Taking the Graphics Processor beyond Graphics

David Geer

For several years, graphics processing units' performance has been increasing faster than the pace predicted by Moore's law. This has occurred because GPUs must meet the demands of increasingly complex visual effects in games and entertainment applications.

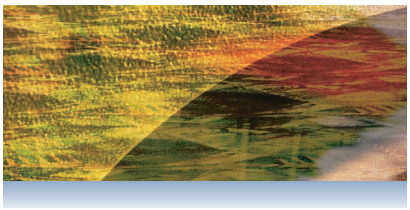
"Instead of doubling every 18 months, GPU performance has been increasing fivefold every 18 months. This is equivalent to doubling in just under every eight months," noted Ian Buck, software architect with Nvidia, a vendor of graphics and digital-media processors. Figure 1 illustrates this trend.

Their performance and functionality have made GPUs potentially attractive as coprocessors for general-purpose computation.

In recognition of this, major graphics chip manufacturers such as Nvidia and ATI Technologies have added support for floating-point computation and released compilers for high-level languages.

Researchers have also developed new algorithms and applications that exploit GPUs' parallelism and vector-processing capabilities, in which one operation can process an entire vector of numbers.

Users and researchers are thus increasingly working with *general-purpose GPUs* (GPGPUs) in areas other than gaming and entertainment, such as geometric, scientific, and database computations; medical imaging; and computer vision.



However, their complex programming environment and other challenges could affect the processors' ultimate popularity.

ABOUT THE GPU

GPUs typically are used in game consoles or as graphics coprocessors to CPUs, mainly for rendering geometric primitives such as polygons.

Researchers have studied the use of graphics hardware for general-purpose computation since the late 1970s, a process that accelerated with GPUs' wider deployment during the past few years.

Several factors have made GPUs more useful for some types of general-purpose computation. For example, vendors have added hardware to support branching, which lets programs alter their instructions based on results from previous instructions, said Nvidia software engineer Mark Harris.

This enables high-level language constructs like *if-then-else statements*—which let a system conditionally execute a group of statements depending on an expression's value—and *while loops*—which let a system repeatedly execute code based on a given Boolean condition. Both are useful for general-purpose computation, Harris explained.

Also, GPUs are widely available, commodity products that typically cost only about $500, noted Tim Purcell, graphics architect with Nvidia's Graphics Architecture Group. As for CPUs, said Jon Peddie, president of Jon Peddie Research, "A high-end Itanium can cost as much as $1,000."

High performance

GPUs frequently have slower clock speeds than premium CPUs, but because graphics chips handle work in parallel, they can offer more performance.

The G70, Nvidia's most recent GPU, performs up to 165 gigaflops. Intel declined to release the performance of its high-end CPUs. However, said University of Virginia assistant professor David Luebke, a 3-GHz, dual-core Intel Pentium 4 Extreme Edition's arithmetic units will theoretically run as much as 24.6 Gflops. Although Intel's fastest chip would offer somewhat more performance, it would still be considerably less than that of a high-end GPU.

"Intense competition between Nvidia and ATI Technologies has driven GPU speeds higher with each new processor release. And the competition by chip makers to have game-console makers use their products has intensified this process," said Tom Halfhill, senior analyst with In-Stat, a market research firm.

GPUs are highly parallel streaming processors optimized for vector operations. Streaming processors present data in a fixed order to processing units with limited memory, explained Suresh Venkatasubramanian, technical staff member of AT&T Labs' Information Visualization Research Group. Each unit performs a fixed set of operations on each data item and passes it on, he said.

A GPU's multiple-instruction, mul-

Published by the IEEE Computer Society

tiple-data (MIMD) pipelines perform vertex processing, which helps render specific points in 3D scenes based on their coordinates. Single-instruction, multiple-data (SIMD) pipelines produce colors and 3D effects for each pixel, the smallest unit of an image displayed on a screen.

Together, the two types of parallel-processing pipelines offer more performance than CPUs.

Strict pipelining, in which systems efficiently process all data items in pipeline order, enables GPUs to easily handle data without extensive caching.

"CPUs need caching because their programs are far more general, and the sequence of memory accesses is far less predictable than with the GPU, for which the program is explicitly constrained," noted Venkatasubramanian.

Reducing the number of on-chip caches leaves GPUs with more room for additional computational units, noted Nvidia's Purcell.

Programmability

Nvidia and ATI have made their commodity GPUs programmable so that the processors can be used more flexibly, such as for general-purpose computation.

Because GPUs are now programmable, chip makers have developed compilers that translate commonly used high-level languages such as C and C++ into the less-familiar languages, such as Cg (C for graphics), that the processors run, noted University of North Carolina professor Ming Lin.

32-bit floating-point capabilities

According to Purcell, his company and ATI recently added floating-point arithmetic logic units to their GPUs. This provides support for floating-point computation, critical for precision in both graphics and general-purpose applications.

Early GPUs could offer only eight-bit color. The eight bits of code available for each color limited a color's dynamic range to only 256 levels, said Lin.

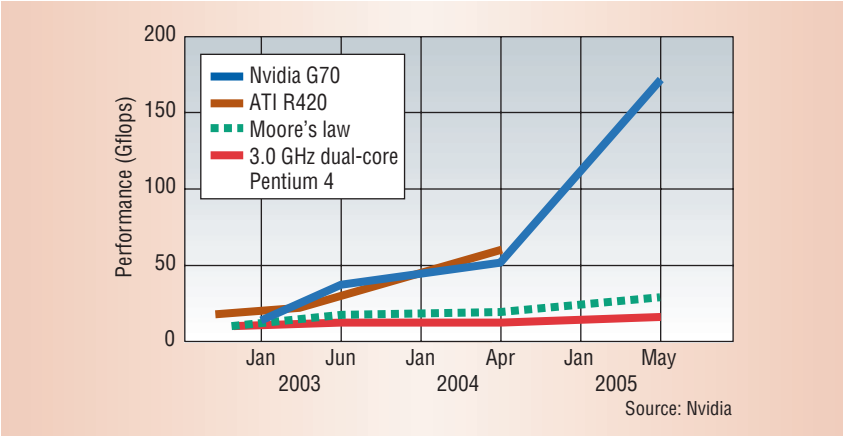


Figure 1. The maximum number of gigaflops produced by two leading graphics processing units, Nvidia's G70 and ATI Technologies' R420, show how GPU performance has improved faster than that called for by Moore's law and that of Intel's Pentium 4.

Current Nvidia and ATI GPUs support 16-bit floating-point color directly in hardware. Nvidia's and ATI's chips support 32-bit and 24-bit color, respectively, via additional programming, noted research assistant professor Naga Govindaraju of the University of North Carolina.

Introducing 32-bit floating-point capabilities added precision and the ability to perform more complex computations and thus made GPUs better able to handle general-purpose functions, explained Lin.

Memory bandwidth

The latest GPU architectures provide considerable memory bandwidth, allowing faster off-chip and on-chip data access and thereby increasing performance, said the University of Virginia's Luebke.

"Peak memory bandwidth is now 38.4 Gbytes per second on the Nvidia 7800 GTX," according to Govindaraju. A high-end CPU has a peak memory bandwidth of only 6.4 Gbps, he noted.

Chip makers have achieved high memory bandwidth in several ways. For example, new GPUs accelerate off-chip memory communications by using the PCI Express bus system, which implements existing peripheral-component-interconnect programming concepts and communications standards over a much faster serial communications system.

GPGPU USES

GPUs rely on the high arithmetic intensity necessary to process graphics, noted Luebke. Thus, he explained, applications that involve numeric computations on large grids of data are well suited to GPGPUs.

This includes linear algebra and the simulation of complex physical processes, said Arie Kaufman, chair of Stonybrook University's Computer Science Department.

Other examples are differential equation solvers and scientific computations, such as the fast Fourier transforms used in real-time MPEG video compression and audio rendering, Luebke noted.

"Signal processing operations are usually computationally intensive and data parallel. GPUs' arithmetic capabilities are suitable for them," Kaufman said.

Other suitable applications involve fluid dynamics, including climate modeling, weather prediction, and oceanic and atmospheric studies; and molecular dynamics, including protein and biomolecular simulations, chemical reactions, and material sciences.

GPGPUs also work well on geometric computations such as Voronoi dia-

grams, distance computations, and robot motion planning and collision detection.

In addition, GPUs' high memory bandwidth and parallel processing can accelerate complex database operations including aggregates, predicates, Boolean combinations, selection queries, and data mining, explained Kaufman.

GPGPUs have cracked encryption used for passwords and other purposes. The processors also are effective for nontraditional graphics-related purposes such as medical imaging, ray tracing, photon mapping, and subsurface scattering.

NOT READY FOR PRIME TIME

GPUs need more work before they are ready for more general-purpose uses. For example, noted Nvidia's Purcell, "GPUs aren't good at all types of general computation. They are highly parallel and thus generally aren't good at executing code that is inherently serial."

GPUs are designed to process graphics and thus are more difficult to program for general-purpose computation than CPUs. "In addition, there are few programming tools and little support for general-purpose programming on GPUs," explained the University of North Carolina's Lin.

Also, GPUs' strictly parallel operations tightly constrain their programming environment. This is a particular challenge for programmers used to working with scalar or sequential applications.

According to Nvidia's Harris, programmers must spend time and effort learning to work with APIs, such as OpenGL, designed specifically for computer graphics, whose core concepts are different than APIs typically used in general-purpose computation.

There are also few debuggers or profilers—which track an application's performance by collecting and checking information during code execution—for use in programming GPGPUs, although this is beginning to change, observed AT&T Labs' Venkatasubramanian.

Purcell noted that GPU's can't currently perform arbitrary memory writes.

Vendors have kept some architectural details secret for competitive reasons, Purcell noted. In some cases, this has kept researchers from having access to information that could help explore new general-purpose uses for GPUs.

According to the University of Virginia's Luebke, "The driving market for GPUs is the video game industry, and the needs of that industry dominate the designs and roadmaps of vendors." However, industry observers say, the processors' changing design will also make them more useful for many types of general-purpose computation.

Jon Peddie said that during the next few years, GPGPUs will expand from 24 to 32 pipelines, and each pipeline will include more floating-point processors and larger cache memories.

"You'll also continue to see the speed increases that we've come to expect from GPUs," said Nvidia's Purcell. He predicted that as games start to integrate general-computing techniques, such as those used in physics applications, developers will create better programming models not tied to graphics APIs.

Meanwhile, Purcell said, GPU clusters might prove useful for high-performance general-purpose computing if the problems they work on are sufficiently parallelizable.

"The rapid growth rate and high-performance capabilities of GPUs are very promising for conducting GPGPU research," said the University of North Carolina's Govindaraju. "The challenge, however, will be redesigning traditional CPU-based algorithms to efficiently exploit the computational power of GPUs." ■

David Geer is a freelance technology writer based in Ashtabula, Ohio. Contact him at geercom@alltel.net.

Editor: Lee Garber, *Computer*, l.garber@computer.org