High-Stakes Battle Rages in Graphics-Chip Marketplace

Neal Leavitt

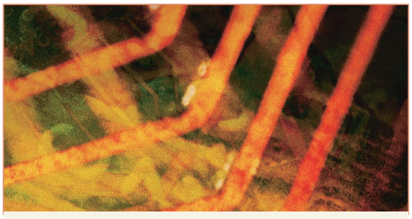
Graphics are becoming ubiquitous, said Tony King-Smith, vice president of marketing for Imagination Technologies, a multimedia and communications system-on-chip firm. "Phones, navigation systems, media players, and TVs are starting to include [graphics] cores," he explained. "In short order, most new devices' designs will include advanced graphics."

In the process, users of these applications are demanding higher performance to make the graphics look better.

Also, over the past few years, graphics processors have been widely used for games and many general-purpose, nongraphics-related applications.

In response, between last year and 2012, market-analysis firm Jon Peddie Research predicts, sales of computer graphics software will rise from $10.8 billion to $15.1 billion; the revenue from the sale of chips with graphics capabilities will grow from $51.1 billion to $57.7 billion; and the number of these processors that are sold will increase, as Figure 1 shows.

Because of these factors, the high-end graphics-chip market has become more important. For years, the market was a two-horse race between ATI Technologies and NVIDIA. But this familiar landscape has changed.



In October 2006, microprocessor maker Advanced Micro Devices (AMD) purchased ATI. Intel, which has made integrated graphics controllers for its chipsets for years, has just entered the fray with its announcement of Larrabee, the company's first stand-alone graphics card, which it plans to release in 12 to 18 months.

The three big vendors are taking divergent approaches.

AMD, for instance, builds smaller graphics chips for mainstream users and puts together two or more of these graphics-processing units for higher-performance uses. NVIDIA builds both smaller graphics chips and large, high-performance GPUs. Both companies make discrete graphics cards with traditional graphics-rendering pipelines.

Intel's Larrabee will use a CPU architecture based on x86 cores, which can be programmed via normal x86 software tools, as well as fixed-function graphics units.

NVIDIA may have an advantage because many people are accustomed to using its chips for high-performance graphics and may be unwilling to try something different, according to Klaus Mueller, assistant professor at Stony Brook University.

AMD spent a lot of money acquiring ATI but has not yet derived a lot of revenue or market buzz from it, he added. "And Intel is catching up with Larrabee, but it's not clear if they will succeed."

BACKGROUNDER

Graphics circuitry has advanced greatly since its inception in the 1960s.

The history

The first graphics chips basically read data from a block of memory and sent it to the monitor, explained Mercury Research principal analyst Dean McCarron.

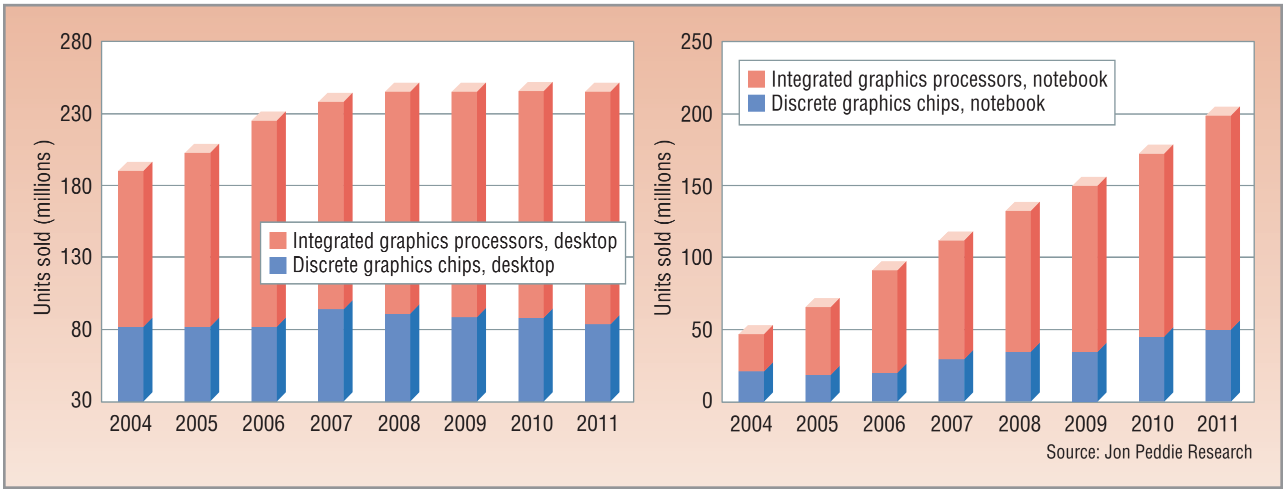
They focused only on converting digital memory into analog signals to drive a display, whereas modern graphics chips handle the entire process of converting a complex 3D scene into a photorealistic depiction, noted University of Illinois at Urbana-Champaign professor John C. Hart.

Chip makers added logic to perform the calculations necessary to manipulate graphics, McCarron said.

Cirrus Logic and Tseng Labs released the first widely used PC graphics accelerators in 1990.

As graphics became more complex and included features such as 3D, the logic portion of graphics chips became much larger and a full processing pipeline was added, McCarron noted.

This process has accelerated as researchers and others have used GPUs and their highly parallel computing capabilities for complex scientific, geometric, and physical simulations, said University of North Carolina professor Dinesh Manocha.



Source: Jon Peddie Research

Figure 1. Sales of chips with integrated graphics capabilities and of freestanding graphics processors will increase during the next few years, according to Jon Peddie Research, a market analysis firm.

Engineering, financial, and other firms, as well as universities, also utilize them for complicated engineering and modeling tasks, including the analysis of geological data to look for the presence of oil or natural gas, the conversion of medical scans into images, and the development of pharmaceuticals.

Today's marketplace

As of mid-2008, Intel had 44.7 percent of the market for PC and laptop chips with graphics capabilities, NVIDIA had 29.7 percent, and AMD had 17.1 percent, according to Jon Peddie, president of Jon Peddie Research. The remainder is divided among several smaller players, including Matrox, Silicon Integrated Systems, and VIA Technologies, he said.

For freestanding PC and laptop GPUs, however, NVIDIA has 63.2 percent of the market, and AMD has 35.4 percent.

AMD'S APPROACH

AMD designs GPUs for every graphics-market segment, said company spokesperson John Taylor. The company uses a modular approach—pairing two or more of its GPUs on the same circuit board—for tasks that require high performance.

The vendor used this strategy for its $549 ATI Radeon HD 4870 X2, released in August, which has two RV770 GPUs on a single card; a core clock speed of 750 MHz; and a total of 1.5 billion transistors, 2 gigabytes of memory, and 2.4 teraflops of processing power. AMD uses PCI Express interconnects to enable the cores to work together.

The company's approach differs from the typical chip-design model of building one large processor for all products, using all of the circuitry for high-performance products, and shutting off portions not needed for mid-range and entry-level products.

According to Taylor, for all but the most powerful processors, buyers pay for a large die and a lot of circuitry they may never use. AMD's approach avoids this and still lets the company save the time and expense of designing separate mid-range and high-end processors.

In addition, Taylor said, two smaller chips on one board use less power and generate less heat than one big processor. AMD also halved the number of bits in its memory interface from 512 to 256 to further reduce power consumption.

To compensate for the resulting performance loss, AMD became the first company to ship products using Graphics Double Data Rate, version 5, high-speed dynamic RAM. This technology uses two parallel data-output links, doubles GDDR4's throughput, and performs real-time error detection and correction.

NVIDIA'S APPROACH

NVIDIA designs fast, large chips to get the highest graphics performance possible from a single processor. The company sells products for markets such as PCs, cell phones, personal media players, and professional workstations.

In June, the company released its flagship GeForce GTX 280 GPU, which sells for $649. The single chip includes 1.4 billion transistors and 240 processor cores with a total clock speed of 1.296 GHz. The memory subsystem provides 142 gigabytes per second of bandwidth over a 512-bit interface, using 1 gigabyte of 1.1-GHz, GDDR3.

Even more powerful versions of GeForce GPUs are included in NVIDIA's Quadro product line for use in markets such as automotive design, oil and gas exploration, medical imaging, and scientific research.

NVIDIA's new GPUs can be used for parallel, nongraphics computation via the firm's Compute Unified Device Architecture, noted Andy Keane, general manager of the company's GPU Computing Group. CUDA is a compiler, programming environment, and toolkit for accessing the GPU features that the driver

exposes. It lets developers write applications in C and leverage the GPU's computational resources.

For some general applications, a computer can use the NVIDIA GPUs for the most complex, easily parallelizable tasks and employ the CPU for others.

INTEL ENTERS THE MARKET

In August, Intel announced its upcoming Larrabee, slated to be the company's first stand-alone graphics card.

"It is not a GPU, as many have mistakenly described it, but it can do most graphics functions," said Peddie.

In Larrabee, Intel is implementing a general-purpose, multicore, x86 CPU architecture that includes special-purpose graphics circuits. The chip implements the graphics pipeline in optimized software running in parallel on the x86 cores.

The cores, based on the same design used in Pentium chips, are essentially 32-bit chips with 64-bit extensions, which make them faster and able to handle more memory. The chips also work with multithreading.

Intel is not publicly discussing most details about Larrabee, including the number of cores the chip will have. Industry observers expect it to have dozens of cores at first and later perhaps hundreds. Intel currently offers quad-core processors and plans to begin producing eight-core chips later this year.

Larrabee will be optimized for graphics processing but can still run nongraphics x86 code. Larrabee won't run ordinary PC applications, though, because it will lack the extensions commonly used in PC software.

However, developers could build applications to run on the chips and program the chips to run PC applications. This is because, unlike GPUs, Larrabee will be fully programmable. A broad range of highly parallel applications, including scientific and engineering software, will benefit from Larrabee's native C/C++ programming model, said Intel Global Communications spokesperson Nick Knupffer.

"Graphics APIs and new graphics algorithms will be easy to develop because central to Larrabee programming is a complete C/C++ compiler that statically compiles [and recompiles] programs to the Larrabee x86 instruction set," he explained. "Application portability could be an enormous productivity gain for developers, especially those working with large legacy x86 code bases."

DOWN THE ROAD

Not surprisingly, there are different opinions as to which approach is best.

NVIDIA has the best single GPU performance in the marketplace, said Nick Stam, the company's director of technical marketing.

AMD's Taylor, on the other hand, said his firm's approach has been tested and proven to deliver leading performance at every price point.

"Intel's approach would offer more flexibility in terms of graphics and nongraphics uses and thus may be more popular with those who use the chips for a range of purposes," said Rob Enderle, president and principal analyst of the Enderle Group, a market-research firm.

AMD and NVIDIA have the advantage of having produced successful graphics chips for years. However, Enderle said, Intel has more money and marketplace clout than both companies combined.

"Selling a few million units may bring Intel significant profit, but it won't mean Larrabee was a good idea from a technical standpoint," stated Peter Glaskowsky, silicon and chip technology analyst for the Envisioneering Group, a market research firm.

Maintaining x86 compatibility may reduce performance and efficiency versus conventional GPUs of the same manufacturing cost, he said. A significant part of each Larrabee core must spend time decoding the complex, lengthy x86 instructions, he explained. This is overhead that consumes a lot of energy but doesn't contribute to rendering, he added.

"Most of Larrabee's theoretical advantages mean nothing to pragmatic gamers," he said, "and I don't expect it to be very successful."

The new graphics chips could change the PC market, according to Enderle. "The lack of good graphics performance is believed to be one of the major reasons that hardware sales have slowed this decade," he explained. "Coupled with compelling visual applications, these [new graphics chips] could drive a resurgence of the PC market."

No matter which approach ultimately proves most successful, the future of graphics chips in general looks bright.

"The number of consumer-level applications that require a powerful 3D visual interface—ranging from virtual navigation of cities to catalogs and libraries with 3D content to games—is rising," explained University of Maryland professor Amitabh Varshney.

Said Enderle, the three fundamentally different graphics-chip approaches will require users to make some interesting choices.

Neal Leavitt is president of Leavitt Communications (www.leavcom.com), a Fallbrook, California-based international marketing communications company with affiliate offices in Brazil, France, Germany, Hong Kong, India, and the UK. He writes frequently on technology topics and can be reached at neal@leavcom.com.

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