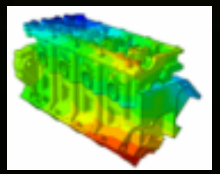
HP Z800 Workstation



NVIDIA Tesla C2070 GPU

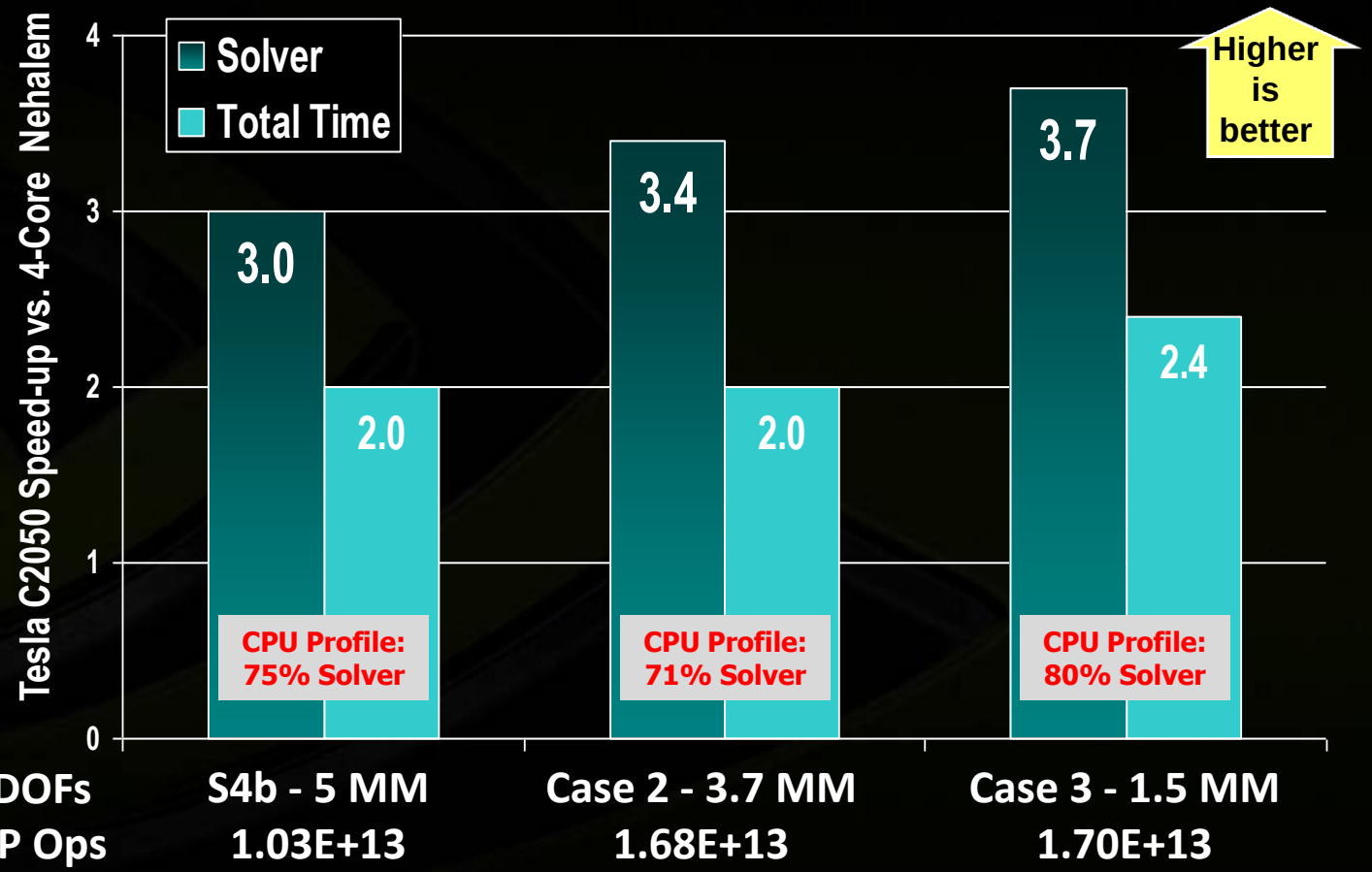
Abaqus/Standard Model – S4b

- Engine geometry, 5,000 K DOF and SOLID187 FE's
- Single load step, static, large deflection nonlinear
- Abaqus/Standard 6.10-EF direct sparse solver



SIMULIA Study on Dell Nehalem-GPU Workstation

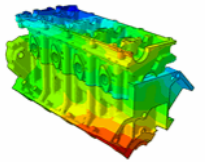
Abaqus/Standard: Based on v6.10-EF Direct Solver – Tesla C2050, CUDA 3.1 vs. 4-core Nehalem



Source: SIMULIA Customer Conference, 27 May 2010:

"Current and Future Trends of High Performance Computing with Abaqus"

Presentation by Matt Dunbar



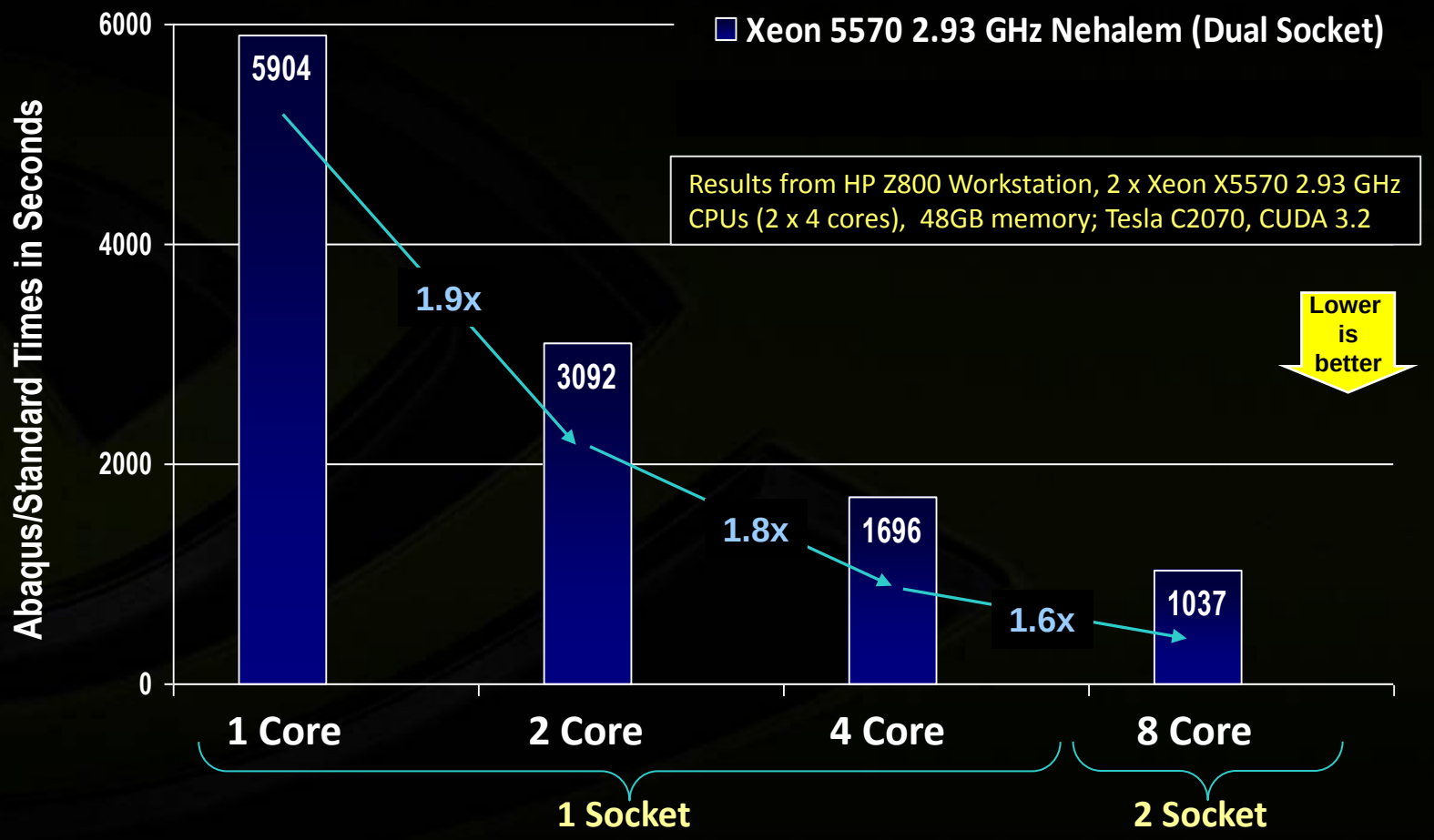
S4b: Engine Block Model of 5 MM DOF

NOTE: Solver Performance Increases with FP Operations

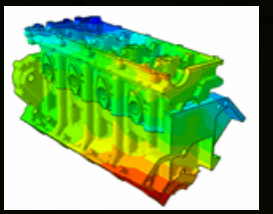
Results Based on 4-core CPU

NVIDIA Study for HP Nehalem-GPU Workstation

NOTE: Results Based on Abaqus/Standard 6.10-EF Direct Solver



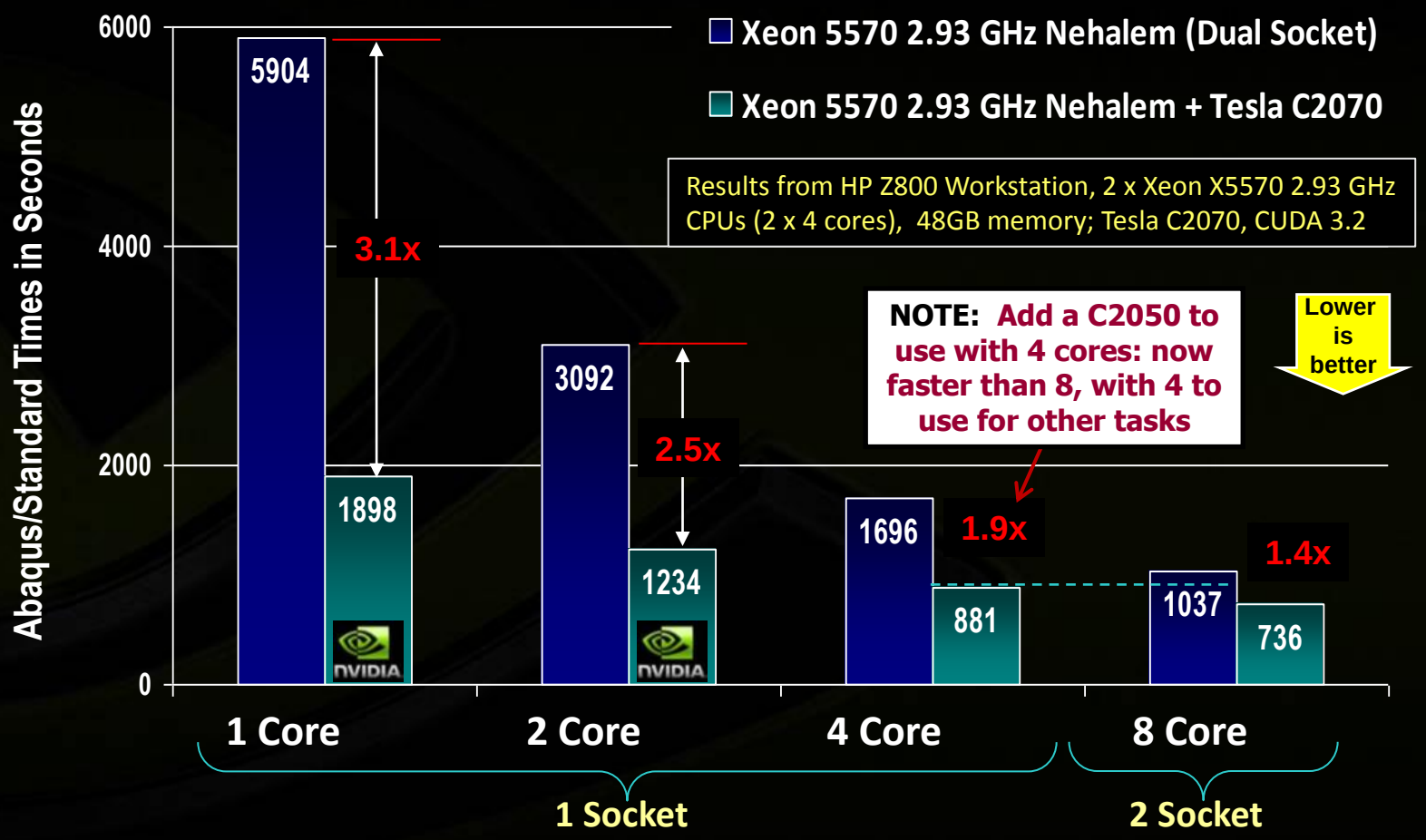
S4b Model



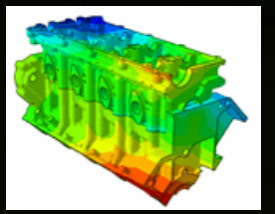
- Engine geometry
- 5,000 K DOF
- Solid FEs
- Static, nonlinear
- Five iterations
- Direct sparse

NVIDIA Study for HP Nehalem-GPU Workstation

NOTE: Results Based on Abaqus/Standard 6.10-EF Direct Solver



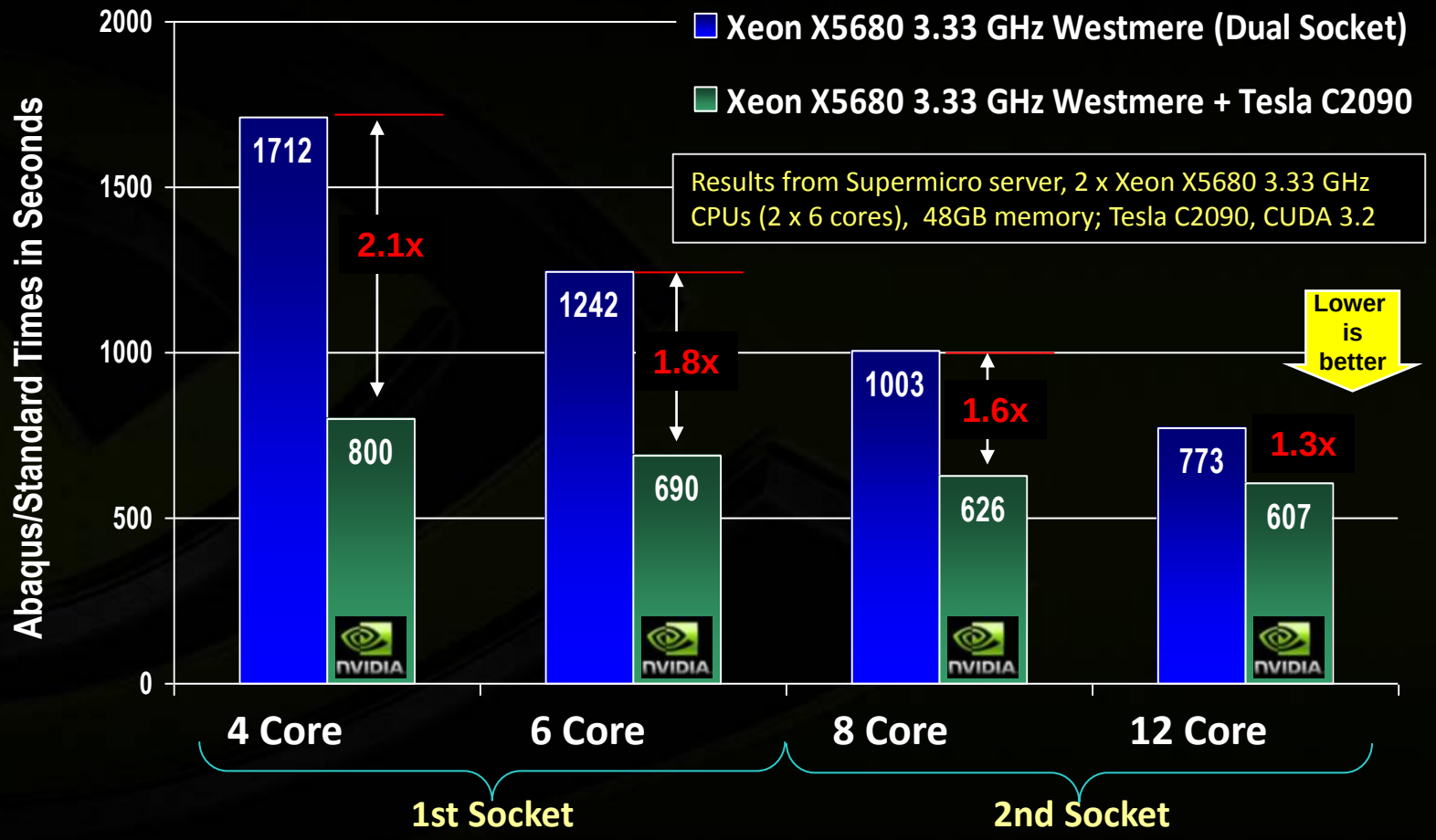
S4b Model



- Engine geometry
- 5,000 K DOF
- Solid FEs
- Static, nonlinear
- Five iterations
- Direct sparse

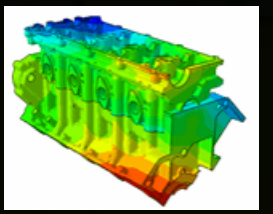
NVIDIA Study for Westmere-GPU Server and C2090

NOTE: Results Based on Abaqus/Standard 6.11 Direct Solver



New!

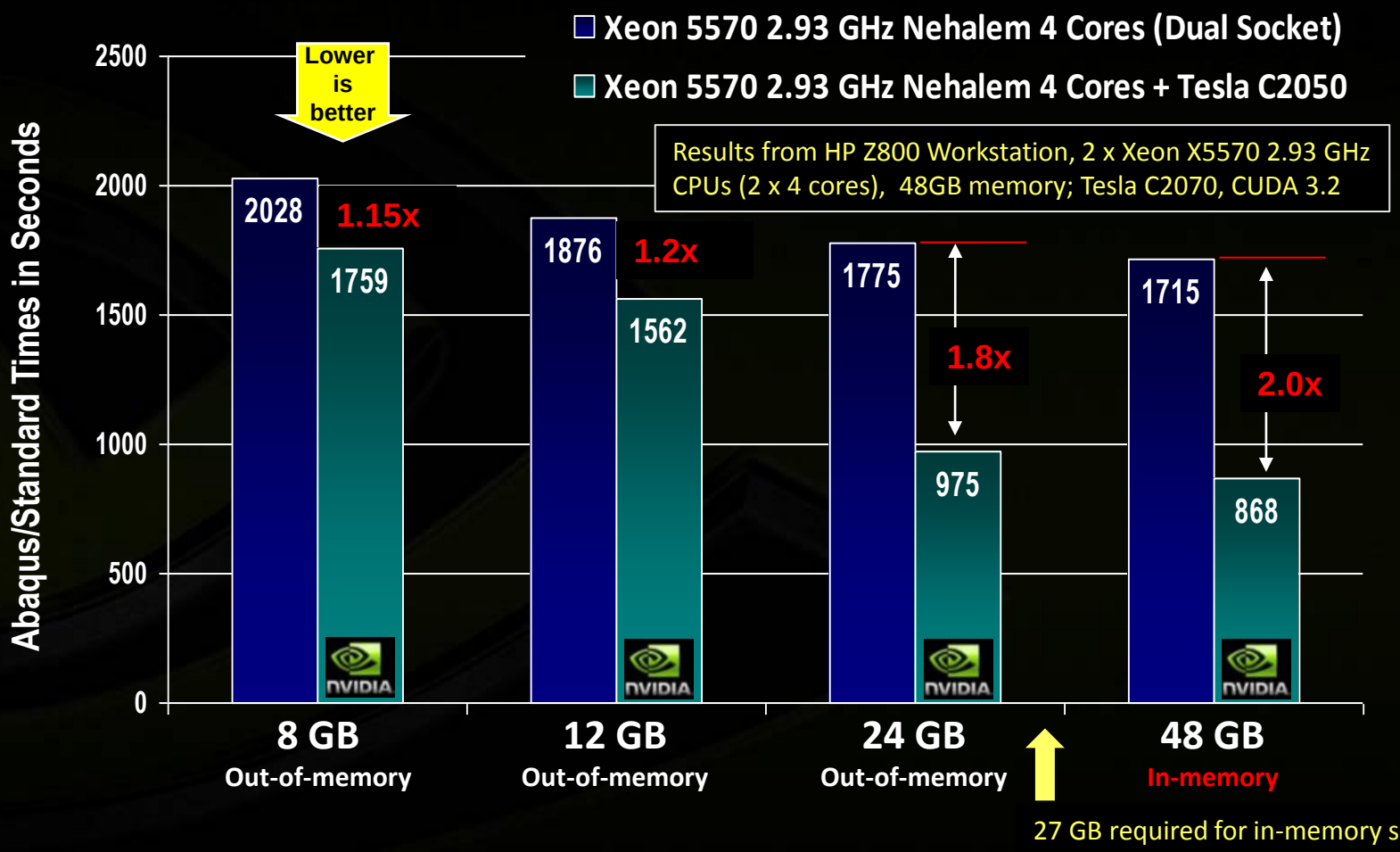
S4b Model



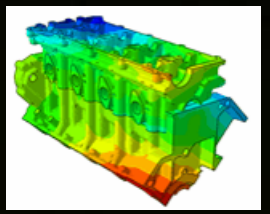
- Engine geometry
- 5,000 K DOF
- Solid FEs
- Static, nonlinear
- Five iterations
- Direct sparse

Effects of System CPU Memory for S4b Model

NOTE: Results Based on Abaqus/Standard 6.10-EF Direct Solver



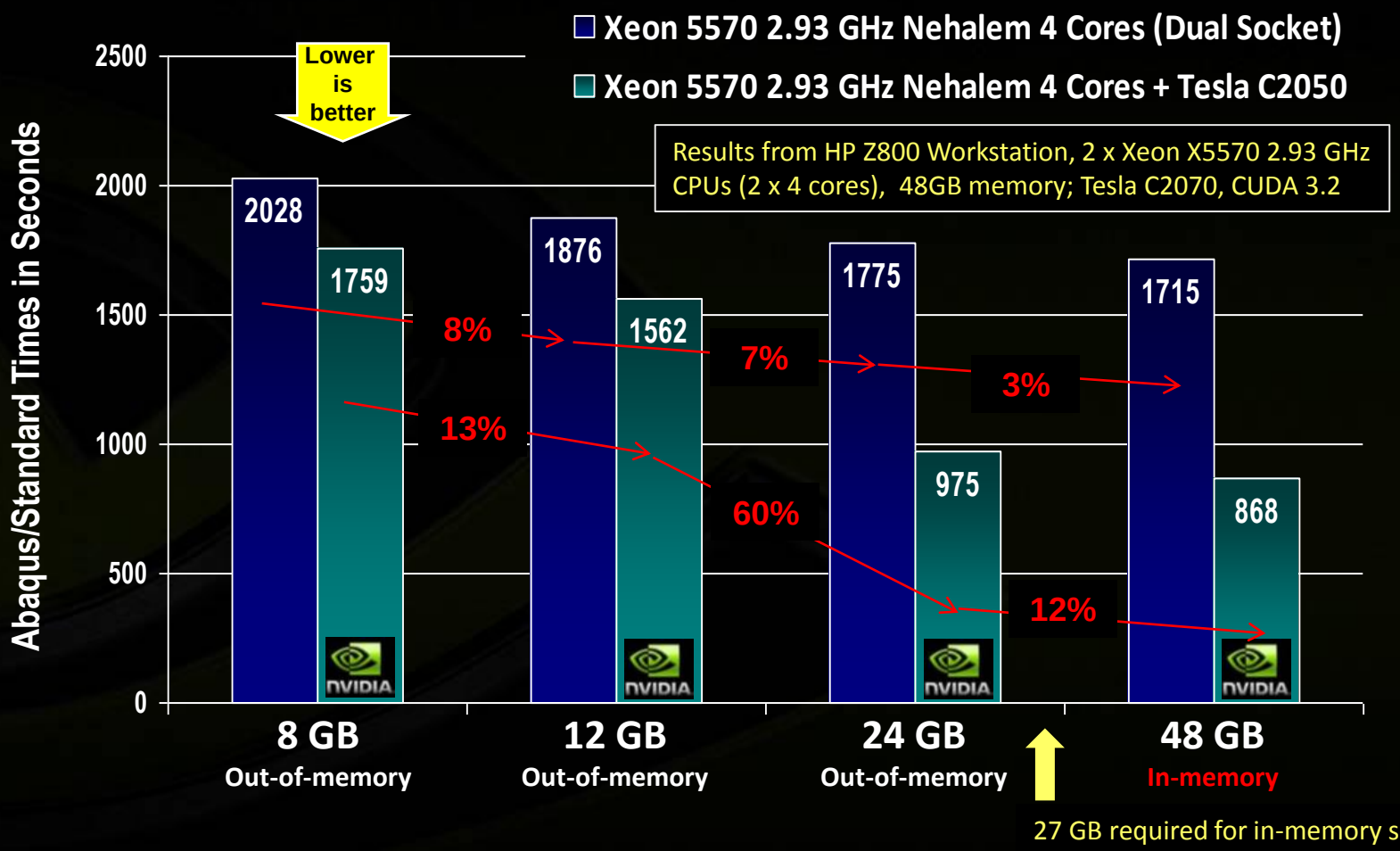
S4b Model



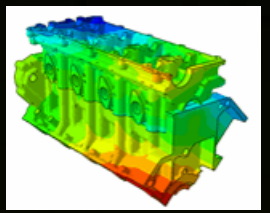
- Engine geometry
- 5,000 K DOF
- Solid FEs
- Static, nonlinear
- Five iterations
- Direct sparse

Effects of System CPU Memory for S4b Model

NOTE: Results Based on Abaqus/Standard 6.10-EF Direct Solver



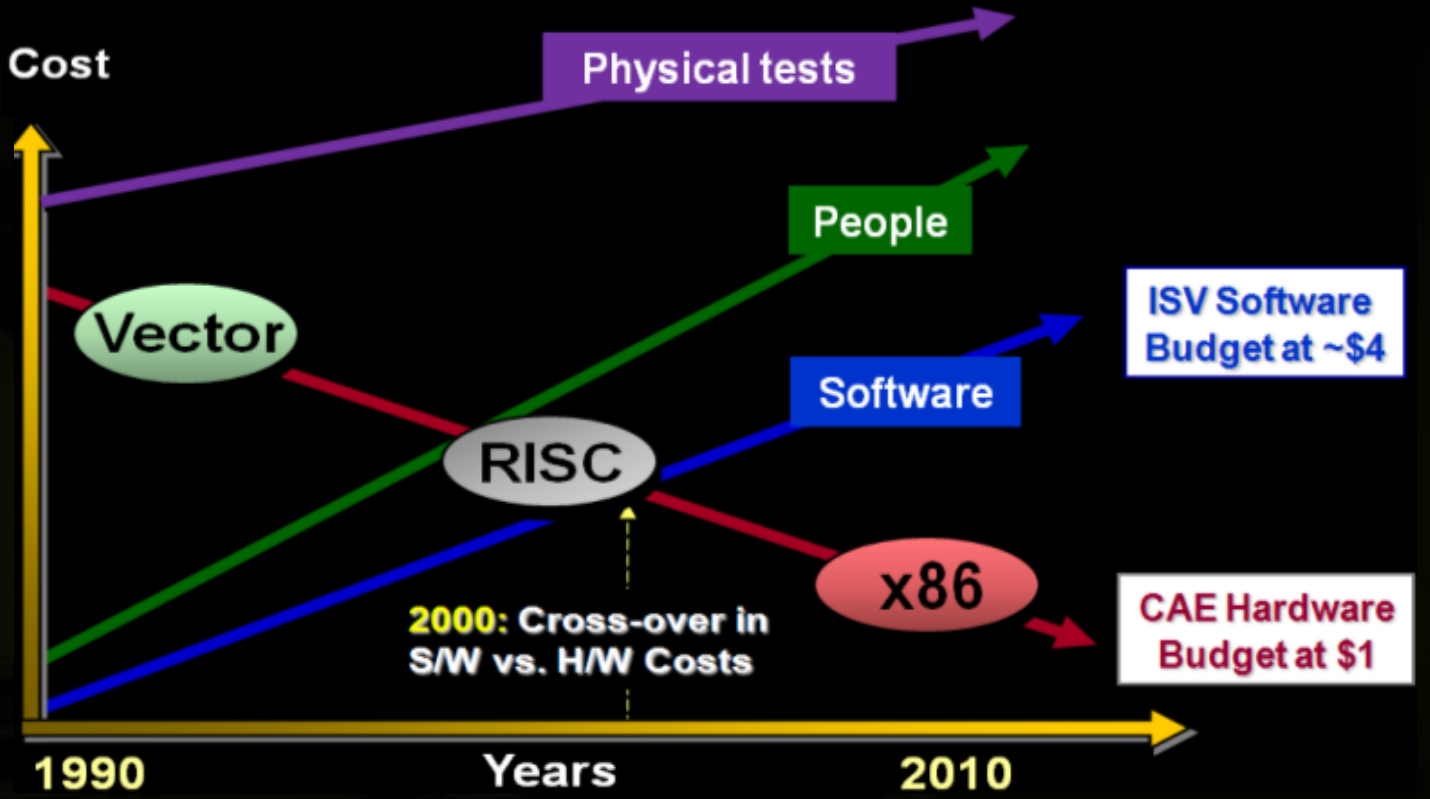
S4b Model



- Engine geometry
- 5,000 K DOF
- Solid FEs
- Static, nonlinear
- Five iterations
- Direct sparse

Economics of CAE Software in Practice

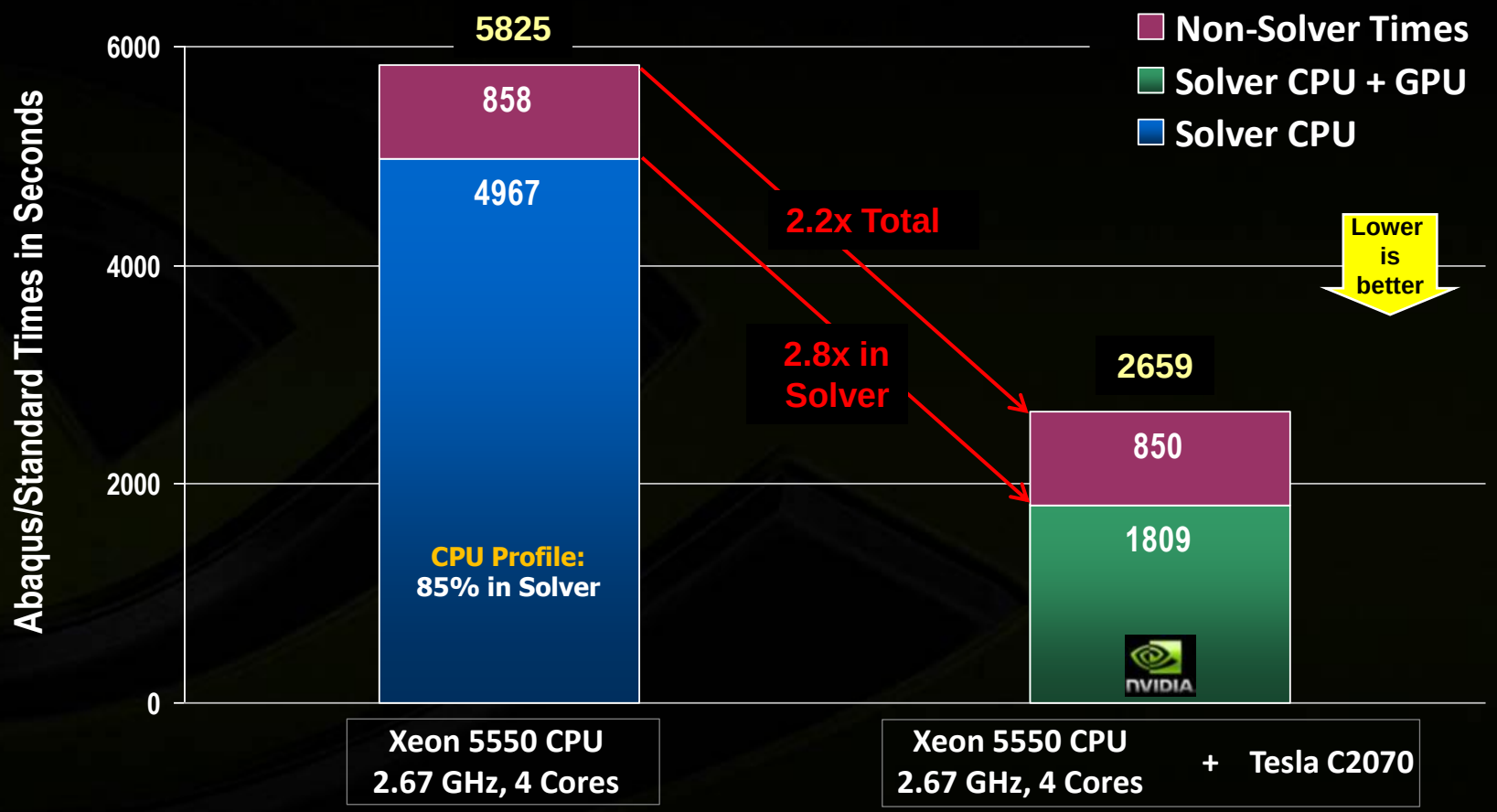
Cost Trends in CAE Deployment: **Costs in People and Software Continue to Increase**



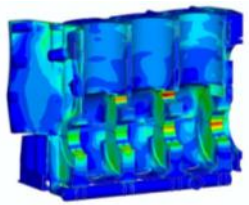
- Historically hardware very expensive vs. ISV software and people
- Software budgets are now 4x vs. hardware
- Increasingly important that hardware choices drive cost efficiency in people and software

Abaqus and Automotive Customer Case Study

NOTE: Results Based on Abaqus/Standard 6.10-EF Direct Solver



Engine Model

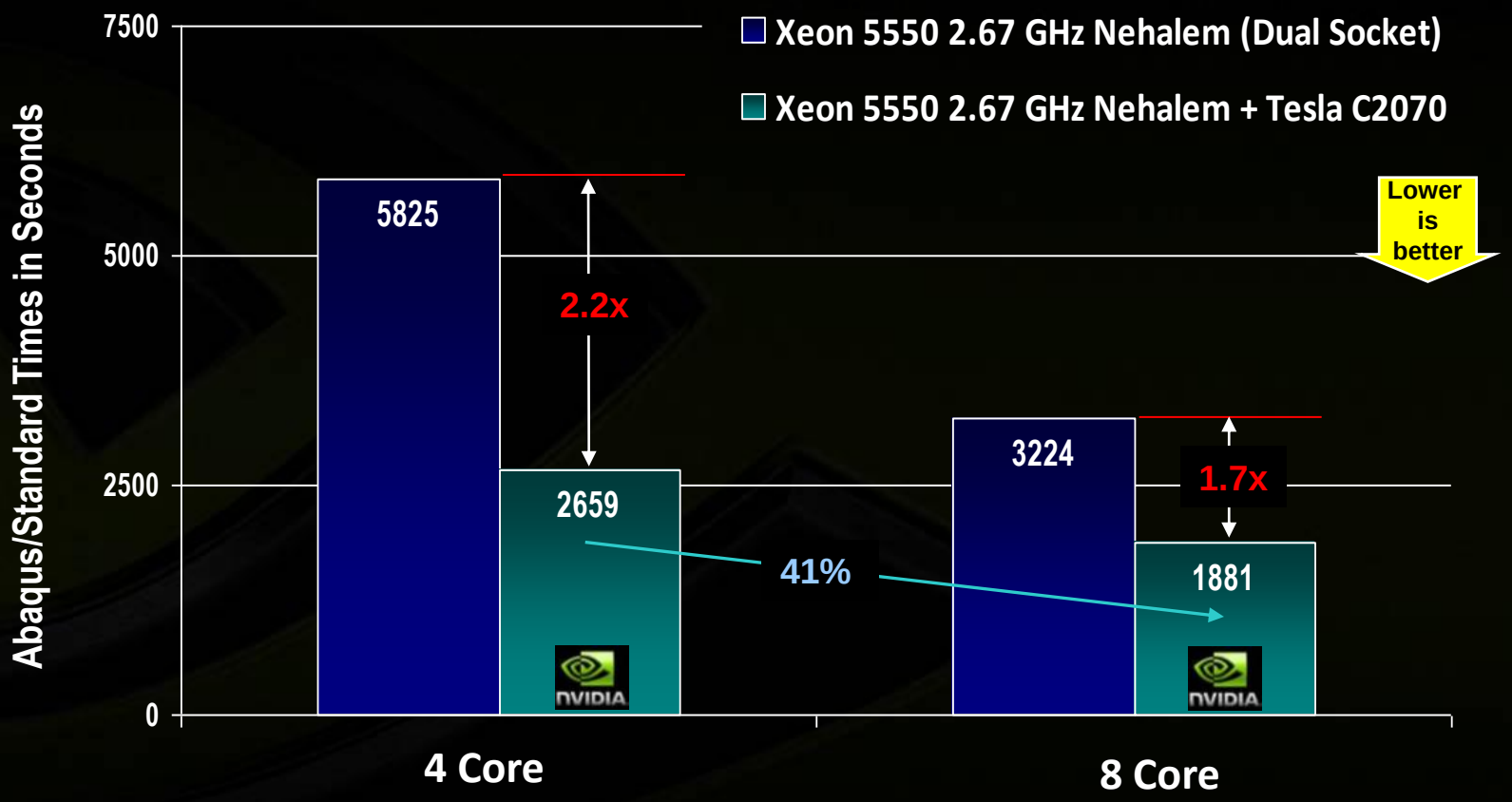


- 1.5M DOF
- 2 Load steps
- 28 Iterations
- 5.8e12 Ops per Iteration

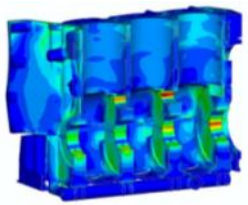
Results from HP Z800 Workstation, 2 x Xeon X5550 2.67 GHz CPUs, 48GB memory, MKL 10.25; Tesla C2070 with CUDA 3.1

Abaqus and Automotive Customer Case Study

NOTE: Results Based on Abaqus/Standard 6.10-EF Direct Solver



Engine Model



- 1.5M DOF
- 2 Load steps
- 28 Iterations
- 5.8e12 Ops per Iteration

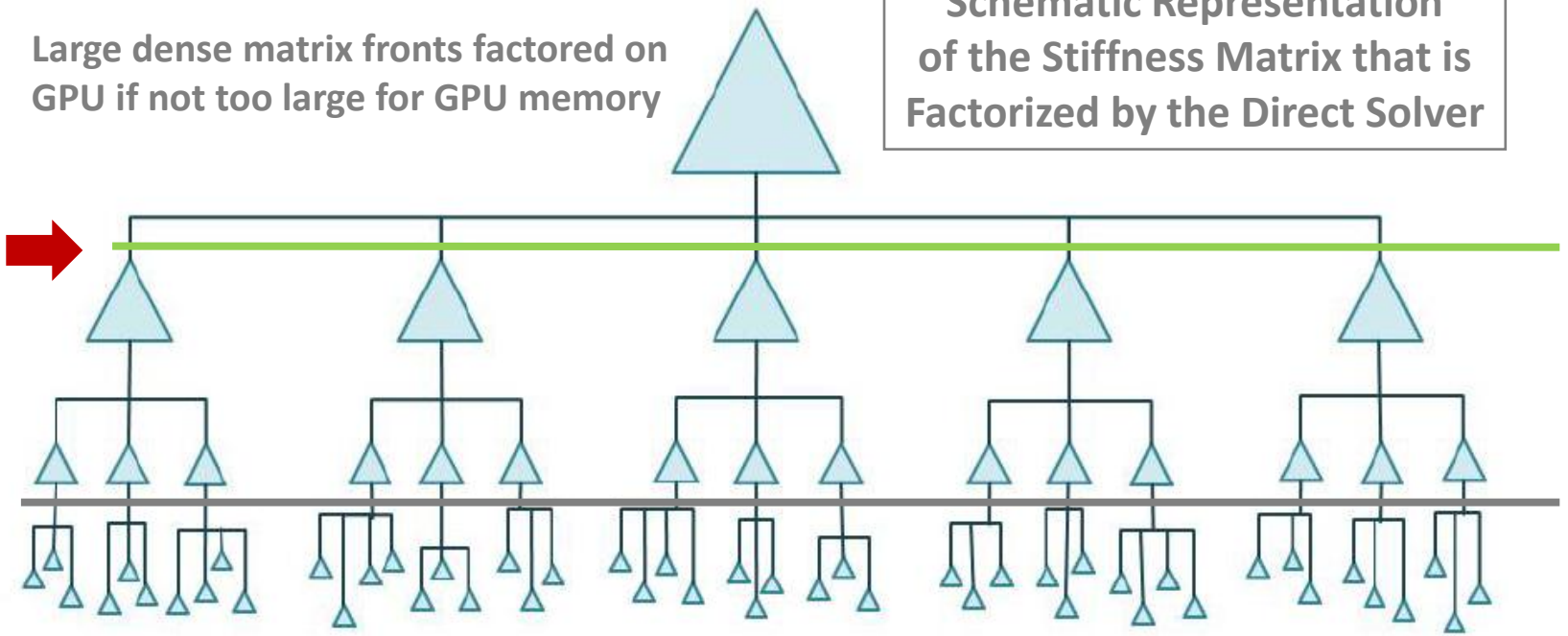
Results from HP Z800 Workstation, 2 x Xeon X5550 2.67 GHz CPUs, 48GB memory, MKL 10.25; Tesla C2070 with CUDA 3.1

How to Remove *Upper Threshold* for Larger Models?

Abaqus/Standard – deployment of multi-frontal direct sparse solver

Large dense matrix fronts factored on GPU if not too large for GPU memory

Schematic Representation of the Stiffness Matrix that is Factorized by the Direct Solver



Upper threshold:
Fronts too large for single GPU memory need multiple GPUs

Lower threshold:
Fronts too small to overcome PCIe data transfer costs stay on CPU cores

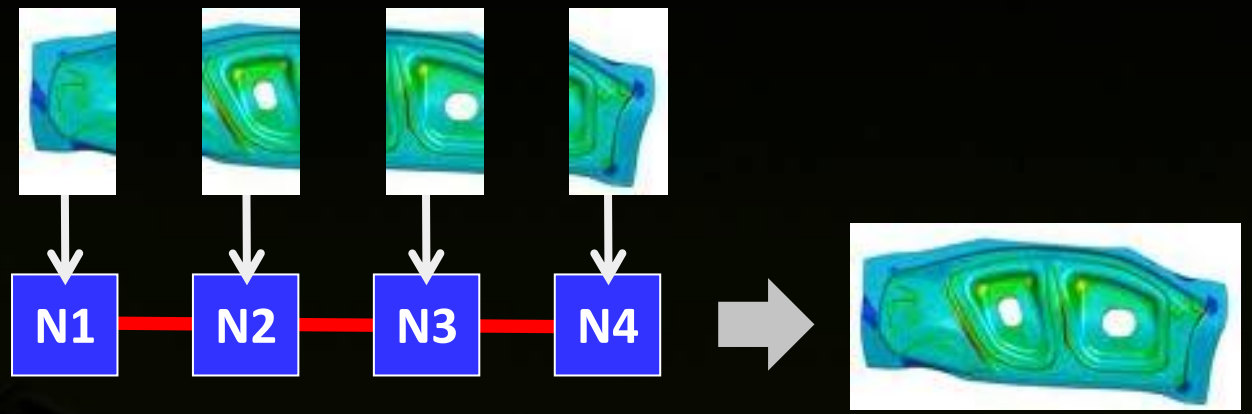
Small dense matrix fronts factored in parallel on CPU – more cores means higher performance

Distributed FEA and NVIDIA GPU Clusters

NOTE: Illustration Based on a Simple Example of 4 Partitions and 4 Compute Nodes

Model geometry is decomposed; partitions are sent to independent compute nodes on a cluster

Compute nodes operate distributed parallel using MPI communication to complete a solution per time step



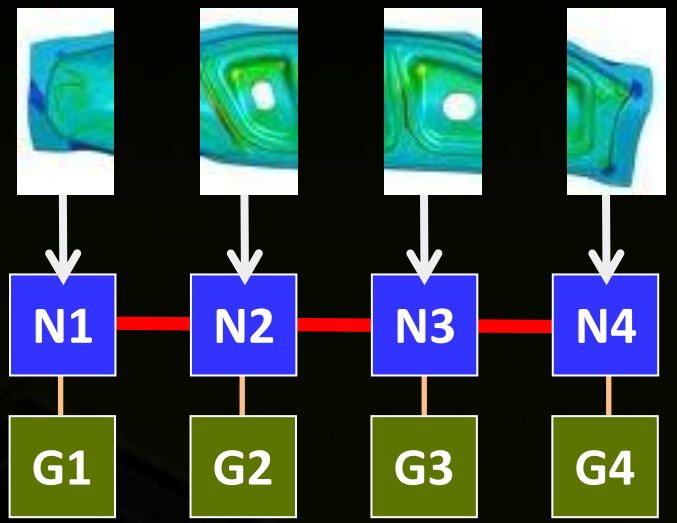
A global solution is developed for the completed job

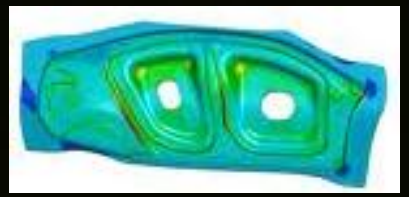
Distributed FEA and NVIDIA GPU Clusters

NOTE: Illustration Based on a Simple Example of 4 Partitions and 4 Compute Nodes

Model geometry is decomposed; partitions are sent to independent compute nodes on a cluster

Compute nodes operate distributed parallel using MPI communication to complete a solution per time step



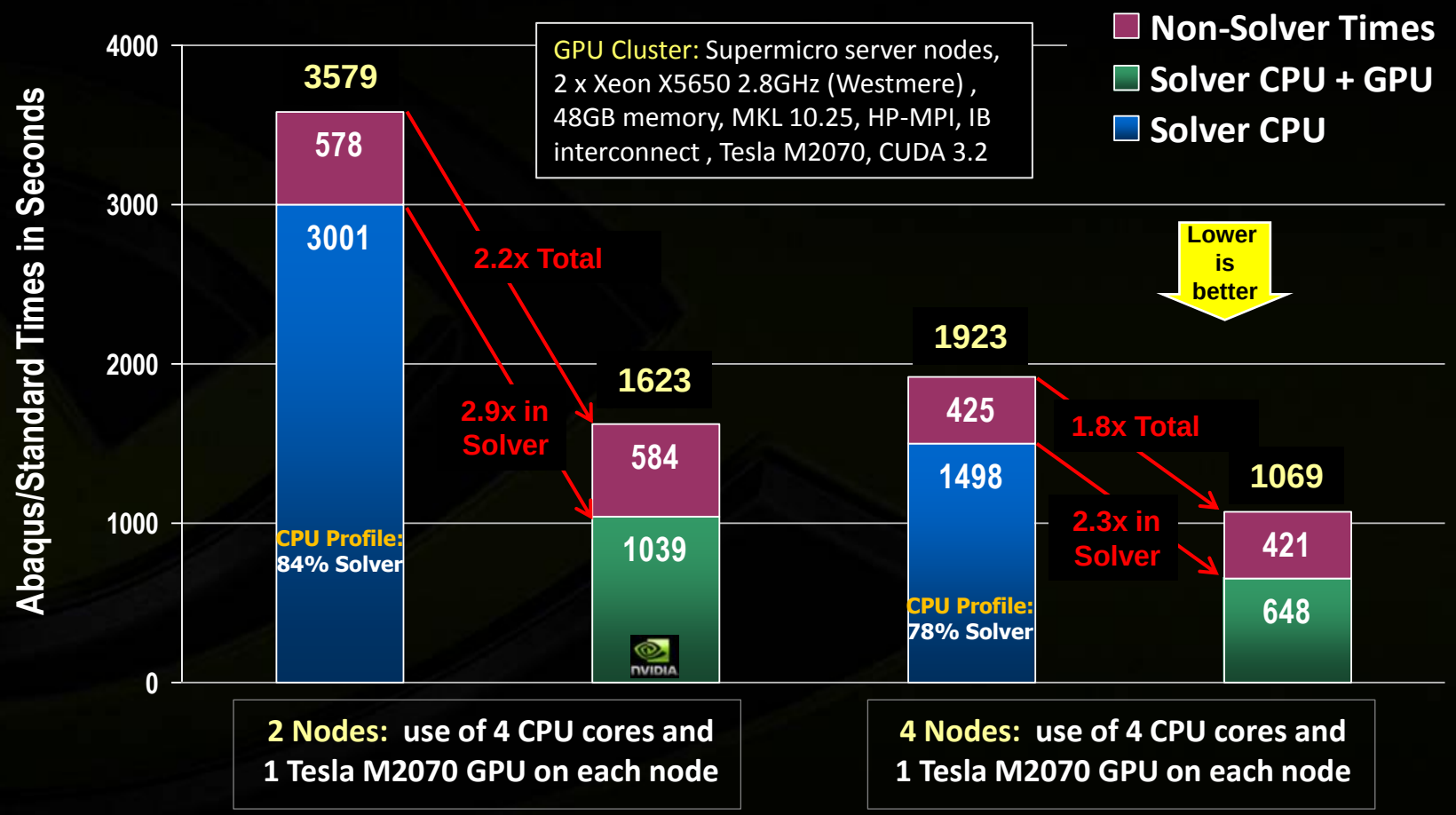


A global solution is developed for the completed job

Model partition is mapped to a GPU and uses shared memory OpenMP, p-threads parallel, a second level of parallelism in a hybrid model

GPU Cluster Demonstration of Abaqus 6.11 DMP

NOTE: Results Based on Abaqus/Standard 6.11 Direct Solver and CUDA 3.2



Engine Model

- Engine geometry
- 4,500 K DOF
- Solid FEs
- Static, nonlinear
- 1 Load step
- 8 iterations

How NVIDIA Tesla GPU Products are Deployed

	Data Center Products		Workstation
	Tesla M2050 / M2070 Adapter	Tesla S2050 1U System	Tesla C2050 / C2070 Workstation Board
			
GPUs	**1 Tesla GPU**	**4 Tesla GPUs**	**1 Tesla GPU**
Single Precision	1030 Gigaflops	4120 Gigaflops	1030 Gigaflops
Double Precision	515 Gigaflops	2060 Gigaflops	515 Gigaflops
Memory	3 GB / 6 GB	12 GB (3 GB / GPU)	3 GB / 6 GB
Memory B/W	148 GB/s	148 GB/s	144 GB/s

Recommended System Configurations

Workstations with Tesla GPUs

Servers with Tesla GPUs

Workstations

Existing System

- Tesla C2050 (3 GB) for < 5M DOF
- Tesla C2070 (6 GB) for > 5M DOF

New System Purchase

- Total 6-8 CPU cores
- Total 32 - 48 GBs of CPU memory
- Disk with minimum 500 GB
- Tesla C2050 or Tesla C2070
  + Quadro FX380 for pre/post

-- OR --

- Quadro 5000 (2.5GB) or 6000 (6GB)

Servers

Existing System

- Tesla S2050 (12 GB or 4 GB/GPU)

New System Purchase

- Total 4 CPUs, 6-8 CPU cores each
- Total 4 x16 PCIe (one for each GPU)
- Total 96 to128 GBs of CPU memory
- Disk with minimum 2000 GB (scratch)
- Tesla M2050 or Tesla M2070

Summary of Abaqus Benefits for GPU Computing

- Abaqus/Standard supports NVIDIA CUDA and Tesla GPUs
  - Single GPU release in 6.11 for CUDA 3.2 with Tesla and Quadro GPUs
- Abaqus/Standard initial developments are only the beginning
  - Solver today, more of Abaqus will be moved to GPUs in progressive stages
- Two talks at NVIDIA's GTC on Abaqus are available for viewing:
  - http://www.nvidia.com/gtc2010-content
    - *GPGPUs in the real world. The ABAQUS experience – SIMULIA Corp.*
    - *Acceleration of SIMULIA's Abaqus Solver on NVIDIA GPUs – Acceleware Inc.*
- SIMULIA and NVIDIA have a well established technology alliance
  - Joint investments in technology exchange to benefit industry

Contributors to the SIMULIA Performance Studies

SIMULIA

- Mr. Matt Dunbar, Technical Staff, Parallel Solver Development
- Mr. Michael Wood, Technical Staff, Development Engineer
- Dr. Luis Crivelli, Technical Staff, Parallel Solver Development

NVIDIA

- Mr. Armando Serrato, Technical Staff, NVIDIA Performance Lab
- Mr. Joe Orr, Technical Staff, NVIDIA Performance Lab

Thank You, Questions ?

Stan Posey | CAE Market Development

Srinivas Kodiyalam| CAE Strategic Alliances

NVIDIA, Santa Clara, CA, USA