Cahners
MICROPROCESSOR REPORT
www.MPRonline.com
THE INSIDER'S GUIDE TO MICROPROCESSOR HARDWARE

WinHEC Sees Great 3D

New Chips, New Software Featured at Windows Hardware Show

By Peter N. Glaskowsky {5/15/00-02}

Microsoft's annual Windows Hardware Engineering Conference (WinHEC) focused this year on the same issues it has focused on over the past few years: ease of use and better graphics. There was one difference this time around, however—OEMs are finally delivering Windows PCs that are actually easier to use and that can display graphics dramatically superior to anything previously seen on the PC platform.

This article describes the new 3D chips and software announced at WinHEC. We'll provide more WinHEC coverage in our next issue.

ATI, Matrox, NVIDIA Show New Graphics Chips

Though ATI has been the most successful company in the PC-graphics industry for the past couple of years, its products have never been the first choice of avid gamers. ATI has been watching NVIDIA's rapid growth with some apprehension, growth due in part to the industry-leading performance of NVIDIA's GeForce 256. ATI's new Radeon 256, announced at WinHEC, is the company's first opportunity to claim 3D-performance leadership.

Unfortunately for ATI, this claim is merely a promise—the Radeon 256 will not ship until June. NVIDIA's new GeForce2 GTS, also unveiled at WinHEC, shipped May 1 from a number of PC OEMs and add-in card makers. As Table 1 shows, the ATI and NVIDIA chips are similar in many ways, combining high-performance rendering engines with transform and lighting acceleration and with HDTV-capable video decoders. Each has unique design details, however, that clearly differentiate the two chips.

Both chips are sure to be very successful and should compete evenly in the midrange and high-end PC market

	ATI Radeon 256	Matrox G450	Nvidia GeForce2 GTS
Bus Interface	PCI/AGP 4x	AGP 4x	PCI/AGP 4x
Memory Width	64/128, DDR/SDR	64 bits, DDR/SDR	64/128, DDR/SDR
Max. Memory Clock	200MHz	Not available	200MHz
Maximum Memory	128M	32M	128M
Geometry Engine	Yes	No	Yes
Core Clock Rate	200MHz	Not available	200MHz
Pixel Pipelines	2	2	4
Multitexturing	3 textures/pixel	No	2 textures/pixel
Shading Engine	Yes	No	Yes
RAMDAC Speed	350MHz	360MHz	350MHz
Max. Res. at 85Hz	2,048 x 1,536	2,048 x 1,536	2,048 x 1,536
MPEG-2 Decoding	iDCT, Motion comp.	No	Motion comp.
Video Outputs	YPrPb, digital	TV out, dual RAMDAC	TV out, digital
Video Inputs	VIP 2.0	VIP 2.0	VIP 2.0
LCD Support	TMDS digital	TMDS digital	TMDS digital
Availability	June	2H00	Now

Table 1. New graphics chips from ATI, Matrox, and NVIDIA have much in common. ATI's geometry-processing and video support is superior, while NVIDIA has the most sophisticated rendering engine. The Matrox chip is little more than a shrink of the company's previous-generation G400.

segments. ATI's Radeon 256 is likely to keep ATI competitive among home and business users who value multimedia capabilities and flexibility over all-out gaming performance. The GeForce2 GTS, on the other hand, will preserve NVIDIA's dominance of 3D gaming and extend the company's reach into the low end of the CAD market.

Matrox's new G450 graphics chip is based on the same 3D core found in its G400 (*see MPR 4/19/99-05*, "3D-Chip Leaders Push the Envelope"). Along with a shrink to 0.18-micron process technology, Matrox gave the G450 an on-chip digital TMDS flat-panel transmitter, an NTSC/PAL video-output encoder, and a second complete RGB output with RAMDAC. The G450 also includes an enhanced local-memory controller that supports 64-bit DDR memory. The DDR interface replaces the G400's 128-bit single-data-rate SDRAM controller. The performance of the new chip matches that of its predecessor.

Geometry Support Accelerates

NVIDIA's GeForce 256, introduced test test test last year (see MPR *9/13/99-msb*, "NVIDIA GeForce Offers Acceleration"), was the first graphics chip to integrate geometry (transform and lighting) acceleration. While the GeForce sold into high-end systems, it was ArtX/ALi's Aladdin 7 that drove geometry acceleration into midrange and low-end machines. (Though S3 announced the mainstream Savage2000 with geometry acceleration at the same time NVIDIA released the original GeForce, S3 has yet to ship a Savage2000 driver that supports the feature) The Radeon 256 gives ATI the ability to meet these competitors head-on. We expect the company to price the new chip at the top end of its product range for the first six months or so, but low-cost versions are sure to be introduced shortly thereafter, which will allow the Radeon 256 to replace the Rage 128.

ATI rates the throughput of the Radeon 256's geometry engine at 30M triangles per second, slightly higher than the 25M triangles/s claimed by NVIDIA for the GeForce2 GTS. The effective performance of the two chips is sure to vary from these numbers, depending on the precise demands of the application software.

Both chips have fixed-function geometry engines. We expect programmable geometry engines to appear in consumer 3D chips later this year. Microsoft is building support for such engines into version 8 of its Direct3D application programming interface (API), scheduled for release this summer. Programmability will give software developers more flexibility in defining 3D models. This flexibility will allow greater variety in the appearance and behavior of 3D games and support more-sophisticated modeling techniques in professional applications.

ATI appears to have included in Radeon 256 a more sophisticated geometry processor than the one NVIDIA designed for the GeForce2 GTS. This accomplishment is all the more impressive considering its slightly higher (claimed) performance and the fact that NVIDIA is now on its second generation of geometry accelerators.

The ATI design, dubbed the Charisma Engine, supports two features not found in NVIDIA's chip—keyframe animation and a four-matrix "skinning" capability. The latter feature produces smooth joints between segments of a flexible model, permitting a more realistic display of human and animal characters. It's possible to perform the same processing steps on the host CPU, but ATI's solution is both faster and easier to program. The GeForce2 GTS supports a two-matrix skinning mode that is less flexible.

ATI's keyframe-animation feature allows the application to define the position of objects at two points in time and have the Radeon 256 generate intermediate positions for each frame in between. This is the same general process used to create classical ink and paint animation.

Rendering Engines Looking Better

The rendering engine in ATI's new chip is greatly improved over its predecessor, the Rage 128. Where the Rage 128 core could draw only one texture per cycle on each of two pixels, the Radeon 256 can apply three textures simultaneously to each of two pixels—and the new chip's clock speed is 200MHz, significantly faster than any of the various Rage 128 speed grades.

For the new chip, ATI has developed a tile-based architecture, one reminiscent of that found in Intel's 740 (*see MPR 2/16/98-01*, "Intel Displays 740 Graphics Chip"), but much more sophisticated. In both designs, the screen is logically divided into tiles, but rendering is done in polygon order—unlike the tile-order rendering in architectures such as PowerVR. The portion of each polygon that touches a tile is rendered completely before the rendering engine moves on to draw the rest of the polygon, and each polygon is fully rendered before moving on to draw the next polygon. This organization increases the locality of references to texture data and the frame buffer, which improves caching efficiency and increases the effective bandwidth of local memory compared with that of more conventional designs.

ATI's "HyperZ" architecture keeps track of the minimum and maximum Z values for each tile and calculates the Z extent of each polygon before beginning to render it. In some cases, these values will indicate that the polygon is further away (has a greater Z distance) than the polygons previously rendered to the tile. As long as the tile is already fully rendered and has no gaps or transparent portions, the rendering engine can skip the tile entirely.

ATI says its testing shows that the HyperZ architecture eliminates some 20% of the possible pixel-drawing operations in typical applications. This reduction leads ATI to claim an effective rendering rate of 1.5 billion texels per second instead of its nominal 1.2-gigatexel rate (200MHz times two pixels times three textures), and an effective local-memory bandwidth of 8GB/s instead of the chip's actual 6.4GB/s peak transfer rate.

Though the HyperZ architecture does increase effective rendering speed, ATI's specific numeric claims are debatable at best. The company could surely have found support for greater or lesser numbers. ATI deserves some credit for choosing these numbers—they are rather conservative when compared with claims made for some other tile-based designs—but in its marketing materials for the Radeon 256, the company often neglects to mention the basis of its estimates, a practice that will lead to some confusion.

NVIDIA's new rendering engine is based on a faster version of that found in the original GeForce design, now running at 200MHz with four parallel pipelines instead of two. As with the previous chip, each pipeline can apply two textures per cycle to each pixel.

Both chips include programmable pixel-shading engines that can perform simple sequences of instructions provided by the application software. Each instruction describes a simple mathematical operation. Operands come from a set of registers containing values previously computed in the pipeline or provided by the application software. Any of these registers may be selected for each operand, and the result may be stored into any register (except for some that contain read-only values).

The NVIDIA part appears to have better peak performance for these shading calculations. In addition to having a slightly more flexible shading unit, the chip has four pixel pipelines, each of which can do about the same amount of work per clock cycle as each of the two pipelines in the ATI chip. The GeForce2 GTS can perform up to 7 shading calculations per pixel, or a total of 28 calculations per clock; the ATI chip offers about half that throughput. If more calculations are required in either chip, additional instructions may be executed over successive clock cycles, with a corresponding reduction in throughput.

Shading algorithms must be strictly sequential, at least in the current generations of these chips. Neither company has implemented any looping or conditional constructs, although future chips are likely to have these capabilities.

Procedural shading, which is supported today by an extension to the OpenGL API and will be supported in DirectX 8, opens the door to a dramatic improvement in the visual quality of 3D scenes. Previous chips can render only objects with uniform surface characteristics, such as a combination of a diffuse (dull) reflection plus a specular (shiny) reflection that appears the same from any angle. A programmable shading engine makes it possible to render complex and variable surfaces such as brushed steel, which reflects light differently, depending on the relative angles of the incident light and the view position. The engine can also be configured to modulate one texture with another, so that only portions of a surface are shiny or textured.

Combined with bump mapping (see *MPR 6/21/99-04*, "A Concise Review of 3D Technology"), a programmable shading engine can duplicate the appearance of complex objects without requiring complex geometry. Naturally, 3D models will continue to become more complex over time; advanced shading techniques will only accelerate this trend.

We expect the industry, in its pursuit of ever better image quality, to focus away from raw polygon and pixel rates and toward greater model and texture sophistication. This quality improvement will require programmable vertex and pixel processors. Before long, each pixel drawn to the screen will be the result of thousands of floating-point calculations.

Techniques, such as ATI's HyperZ, that eliminate the need to draw hidden pixels will become essential—but at the same time, these techniques will account for only a small fraction of the work done in a 3D chip. The Z complexity of 3D scenes will not grow as quickly as effective screen resolution, while model and lighting complexity will increase faster still. Meeting these rapidly changing needs will require a great deal of work by 3D architects and software developers in the coming years. We expect NVIDIA and other 3D-chip companies to introduce features like HyperZ, but we don't expect these features to give any company a significant competitive advantage.

Video Support: Good News for Couch Potatoes

The ATI and NVIDIA chips both include powerful video-processing subsystems that accelerate MPEG-2 decoding. Both can handle MPEG-2 motion compensation, and ATI's chip also includes an inverse discrete-cosine-transform (iDCT) engine. Both chips can handle the resolutions and bit rates required by high-definition TV—about six times faster than those required for DVD playback. ATI says the Radeon 256 is about nine times faster than needed for DVD playback, giving it the ability to decode a second stream at the same time for picture-in-picture display. A PC equipped with these chips lacks only an HDTV-compatible tuner board to receive and display high-definition video.

Both companies tout their chips' ability to handle what they call motion-assisted de-interlacing. This technique improves the display of interlaced video (such as that found on DVDs and in some HDTV broadcasts) on non-interlaced (progressive-scan) computer monitors; it uses information from the motion-compensation task to select between two different video filters when displaying each part of the image. The result on screen can be superior to all but the very best consumer televisions.

ATI goes one step further, including in the Radeon 256 a hardware unit that looks for interlacing artifacts in video windows, even when motion vectors are not available—for example, when the video comes from an analog source. The unit identifies artifacts, such as feathered edges, that can be created during the de-interlacing process and applies a suitable video filter to each portion of the screen as needed.

One relatively simple feature of ATI's Radeon 256 seems, in retrospect, overdue. Though mainstream graphics chips have for years supported TV-video outputs, through off-chip as well as integrated RGB-to-NTSC, -PAL, and

SECAM encoders, the Radeon 256 is the first chip to support the new generation of component-video TV monitors. The ATI chip's RAMDAC, supported by on-chip color-space conversion hardware, can be programmed to generate the Y-Pr-Pb component signals used by these monitors instead of the normal RGB analog signals.

This is a relatively simple feature, but one that could benefit customers who own, or will soon buy, high-definition televisions. These TVs usually have component-video inputs capable of matching the resolution of common computer displays—1,280 x 768 pixels with 60Hz progressive-scan refresh or 1,920 x 1,080 pixels with 30Hz interlaced scanning. The price of high-definition televisions is currently comparable to RGB-input big-screen monitors, such as those offered by Mitsubishi and others, but HDTVs tend to have larger screens, and their prices are likely to drop rapidly as they become more popular.

New Chips Will Hasten Market Consolidation

The new chips from ATI and NVIDIA, taken together with these companies' current products, leave little room for competition. There will always be smaller companies with proprietary technology or a unique market focus looking to survive as niche players, but we expect ATI and NVIDIA to dominate 3D-chip sales for at least the next year.

ATI sells board-level products as well as discrete chips, both mostly to PC OEM customers. The company's board business gives it higher revenues than NVIDIA, which sells only chips. NVIDIA has higher margins, however, and the two companies are roughly evenly matched in R&D spending.

The growth of these two companies has put tremendous pressure on several vendors that once held respectable market shares, notably 3dfx, Matrox, and S3. The Voodoo 4 and 5 families from 3dfx are only now reaching store shelves, some six months after their Comdex introductions (see *MPR 12/6/99-02*, "Napalm Ignites Graphics Market"). The new Voodoo boards will do well in the retail channel as upgrade products, but they are unlikely to have substantial sales to OEMs. The future of 3dfx, even with the infusion of talent and intellectual property from Gigapixel (see *MPR 04/03/00-10*, "GigaPixel Bought by 3dfx"), remains uncertain.

Matrox's G450 will do little to reverse the company's slow decline as ATI and NVIDIA take away more of its OEM and retail business. Though the G400 was quite competitive against ATI's Rage 128 and NVIDIA's TNT2, the G450 looks old and slow next to today's alternatives.

S3's graphics-chip group, recently spun off to create a joint venture with VIA (see *MPR 5/1/00-03*, "S3, VIA Split Out Graphics Group"), has yet to introduce a successor to the Savage2000, apart from the ProSavage PM133, an integrated-graphics chip set previewed at WinHEC.

The GeForce2 GTS is likely to outsell the Radeon 256 among gamers, if only because of the reputation of NVIDIA's earlier products. The delayed availability of the ATI chip probably won't have much effect on its overall sales; ATI's customer base has always seemed remarkably patient with the company.

The relative performance of the two chips and their slightly different feature sets will be the least important factors in their success. Speed records never last long in this industry, and developers are reluctant to support new hardware features until they become widely adopted. The factors that will really matter are the ones that always matter—meeting price and production commitments, delivering quality driver code, and building consumer brand awareness. ATI and NVIDIA are fairly evenly matched in these areas, and both are sure to experience strong demand in the coming year. ◇

Price & Availability

ATI's Radeon 256 is sold in add-in board products through OEM and retail channels and as a discrete device to some OEMs. The company has not released pricing information for the new chip but says products based on the Radeon 256 will be available in June.

Matrox sells its graphics chips on its own graphics cards only, and does not release chip-level pricing. Matrox says G450-based graphics cards will ship in the second half of 2000.

NVIDIA has been shipping the GeForce2 GTS to its add-in-board and OEM partners for about a month, and products from several companies are available now. NVIDIA also has not released chip-level pricing.