

# Built for Speed

## Graphics Processors for General-Purpose Computing

**W**HILE computer gamers are eagerly awaiting the next generation of platforms, the computer scientists of Lawrence Livermore’s Graphics Architectures for Intelligence Applications (GAIA) project are tracking the rapidly changing technology, but for a different reason. A team, led by John Johnson of the Computation Directorate, is researching graphics processing units (GPUs)—the highly specialized, low-cost, rendering engines at the heart of the gaming industry—to determine how they might be programmed and used in applications other than virtual entertainment.

“Graphics processors are accelerating in performance much faster than other microprocessors,” says Sheila Vaidya, project leader for GAIA. “We have an opportunity to ride the wave of innovations driving the gaming industry.” These processors—traditionally designed for fast rendering of visual simulations, virtual reality, and computer gaming—could provide efficient solutions to some of the most challenging computing needs facing the intelligence and military communities. Real-time data-processing capabilities are needed for applications ranging from text and speech processing to image analysis for automated targeting and tracking.

### Gaming the System

The GAIA team, including collaborators from Stanford University, the University of California at Berkeley and Davis, and Mississippi State University, is researching graphics processors used in the computer gaming and entertainment industries to determine how they might be used in knowledge-discovery applications of relevance to national security.

Why bother with this class of processors when plenty of central processing units (CPUs) exist to do the heavy-duty work in high-performance computing? Two words: speed and cost.

The ever-growing appetite in the three-dimensional (3D) interactive gaming community has led to the development and enhancement of GPUs at a rate faster than the performance of conventional microprocessors predicted by Moore’s Law. This



Dave Bremer, a member of the Graphics Architectures for Intelligence Applications team, holds a commercial graphics processing card. These units used for virtual entertainment are providing low-cost solutions for high-performance data processing.

acceleration in improved performance will likely continue as long as the demand exists and integrated-circuit technologies continue to scale.

During the past 2 years, the GAIA team has implemented many algorithms on current-generation CPUs and GPUs to compare their performance. The benchmarks that followed showed amazing performance gains of one to two orders of magnitude on GPUs for a variety of applications, such as georegistration, hyperspectral imaging, speech recognition, image processing, bioinformatics, and seismic exploration.

GPUs have a number of features that make them attractive for both image- and data-processing applications. For example, they are designed to exploit the highly parallel nature of graphics-rendering algorithms, and they efficiently use the hundreds of processing units available on-chip for parallel computing. Thus, one operation can be simultaneously performed on multiple data sets in an architecture known as single-instruction, multiple data (SIMD), providing extremely high-performance arithmetic capabilities for specific classes of applications. Current high-end GPU chips can handle up to 24 pipelines of data per chip and perform hundreds of billions of operations per second.

Today’s commercial GPUs are relatively inexpensive as well. “National retailers charge a few hundred dollars for one, compared to the thousands of dollars or more that a custom-built coprocessor might cost,” says Johnson.

The performance of these GPUs is impressive when compared with that of even the newest CPUs. “A modern

CPU performs about 25 billion floating-point operations per second,” says Johnson. “Whereas a leading-edge GPU, such as the NVIDIA® GeForce™ 7800 GTX video card or the upcoming successor to the ATI Radeon® X850, performs six times faster at half the cost of a CPU.” These GPUs are optimized for calculating the floating-point arithmetic associated with 3D graphics and for performing large numbers of operations simultaneously.

GPUs also feature a high on-chip memory bandwidth, that is, a large data-carrying capacity, and have begun to support more advanced instructions used in general-purpose computing. When combined with conventional CPUs and some artful programming, these devices could be used for a variety of high-throughput applications.

“GPUs work well on problems that can be broken down into many small, independent tasks,” explains GAIA team member Dave Bremer. Each task in the problem is matched with a pixel in

an output image. A short program is loaded into the GPU, which is executed once for every pixel drawn, and the results from each execution are stored in an image. As the image is being drawn, many tasks are being executed simultaneously through the GPU’s numerous pipelines. Finally, the results of the problem are copied back to an adjacent CPU.

However, general-purpose programming on GPUs still poses significant challenges. Because the tasks performed on a GPU occur in an order that is not controlled by a programmer, no one task can depend on the results of a previous one, and tasks cannot write to the same memory. Consequently, image convolution operations work extremely well (100 times faster) because output pixels are computed independently, but computing a global sum becomes very complex because there is no shared memory. “Data must be copied in and out of the GPU over a relatively slow transmission path,” says GAIA team member Jeremy Meredith. “As a result, memory-intensive computations that require arbitrary access to large amounts of memory off-chip are not well suited to the GPU architecture.”

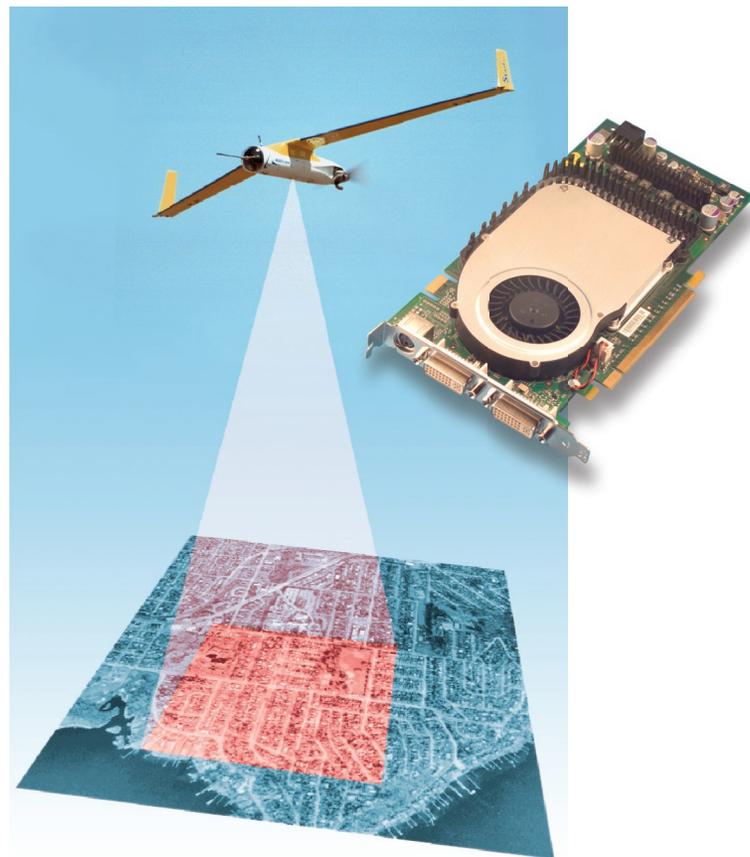
Today’s GPUs are power hungry. But designers, faced with the growing demand for mobile computing, are rapidly evolving chip architectures to develop low-power versions that will approach the performance of high-end workstations.

### What’s in the Pipeline

“GPUs are beginning to more closely resemble CPUs with every evolution,” notes Johnson. “The drawbacks for general-purpose programming are being tackled by the industry, one by one.” Next-generation CPU architectures are adopting many features from GPUs. “Emerging architectural designs such as those found in Stanford’s Merrimac and the IBM–Toshiba–Sony Cell processor look similar to the architecture of GPUs,” says Johnson. “These designs could be the next-generation technology for real-time, data-processing applications. Our work with GPUs will help us evaluate and deploy the emerging devices.”

The Cell processor, which is a crossover GPU–CPU chip, is scheduled to hit the gaming market soon. But the Cell might also prove to be useful in defense and security computing environments. The scientists of GAIA—just like the gamers—are eager to test and scale its limits.

—Ann Parker



The use of graphics processing units (inset) for real-time georegistration of high-resolution wide-area imaging was demonstrated by Livermore’s Sonoma project. Graphics processing on remote-collection platforms, such as in the one in this aircraft, reduces the volume of data transferred to users, enabling real-time analysis.

**Key Words:** central processing unit (CPU), general-purpose programming, Graphics Architectures for Intelligence Applications (GAIA) project, graphics processing unit (GPU), knowledge discovery, streaming architectures.

**For further information contact John Johnson (925) 424-4092 (jjohnson@llnl.gov) or Sheila Vaidya (925) 423-5428 (vaidya1@llnl.gov).**