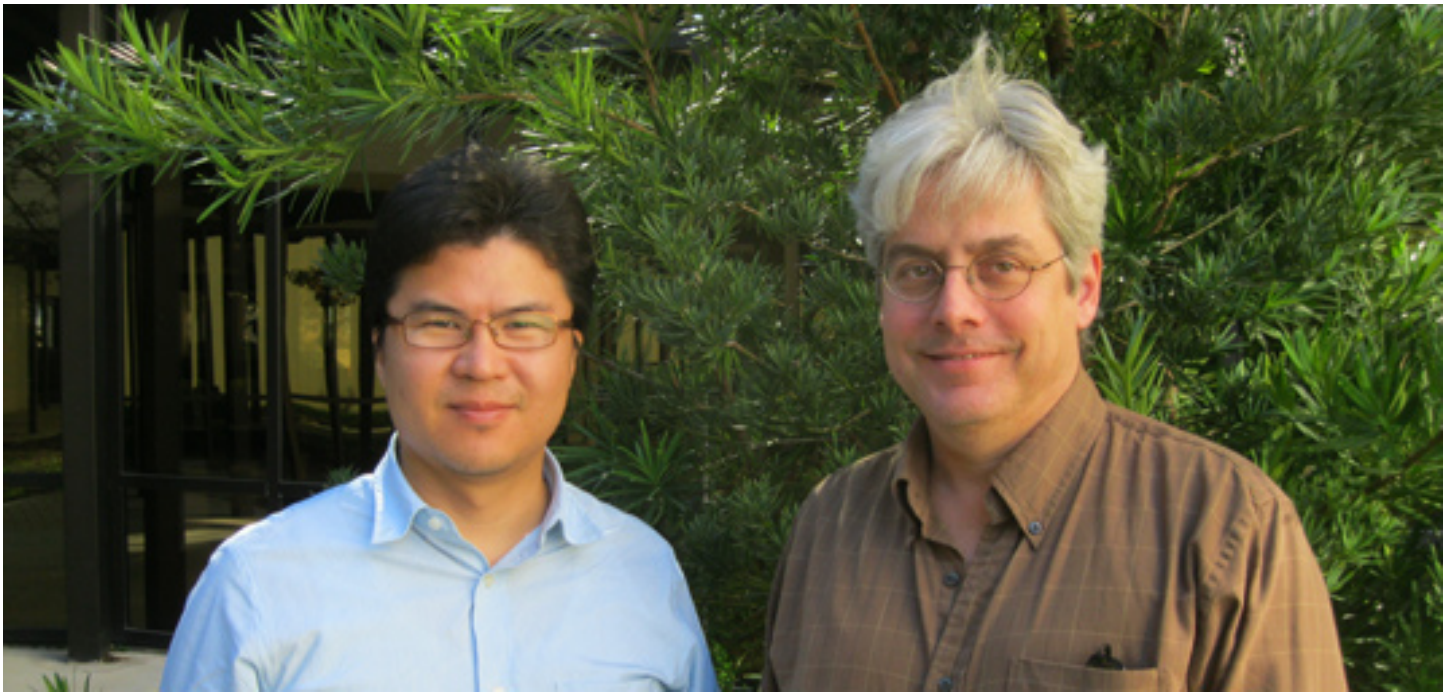
THE UNIVERSITY OF TEXAS AT AUSTIN
TEXAS ADVANCED COMPUTING CENTER

8 Things You Should Know About GPGPU Technology

Q&A with TACC Research Scientists



Byungil Jeong (left) and Greg Abram (right) in TACC's Data and Information Analysis Group.

1 **What is GPGPU computing?**

GPGPU (general purpose computing on graphics processing units) is a methodology for high-performance computing that uses graphics processing units to crunch data. The characteristics of graphics algorithms that have enabled the development of extremely high-performance special purpose graphics processors show up in other HPC algorithms. This same special-purpose hardware can be put to use accelerating those algorithms as well.

2 **What applications work best with GPGPUs? Why?**

Algorithms well-suited to GPGPU implementation are those that exhibit two properties: they are data parallel and throughput intensive. Data parallel means that a processor can execute the operation on different data elements simultaneously. Throughput intensive means that the algorithm is going to process lots of data elements, so there will be plenty to operate on in parallel. Taking advantage of these two properties, GPUs achieve extreme performance by incorporating lots (hundreds) of relatively simple processing units to operate on many data elements simultaneously.

Perhaps not surprisingly, pixel-based applications such as computer vision and video and image processing are very well suited to GPGPU technology, and for this reason, many of the commercial software packages in these areas now include GPGPU acceleration. Physical simulation applications, which often rely on numerical solutions to PDEs and on the heavy use of linear algebra operations, are also very well-suited to GPGPU technology. In the area of rendering, ray tracing (a computationally intensive algorithm for highly realistic image synthesis) uses GPGPU technology to compute ray-object intersections, making real-time ray tracing achievable.

3 **How is a GPU different from a CPU?**

GPU properties lead to a very different processor architecture from traditional CPUs. CPUs devote a lot of resources (primarily chip area) to make single streams of instructions run fast, including caching to hide memory latency and complex instruction-stream processing (pipelining, out-of-order execution and speculative execution). GPUs, on the other hand, use the chip area

area for hundreds of individual processing elements that execute a single instruction stream on many data elements simultaneously. Memory latency is hidden by very fast context switching; when a memory fetch is issued while processing one subset of data elements, that subset is set aside in favor of another subset that is not waiting on a memory reference.

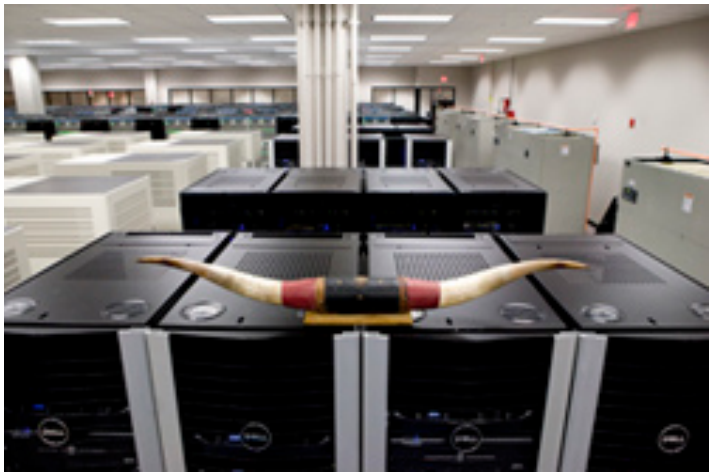
4 **What advantages do GPUs offer? Are there any disadvantages?**

GPUs can run certain algorithms anywhere from 10 to 100 or more times faster than CPUs—a huge advantage.

There are two disadvantages: 1) Gaining this speedup requires that algorithms are coded to reflect the GPU architecture, and programming for the GPU differs significantly from traditional CPUs. In particular, incorporating GPU acceleration into pre-existing codes is more difficult than just moving from one CPU family to another; a GPU-savvy programmer will need to dive into the code and make significant changes to critical components. 2) Incorporating GPU hardware into systems adds expense in terms of power consumption, heat production, and cost. Some job mixes may be served more economically by systems that maximize the number of CPUs that can be brought to bear.

How is the field of advanced computing being impacted by GPGPU technology?

5 GPGPU computing is making a significant impact on high-performance computing in a wide range of application domains. Widely used HPC codes in areas including weather forecasting, molecular dynamics, and fluid-flow are being updated to incorporate GPGPU acceleration. Popular commercial scientific and engineering applications now provide GPGPU acceleration, notably MATLAB and ANSYS; open-source systems including AMBER, LAMMPS, NAMD; and Gromacs are now using GPGPU technology. Even more significant is the wide acceptance of GPGPU technology in new code development for leading-edge scientific and engineering research.



Longhorn, the largest hardware-accelerated interactive visualization cluster in the world, is located at TACC.

How is TACC using GPGPU technology?

6 High-performance computing has long relied on parallel systems to achieve the performance required by leading-edge science and engineering research. With the advent of GPGPU technology, the capability of single nodes is greatly enhanced, but still does not provide sufficient performance for current HPC applications, which currently rely on systems of tens of thousands of individual CPUs working in concert to achieve the necessary performance.

TACC's HPC servers meet these needs by using a distributed-memory, message-passing system architecture. This approach achieves performance goals by combining relatively low-cost, multi-core nodes with a high-speed interconnect for explicit message passing. This approach is well suited to GPGPU technology—some or all of a cluster's nodes can be enhanced with GPUs enabling two levels of parallelism to achieve extreme performance levels. In January 2010, TACC installed Longhorn, a 256- node Dell visualization and data analysis cluster with two powerful NVIDIA GPUs in each node. In addition to its use as a production visualization supercomputer, it also serves as a research and development platform for investigating how cluster and GPGPU parallel technologies can be effectively hybridized.

What tools are deployed on these GPU clusters?

7 TACC supports both OpenCL, the NVIDIA implementation of an open standard for GPGPU computing, and CUDA, the NVIDIA proprietary GPGPU tool chain. NVIDIA has invested heavily in developing a complete, robust set of tools including a compiler, debugger and profiler. In addition, NVIDIA's GPGPU implementation of several core HPC libraries providing linear algebra (CUBLAS) and the fast Fourier transform (CUFFT) are available on Longhorn.

Where can I learn more about GPU programming?

8 NVIDIA provides a useful website where you will find full documentation on the CUDA toolset.