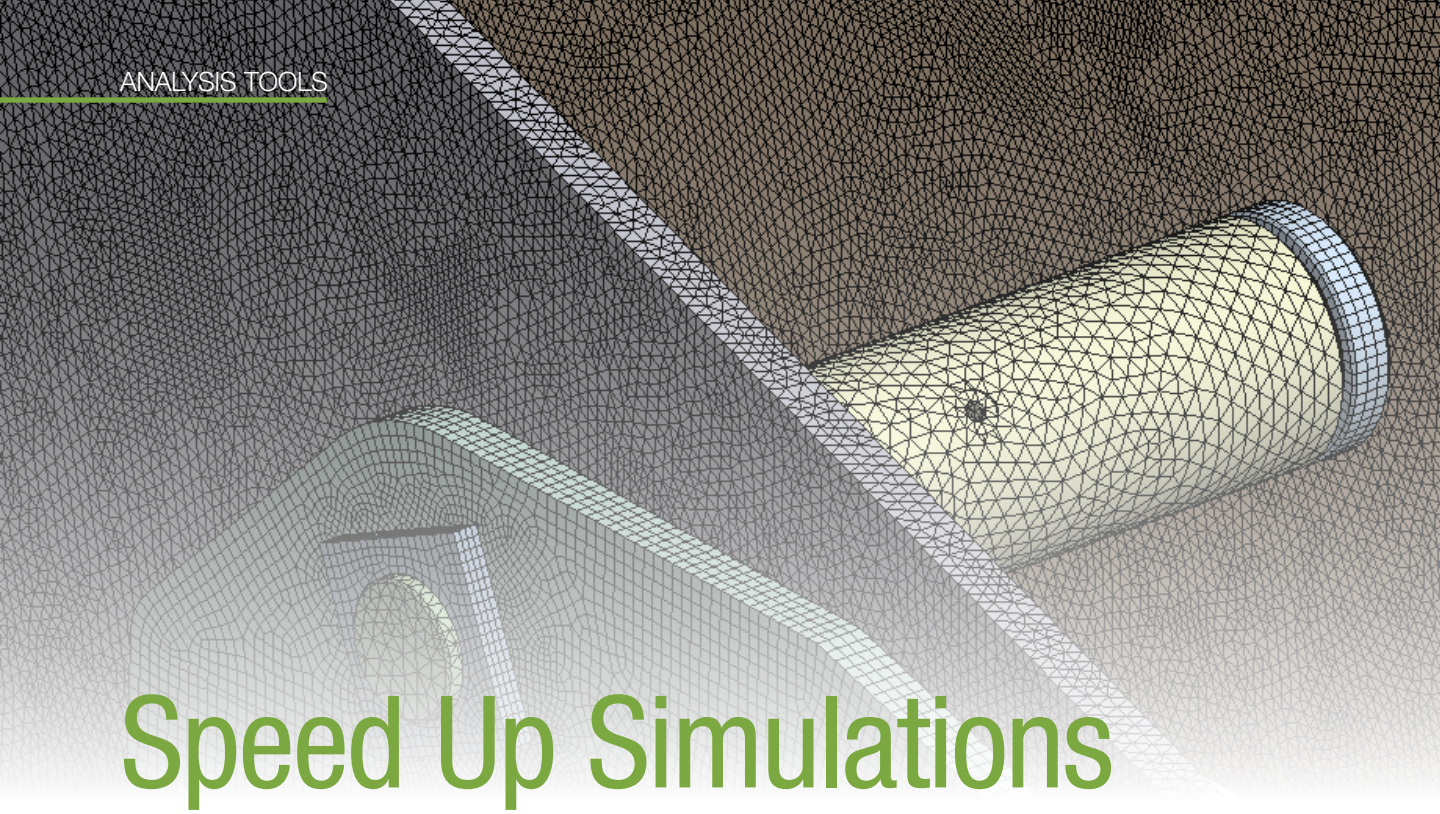


Speed Up Simulations with a GPU

A new feature in ANSYS Mechanical leverages graphics processing units to significantly lower solution times for large analysis problem sizes.

By Jeff Beisheim, Senior Software Developer, ANSYS, Inc.

Engineers looking to improve turnaround times for increasingly complex engineering simulations — especially those involving models with complicated multiple physics and greatly refined meshes — may want to investigate using a graphics processing unit (GPU). These devices have been around for years working alongside the computer's CPU to speed up graphics operations. Only recently have GPUs been specifically developed with the computational precision required for finite element analysis (FEA) solutions as well as the computational power to effectively complement the performance of the latest CPUs. With hundreds of low-power cores on a single socket, GPUs have the potential to dramatically increase computing capacity, provided that the compute workload will fit in the available memory of the GPU.

A new feature called the general-purpose GPU accelerator capability is available in a preview version of ANSYS 13.0 to take advantage of these high-end devices when performing structural mechanics simulations. The accelerator works by offloading some of the most numeric-intensive algorithms from the CPU onto the GPU. These algorithms are part of the equation solutions occurring during a simulation. Other computations during simulation remain on the CPU. In this way, the GPU and CPU work in a collaborative fashion to help speed up the time to solution. While limited to simulations using the shared-memory solvers of the prerelease version ANSYS 13.0, this initial support of GPU computing for structural mechanics simulations is an important first step in leveraging the devices as a significant new resource in high-performance computing (HPC).

Initial GPU Implementation

The GPU accelerator capability accelerates only the shared-memory equation solvers that support the usage of a GPU — the sparse direct and preconditioned conjugate gradient (PCG)/Jacobi conjugate gradient (JCG) iterative solvers. This includes the use of block Lanczos and PCG Lanczos solvers in an eigenvalue buckling or mode frequency analysis. Other equation solvers will continue to run on only the CPU and will not see any performance benefit when using the GPU accelerator capability.

Other limitations when using the GPU accelerator capability include:

- Windows® x64 and Linux® x64 are currently the only platforms supported. Windows users should be aware that use of remote desktop disables the use of the GPU to accelerate structural mechanics simulations.

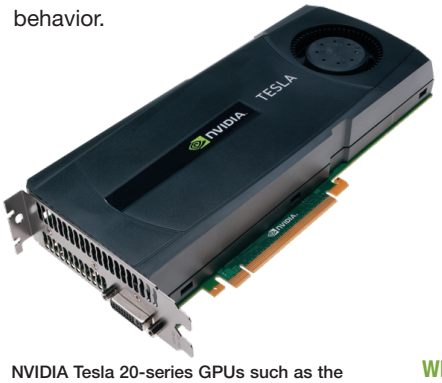
- Only NVIDIA® Tesla™ GPU is currently supported for use when accelerating ANSYS structural mechanics simulations. Only the more powerful 20-series (Fermi) GPUs are recommended, as these are the most computationally powerful and, therefore, the most likely to produce faster solution times.
- The GPU accelerator capability is not currently supported when using Distributed ANSYS.

Activating the New Feature

For commercial license users, the GPU accelerator capability is enabled using the ANSYS HPC Pack licensing model. For academic license users, the GPU capability is included within the base ANSYS Academic product (that provides access to ANSYS Mechanical or higher capability) and no add-on Academic HPC licenses are required. Engineers can use a GPU to accelerate computations on conventional multicore processors without any additional GPU-specific licensing required. During structural mechanics simulations, ANSYS Mechanical APDL software makes use of only a single GPU per simulation.

ANSYS Mechanical APDL users can activate the accelerator capability simply by selecting the `High-Performance Setup` tab in the launcher and then checking the `GPU Accelerator Capability` box. Alternatively, `-acc nvidia` can be added to the list of arguments supplied on the ANSYS Mechanical APDL command line. ANSYS Workbench users can choose to activate the GPU accelerator capability during solution by modifying the GPU acceleration option on the `Advanced Properties` page of the `Solve Process Settings`.

Once the GPU accelerator capability is activated, when ANSYS Mechanical APDL is launched it should accelerate the solution, when possible, without requiring input from the user. For cases in which it does not apply, this new feature will simply have no effect on the program behavior.



NVIDIA Tesla 20-series GPUs such as the C2050 and C2070 are the most computationally powerful and, therefore, most likely to produce faster solution times in ANSYS simulations.

Optional Control Settings

A new ACCOPTION command is available for users who want additional control over various settings related to the GPU accelerator capability:

- `Activate` to control which analysis will use/not use the GPU accelerator capability
- `MinSzThresh`, a threshold parameter to determine when the sparse direct solver data size is large enough to justify using the GPU
- `SPkey` to control the use of single- or double-precision math operations when running the sparse direct solver on the GPU

In addition, some hardware settings for NVIDIA GPU cards can be useful under certain scenarios:

- Environment variables are available in ANSYS Mechanical APDL to help avoid over-subscribing the GPU hardware for users with multiple GPU cards or users who run in a multi-user environment, such as a server.
- NVIDIA GPU users can consider switching their hardware to exclusive mode, which allows only one process — for example, one ANSYS Mechanical simulation — to be run at a time on the GPU.
- Another hardware setting for NVIDIA Tesla 20-series GPUs disables error correcting (ECC) memory to make use of all the memory on the GPU card as well as to increase overall memory bandwidth and GPU performance. To ensure FEA result accuracy, however, it is recommended that users keep the default setting of ECC memory enabled.

When to Use a GPU

The amount of acceleration achievable when using the GPU accelerator capability will vary greatly depending on the hardware being used and the model being simulated. The following guidelines can help determine whether use of the GPU accelerator capability will provide a performance boost. In general, the capability provides the greatest reductions in overall simulation time when the following conditions are met:

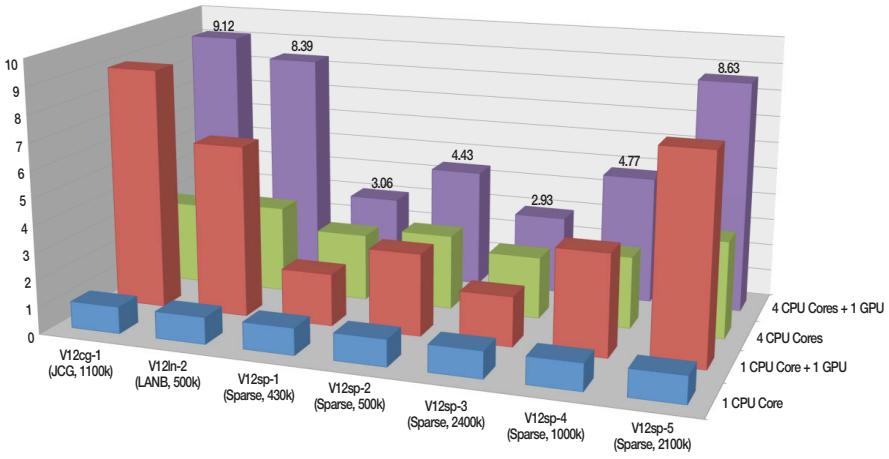
- The simulation spends most of its time on the numerical analysis solution rather than other tasks, such as pre- and post-processing. Only the operation of the solver is accelerated with a GPU, including analyses that use the sparse direct or PCG /JCG iterative solvers (including block Lanczos and PCG Lanczos eigensolvers).
- The problem size is in the following ranges:
  - 500K to 5,000K DOFs for the sparse direct solver
  - 500K to 3,000K DOFs for PCG/JCG iterative solvers

Size guidelines listed above represent the general range many users now routinely work within and are based on the NVIDIA Tesla C2050,

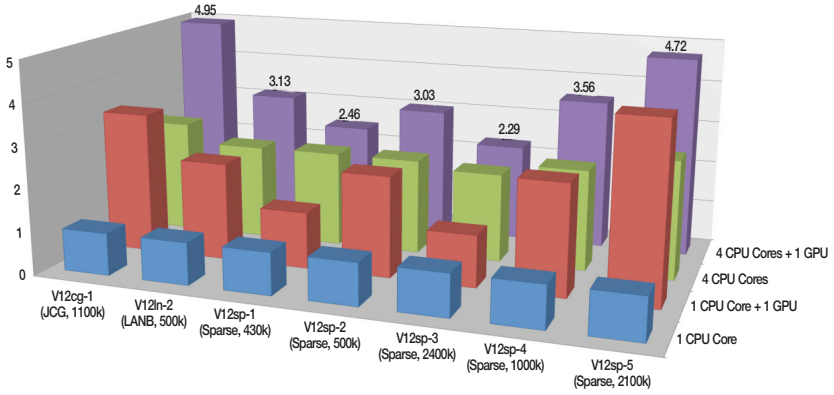
which offers 3 GB of memory. GPUs containing more or less memory can be expected to adjust these sizes accordingly. For simulations involving model sizes outside these ranges, some acceleration may be achieved, but generally the code will avoid using the GPU and run the entire simulation using the CPU cores.

When using the sparse direct solver, all analysis types are supported except substructuring. Models that create nonsymmetric matrices — such as frictional contact models that use the NROPT, UNSYM command — are supported with the GPU accelerator capability. However, models that require the use of partial pivoting are not supported. Partial pivoting is activated by the solver automatically when certain element types and options are included, such as current-technology elements containing the mixed u-P formulation option and contact elements with the pure Lagrange formulation option. In these cases, the GPU accelerator capability is deactivated and the solution proceeds using the CPU.

When using the PCG/JCG solver, all analysis types supported by these solvers are also supported when using the GPU accelerator capability. The use of the memory-saving option (MSAVE command) for the PCG solver will deactivate the capability.



GPU accelerator capability of solver kernel speedups using a prerelease version of ANSYS 13.0



GPU accelerator capability of overall speedups using a prerelease version of ANSYS 13.0

Comparing GPU and CPU Performance

Concerning the amount of speedup obtained when using a GPU, it is important to clarify what is being evaluated. Comparison to a single CPU core, for example, may look impressive for software like the ANSYS Mechanical product, which scales on multiple CPU cores, but it is not an accurate basis for comparison of performance. In the interest of accuracy, ANSYS performance benchmarking was done on an HP® Z800 Workstation with 32 GB of RAM. The process focused on using all four cores of an Intel® Xeon® 5560 series processor (2.8 GHz) as well as double precision computations. The GPU used for this benchmarking was a Tesla C2050 with ECC memory enabled.

Results are shown in the accompanying charts for seven of the 10 problems contained in the ANSYS 12.0 benchmark set using a prerelease version of ANSYS Mechanical APDL software. Of the three remaining benchmarks, two use the MSAVE option, making them invalid with the GPU accelerator capability, and one was too large to be run using the Tesla C2050. Results demonstrate that using a GPU with one or more CPU cores can lead to impressive speedups in number-crunching equation solver kernels, resulting in impressive acceleration for overall simulation times. Since the GPU accelerates only the key equation solver kernels and nothing else, the overall speedups are expected to be lower as other simulation processes are involved in the timings.

Future Directions

As GPU computing trends evolve, ANSYS will continue to enhance its offerings as necessary for a variety of simulation products. Certainly, performance improvements will continue as GPUs become computationally more powerful and extend their functionality to other areas of ANSYS software. ANSYS is investigating the use of AMD/ATI GPU cards to accelerate simulation. The company is also investigating the potential for supporting multiple GPUs and Distributed ANSYS. ■