

The GPU Enters Computing's Mainstream

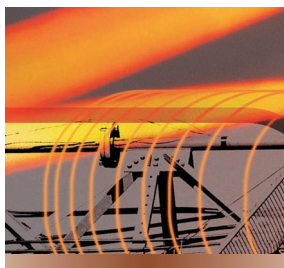
Michael Macedonia, Georgia Tech Research Institute

The Siggraph/Eurographics Graphics Hardware 2003 workshop (graphicshardware.org), held in San Diego, will likely be remembered as a turning point in modern computing. In one of those rare moments when a new paradigm visibly begins changing general-purpose computing's course, what has traditionally been a graphics-centric workshop shifted its attention to the nongraphics applications of the graphics processing unit.

GPUs, made by Nvidia (www.nvidia.com) and ATI (www.ati.com), function as components in graphics subsystems that power everything from Microsoft's Xbox to high-end visualization systems from Hewlett-Packard and SGI. The GPUs act as coprocessors to CPUs such as Intel's Pentium, using a fast bus such as Intel's Advanced Graphics Port. AGP8x has a peak bandwidth of 2.1 gigabytes per second—speed it needs to avoid inflicting bus starvation on data-hungry GPU coprocessors.

RAW PERFORMANCE

The new GPUs are very fast. Stanford's Ian Buck calculates that the current GeForce FX 5900's performance peaks at 20 Gigaflops, the equivalent of a 10-GHz Pentium—with, according to Nvidia, even more speed on the horizon. Performance growth has multiplied at a rate of 2.8 times per year since 1993, a pace analysts expect the industry to maintain



The graphical processing unit is visually and visibly changing the course of general-purpose computing.

for another five years. At this rate, GPU performance will move inexorably into the teraflop range by 2005.

What does this mean for consumers? For less than \$400 they can buy an off-the-shelf graphics card today with performance comparable to a top-of-the-line image generator that cost hundreds of thousands of dollars in 1999. Intense competition has driven this explosive growth, with Nvidia and ATI leapfrogging each other's introduction of a new GPU generation every six months.

Recently, this battle expanded beyond the PC to include video game consoles. Nvidia, which years ago staged a coup by locking down the rights to make the GPU for Microsoft's Xbox, lost out when Microsoft awarded the GPU manufacturing rights for the Xbox's next-generation successor to ATI.

STREAM PROCESSING

Stream processing gives GPUs their impressive speed. Streams constitute a "computational primitive because large amounts of data arrive continuously,

and it's impractical or unnecessary to retain the entire data set" (graphics.stanford.edu/papers/humper_thesis/humper_thesis.pdf; p.30). According to Fred Brooks, a developer of IBM's Stretch supercomputers, stream processing goes back to the 1950s when IBM built the Harvest variant of the Stretch system for the National Security Agency. Harvest treats large volumes of encrypted text as a stream and has a special coprocessor designed specifically to handle this task.

DARPA has also backed Bill Dally's Imagine stream processor work at

Stanford (U.J. Kapasi et al., "Programmable Stream Processors," *Computer*, August 2003, pp. 54-62). Imagine, a programmable architecture, "executes stream-based programs and is able to sustain tens of Gflops over a range of media applications with a power dissipation of less than 10 Watts" (cva.stanford.edu/imagine/).

Similarly, GPUs treat computer graphics primitives such as vertices and pixels as streams. Multiple programmable processing units connect via data flows. In a GPU, a vertex processor transforms and processes points—as a four-component vector—and a fragment processor computes pixel color. As a stream processor, a GPU performs simple operations and exploits spatial parallelism. For example, the fragment processor runs the same program for each pixel. In the case of the ATI R300, eight-pixel pipelines handle single-instruction, multiple-computing processing.

GPUs' simple architectures devote large areas of chip real estate to the computational engines. For example,

the 0.15-micron-process ATI R300 chip has more than 110 million transistors, while Intel's Xeon microprocessor has 108 million. However, the Xeon chip devotes more than 60 percent of its transistors to cache (www.anandtech.com/printarticle.html?i=1749). The Nvidia NV40, to be announced this fall and manufactured using a 0.13 process, is rumored to have 150 million transistors.

HACKING THE GPU

Two years ago, Nvidia and ATI shook the computer graphics community by making their GPUs programmable. Programmability—nothing new in the graphics world, as a description of the Pixel Flow architecture reveals (www.cs.unc.edu/~molnar/Papers/PxFHhws97.pdf)—provided a dramatic leap forward for commodity graphics processors. With programmability, developers could code small programs, known as shaders, to perform real-time rendering effects such as bump mapping or shadows. These shaders can be written in assembly or languages such as Nvidia's Cg (www.cgshaders.org/) or Microsoft's High-Level Shader Language (msdn.microsoft.com/directx/).

Microsoft has been a driving force in making GPUs programmable, starting with its introduction of DirectX 8.0 as the 3D graphics API for MS Windows and, in variant form, for the Xbox. Previous DirectX versions worked as fixed-function pipelines, meaning developers could not program their hardware and had to rely on static API functions.

The DirectX 9.0 specification provided the next major boost in GPU development. First, DX9 allows much longer shader programs and gives developers a greatly expanded pixel-shading instruction set. Second and most important, DX9 breaks the mathematical precision barrier that had previously limited PC graphics. Precision, and therefore visual quality, increases when displayed with 128-bit floating-point color per pixel (www.nvidia.com/object/dx9_tb.html).

Floating-point color has opened up

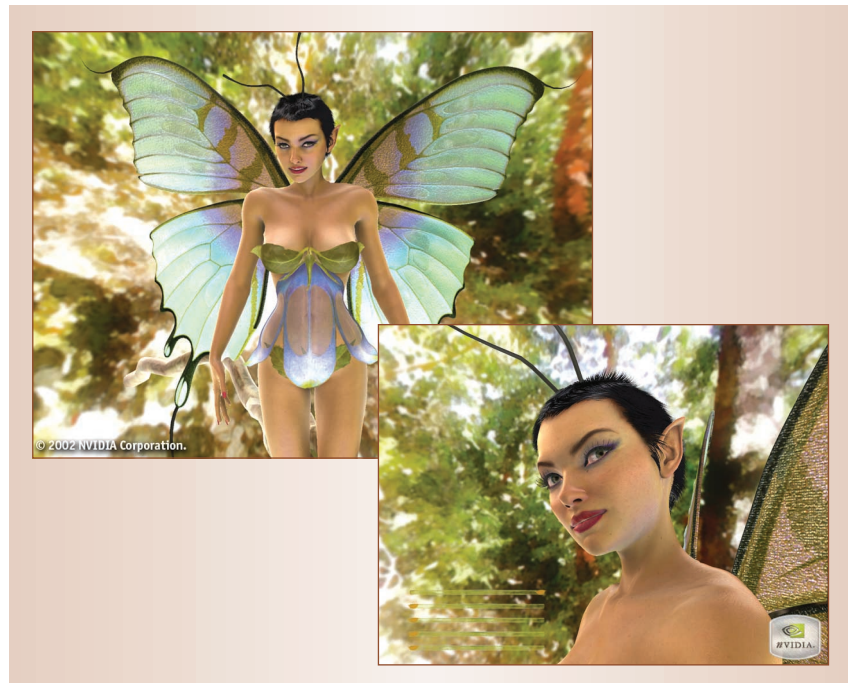


Figure 1. Nvidia's cutting-edge Siggraph demo. Developers rendered this exquisitely animated fairy in real time, leveraging the power of today's GPUs.

the new field of high dynamic range imaging and image-based lighting. Image-based lighting illuminates objects with images of light from the real world, and HDRI techniques imbue the scene with the same dynamic range of real light. Earlier GPUs could provide only 8 bits for each red, green, and blue color value, which limited the dynamic range to 256 levels. Floating-point color removes this restriction, allowing a dramatically better representation of real light values. Most importantly, the introduction of 32-bit floating-point capability brings GPUs closer to true general-purpose computing.

THE ULTIMATE MEDIA PROCESSOR

This performance and programmability revolution has already resulted in spectacular computer graphics. Nvidia markets its GPUs as hardware capable of enabling a "cinematic computing experience" because interactive graphics now approach the quality of fully rendered motion-picture animation. As Figure 1 shows, Nvidia supplied proof of this progress at Siggraph

2003 with its *Dawn* fairy demo, produced by Joe Demers. The real-time demo is available online at www.nvidia.com/object/demo_dawn.html.

Dawn stands at the threshold of a new art form. In a case of art imitating reality, a new class of movies called *machinima* uses video game engines such as *Unreal Tournament* to render their visuals (www.machinima.com/). High-quality, real-time rendering may soon give way to creating full-length commercial movies with this technique.

BEYOND GRAPHICS

In a recent *Wired* magazine interview, Nvidia CEO Jen-Hsun Huang said, "The microprocessor will be dedicated to other things like artificial intelligence. That trend is helpful to us. It's a trend that's inevitable." (www.wired.com/wired/archive/10.07/Nvidia.html?pg=1&topic=&topic_set=) Surprisingly, GPUs—not just CPUs—are making this vision real. The latest GPUs now perform nongraphics functions such as motion planning and physics for game AI, making them the

NEW FOR 2004!

IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING

This new journal will provide original results in research, design, and development of **dependable, secure computing** methodologies, strategies, and systems including...

Architecture for secure systems
■
Firewall and network technologies
■
Intrusion and error tolerance
■
Modeling and prediction
■
Emerging technologies

IEEE Computer Society
Members may **subscribe**
for the low member rate
of \$31!

DEPENDABLE
AND SECURE
COMPUTING

<http://computer.org/tdsc>



IEEE



IEEE
COMPUTER
SOCIETY

Entertainment Computing

ultimate processor for entertainment applications. In addition, GPUs can compute fast-Fourier transform functions such as real-time MPEG video compression or audio rendering for Dolby 5.1 sound systems.

This year's Graphics Hardware Workshop also provided a forum for speculating about the potential for GPU and stream-processor use in nonentertainment applications. Mark Harris has dedicated a Web site (www.gpgpu.org) to these and other general-purpose GPU applications. The site lists several papers describing GPU uses that range from linear algebra to simulating ice crystal growth.

For example, Dinesh Manocha and Ming Lin's research group at the University of North Carolina, Chapel Hill, has shown how GPUs can perform fast computation of Voronoi diagrams (www.cs.unc.edu/~geom/voronoi/). These diagrams partition a plane with n points into n convex polygons so that each polygon contains exactly one point, and every point in a given polygon is closer to its central point than to any other. Computations such as these can be useful in applications like robotics.

A pain to program

GPUs have a long way to go, however, before they become truly general purpose. Developers find programming current models painful because these processors lack many essential features such as real debuggers or profilers. Nor do they support branching. Nvidia's primitive language, Cg, offers no support for pointers. Worse, GPU vendors have limited programming on these devices to the DirectX or OpenGL APIs and have kept internal functions secret. If made public, researchers could use this information to explore new uses for these inexpensive but powerful devices.

On the horizon

DARPA may provide a solution to this mess by funding both new tools and new processors that exploit stream computing. For example, the agency is

funding Ian Buck's development of the Brook stream-processing language. Brook hides current GPUs' graphics abstractions, such as texture memory, and is intended to pave the way for future stream processors such as the DARPA-sponsored Tera-op reliable intelligently adaptive processing system (www.graphicshardware.org/presentations/Buck.ppt). The Trips chip, jointly developed by IBM and the University of Texas, will serve as part of an overall supercomputer design effort at IBM to develop a productive, easy-to-use, reliable computing system (www.research.ibm.com/resources/news/20030827_trips.shtml). DARPA also contributes around \$11 million annually to the project's funding. The system will reportedly perform one quadrillion operations per second and, by 2010, will be able to analyze its own workflow and optimize its own hardware and software resources.

Until GPUs mature and become much easier for developers to use, IBM will continue to help build supercomputers for entertainment. Console game market leader Sony invited IBM and Toshiba to collaborate on development of the Cell, a chip that will serve as the GPU for Sony's PlayStation 3.

Analysts estimate that the Cell will use a 50-nanometer process and operate at 4 GHz. Designing the chip required 300 engineers and cost \$400 million. Estimates show that developing the plant to produce these supercomputers-on-a-chip will cost more than \$1.6 billion (news.com.com/2100-1041_3-997596.html). Why is Sony willing to spend so much? To leverage the vast market it has created by selling 80 million PlayStation 2 consoles. ■

Michael Macedonia is a senior scientist at the Georgia Tech Research Institute, Atlanta. Contact him at macedonia@computer.org.