1997 COMPUTER PRESS ASSOCIATION WINNER ♦ BEST COMPUTER NEWSLETTER

MICROPROCESSOR REPORT

VOLUME 12, NUMBER 10
AUGUST 3, 1998

THE INSIDERS' GUIDE TO MICROPROCESSOR HARDWARE

3D Vendors Aim High

3Dlabs, Nvidia, Number Nine Expand Performance Envelope

by Peter N. Glaskowsky

Professional users demand more features and performance than the gaming-oriented consumers for whom the chips described in our previous 3D article (see MPR 7/13/98, p. 16) were designed. New 3D engines from 3Dlabs, Nvidia, and Number Nine are meant to satisfy these demands.

The 3Dlabs Permedia 3 and Nvidia's new RIVA TNT are closely matched in performance and should meet the needs of 3D artists, game developers, and the most avid game players. Number Nine's Ticket To Ride 4, the sequel to last year's Ticket To Ride, is aimed at high-end business desktops and applications that require exceptionally high resolution displays. Table 1 compares these three parts.

For Permedia, the Third Time's a Charm

When it introduced the Permedia 2 in 1997 (see MPR 6/2/97, p. 16), 3Dlabs hoped to dominate the consumer 3D business the way its Glint chips have dominated the high end of the market (see MPR 8/3/98, p. 5). Unfortunately, the 40-Mpixel/s P2 simply wasn't fast enough to compete effectively with Nvidia's 100-Mpixel/s RIVA 128. Because Permedia 2 came with high-quality OpenGL drivers not available with most consumer-grade 3D chips, 3Dlabs achieved some design wins for business PCs and low-end CAD systems, but these couldn't match the volume potential of mainstream OEM sales.

With the much-improved Permedia 3 (P3), 3Dlabs has solved the performance problem, providing increased competition for 3D-gaming market leaders 3Dfx and Nvidia.

Like 3Dfx's Voodoo2, the Permedia 3 is equipped with two texture-mapping units that work together on one pixel at a time, as Figure 1 shows. Each texture unit can perform bilinear filtering on two separate textures, one of which may be a bump map, to be applied to the same pixel—a technique known as multitexturing—or they can work together to apply a single texture, using a trilinear filtering algorithm at the same pixel rate.

The 125-MHz P3 includes 2D and video features, making it a better feature match to 3Dfx's Voodoo Banshee, also a single-chip solution. Banshee has only one texture unit and runs at just 100 MHz, though a 125-MHz Banshee is due later this year.

For trilinear-filtered textures, the Permedia 3 is more than six times faster than the Permedia 2 and matches the pixel throughput of S3's Savage3D or Nvidia's RIVA TNT. The P3 also matches the TNT's throughput on bilinear-filtered dual-textured polygons. The one mode where the TNT excels is in bilinear filtering without multitexturing, where its peak throughput is doubled to 250 Mpixels/s. The trend toward multitexturing on new 3D games, both for game play and game development, 3Dlabs believes is strong enough to reduce the significance of Nvidia's advantage, though older games are sure to run faster on TNT. (Perhaps faster than necessary—some of these titles already achieve frame rates exceeding 100 Hz on Voodoo2.)

Fast Memory Supports Fast Engine

The Permedia 3's 128-bit memory interface supports up to 16M of SDRAM or SGRAM. The memory controller is designed to operate at speeds up to 200 MHz. Today, only MoSys makes SGRAMs at this speed (see MPR 8/3/98, p. 9),

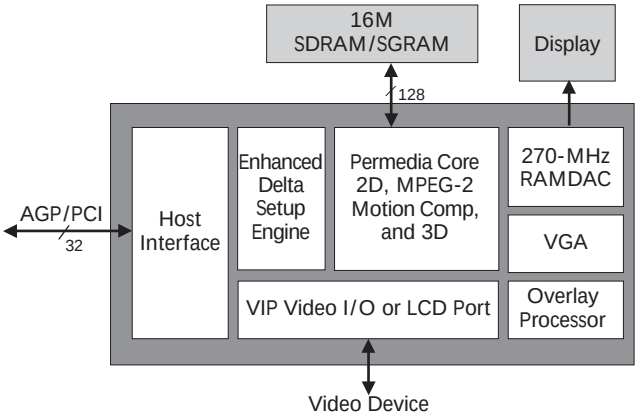

Figure 1. 3Dlabs' Permedia 3 looks much like its predecessor but includes dual texturing pipelines and much faster operation.

and 3Dlabs expects most customers will use 125-MHz memory chips. Memory timing is independent of the core clock rate, providing more flexibility for OEM customers. The P3 uses SGRAM's block-write and write-per-bit functions if available, emulating these functions in hardware when SDRAM is used. This trick, also used by 3Dfx in its Voodoo Banshee, allows the use of the same drivers for either memory configuration, and it eliminates the need to write and support different drivers for each configuration.

To reduce demand on the local memory, the P3 includes 8K of on-chip texture cache, a typical amount for recent 3D chips. The Permedia 3, however, is the first to use its local memory as a second-level cache for texture accesses to main memory. Unlike other chips that require applications to move textures explicitly from main memory to local texture storage, the Permedia 3's programming model allows applications to treat main memory as the only software-visible texture-storage location.

As textures are requested during the rendering process, the P3 automatically transfers them from main memory to its local memory and on-chip texture cache. Even when new textures bump older ones from the on-chip cache, the P3 is likely to keep the older texture data in its local SGRAM to satisfy any repeated requests for the same data. This gives the P3 the software benefits of AGP-only texturing, as found in the Intel740, plus the reduced AGP usage of chips such as 3Dfx's Voodoo that move textures across the bus only once before using them multiple times in the same scene.

The texture page size is 4K bytes, enough to store a 32 × 32-texel fragment of a texture map. The P3 also allows explicit texture management on a per-texture basis, so applications can either move an entire texture to the P3's local memory before use or load a texture from system memory without local caching. The latter feature is necessary for video and other texture maps that change rapidly, since P3 offers no mechanism for cache coherency with system memory.

	3Dlabs Permedia 3	Nvidia RIVA TNT	Number Nine T2R4
Bus Interface	AGP 2×	AGP 2×	AGP 2×
Local Memory Type	SD/SGRAM	SD/SGRAM	SD/SG/WRAM
Memory Width	128 bits	128 bits	128 bits
Memory Clock Rate	125 MHz	200 MHz	100 MHz
Maximum Memory	16M	16M	32M
Texture Cache	8K	12K	8K
Core Clock Rate	125 MHz × 2	125 MHz × 2	100 MHz
Peak Bilinear Rate	125M	250M	n/a
Peak Trilinear Rate	125M	125M	n/a
Peak Triangle Rate	8M	8M	n/a
RAMDAC Speed	270 MHz	250 MHz	250 MHz
Video I/O Ports	In/Out	In/Out	None
LCD Support	Yes	No	Yes
Availability	4Q98	3Q98	3Q98
Volume Price	$45	$45	$65

Table 1. Permedia 3 and the RIVA TNT will contest high end content-creation and gaming. Number Nine remains focused on business desktops. n/a: not available. (Source: vendors)

While the Permedia 3 supports texture compression, its algorithm is not compatible with the S3 method licensed by Microsoft for DirectX 6.0. Instead, 3Dlabs offers a lossless run-length encoding (RLE) technique that should be a better fit for CAD, where the demand for visual accuracy may preclude the use of lossy compression algorithms such as S3's. RLE-compressed textures can still be accessed via version 6.0 of Microsoft's Direct3D API, but such textures will require specific support from each application.

CRT and Flat-Panel Monitors Supported

For 2D graphics, the Permedia 3's 270-MHz RAMDAC can drive an RGB monitor at resolutions up to 1,920 × 1,200 pixels in true color with a 90-Hz refresh rate. The P3 can render 3D graphics in 16-bit color with a 16-bit Z buffer at a resolution of 1,600 × 1,200 pixels, allowing supersampled antialiasing for 3D displays at 800 × 600 pixels. True-color rendering and a 24- or 32-bit Z buffer are also supported at lower resolutions. The P3 also provides the necessary logic to connect to LCD glasses for stereoscopic 3D.

The P3 can drive a flat-panel display at speeds up to 135 MHz through a 24-bit digital RGB output port. This port may be configured as a bidirectional VIP interface or to receive Zoom Video from a PC Card port in notebook applications. The chip also includes a genlock function to synchronize graphics and video streams. Video-decoding performance is further enhanced by the inclusion of MPEG-2 motion-compensation circuitry, which reduces CPU loading and system power consumption during DVD playback.

Another notebook-oriented feature is a flexible power-management controller that can shut down each major function in the P3 independently of the others, allowing video playthrough with no power consumed in the 3D unit, or vice versa. Though the Permedia 3 consumes significantly more power than popular mobile-graphics accelerators such as NeoMagic's MagicGraph 256v and its performance would be overkill for most laptops, there is a market for high-end notebooks with 3D acceleration and DVD playback. The P3 is well-suited to such systems.

With its support for multitexturing, bump mapping, and full-screen antialiasing, the P3 is well adapted for the Direct3D 6.0 API. Without support for anisotropic texturing, however, it may lose some ground to the TNT among designers and users of 3D games that utilize that feature. However, 3Dlabs will continue to provide OpenGL drivers for the P3, serving its workstation customers. The chip includes direct hardware support for both D3D and OpenGL polygon-vertex formats plus vertex arrays, eliminating any need for its drivers to translate from one format to another. Such translation causes significant performance penalties for less-flexible 3D chips.

The asking price for the Permedia 3 will be $45 when it ships in the fourth quarter. As if to indicate that this is the real price of the chip, rather than the often highly negotiable prices quoted by other vendors, 3Dlabs says the $45 price will

require volume purchases of 100,000 units per month. Unlike Permedia 2, the new chip will not be available from alternate sources; only 3Dlabs will be selling it. Buyers will include Diamond, Elsa, STB, and other current 3Dlabs customers. Some new OEM accounts may find Permedia 3 more interesting than 3Dlabs' previous products.

Nvidia's TNT Is Da Bomb

The new RIVA TNT's 250 Mpixels/s of peak rendering performance is unmatched by any other single-chip engine. Even 3Dfx's Voodoo2 chip set, when used in pairs via 3Dfx's unique scan-line interleaved mode, offers only 180 Mpixels/s for single-textured polygons, despite having six ASICs and six 64-bit local memory arrays. Interleaved Voodoo2 cards still have an advantage for dual-textured polygons, where they can sustain the same 180-Mpixel/s rate, compared with the TNT's reduced 125-Mpixel/s throughput in this mode.

Figure 2 shows the TNT's single-chip design with one 128-bit memory bus. With seven million transistors in a 0.35-micron process, the TNT is easily the most complex mainstream 3D chip yet announced. The new chip can drive 16M of SGRAM or SDRAM at speeds up to 200 MHz, though Nvidia has suggested that the earliest TNT chips may not support the 200-MHz rate. Nvidia plans to move the TNT quickly to 0.25-micron technology, ensuring 200-MHz memory support; the new process could also boost the TNT's initial 125-MHz core clock rate.

Polygon Processing Still CPU-Limited

While both the TNT and the P3 claim a peak polygon-processing rate of 8 Mtriangles/s, application-level performance will be much lower. Both chips have improved 3D setup engines to reduce CPU loading. Their vendors hope to catch up to 3Dfx's Voodoo2, which currently offers up to three times the peak polygon rate of other chips, due to its industry-leading setup efficiency.

Even with the streamlined Voodoo2 drivers, a 400-MHz Pentium II can only process 1.7 million polygons per second. Faster CPUs can perform geometry calculations for more polygons, but over the life of the TNT and P3, Intel's fastest CPUs are unlikely to exceed about 3 Mtriangles/s of peak performance. Nvidia's claimed 10-GFLOPS performance from the TNT's setup engine might appear to be 40× faster than the P3's nominal 250-MFLOPS throughput, but the similar claims for polygon throughput suggest that these FLOPS ratings are not directly comparable.

With an integrated 250-MHz RAMDAC, the TNT falls slightly short of the maximum display resolution of the Permedia 3, but the TNT can still drive a 1,600 × 1,200-pixel monitor at a refresh rate of 85 Hz, more than enough for most users.

Reflecting Nvidia's interest in the consumer market, the RIVA TNT includes a video I/O interface, color-space conversion logic, and other features to enable and accelerate digital-video operations. The chip lacks the motion-compensation logic found in the Permedia 3. However, 350-MHz and faster Pentium II processors capable of software-only DVD decoding will be moving into the mainstream over the next year. This advance will reduce the importance of DVD motion compensation hardware.

Like 3Dlabs, Nvidia now provides both Direct3D and OpenGL drivers for its chips. The RIVA TNT is likely to compete with the P3 for PC workstation designs, where OpenGL compliance is mandatory and applications can use as much performance as any chip can deliver.

The TNT's performance and features will not come cheaply. The chip will sell for $45, the same price as Permedia 3, though in more moderate 10,000-unit quantities. TNT-based graphics cards are planned by STB and Elsa among others, and we expect these cards to sell for $250 or more, well above typical 3D-card prices of $149 or less. Nvidia's RIVA 128ZX provides a less expensive alternative for cost-sensitive OEMs, but as the 128ZX becomes effectively obsolete, Nvidia will have to cut the price of the RIVA TNT to more mainstream levels.

Number Nine's Number Four Shows Improvement

Perhaps the most perplexing product naming scheme in the 3D market today comes from Number Nine Visual Technology. While we thought it was odd for PowerVR to describe its third-generation chips as the PowerVR Second Generation family, Number Nine has named its second Ticket To Ride chip the Ticket To Ride 4 (T2R4), because it is the company's fourth 128-bit graphics accelerator and the fourth to be used in the company's Revolution series of graphics cards.

With a name like that, it has to be good. Fortunately, Number Nine has made significant improvements to the T2R, producing a chip that should be even more popular among Number Nine's customer base.

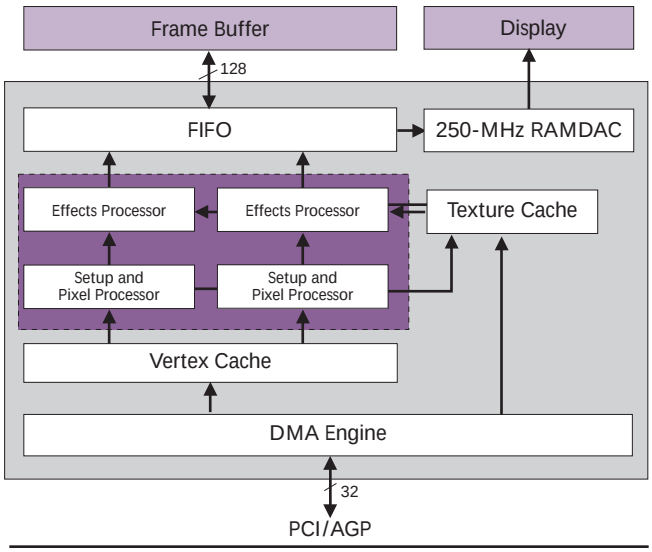

Figure 2. Nvidia's RIVA TNT also has dual rendering pipelines, but unlike the P3 the TNT can render two pixels at the same time.

Like Matrox and 3Dfx, Number Nine claims its new chip has the fastest 2D engine in the world. It would probably be fair to say that all of the new graphics chips we've seen this year are fast enough for any 2D application. The 128-bit VGA core in the Voodoo Banshee appears to have a substantial lead over the others for old DOS games, but this merely confirms 3Dfx's dominance in the game market.

Number Nine cares more about professional customers, and the T2R4 was clearly designed to meet their needs. Like the other chips covered here, the T2R4 is a 128-bit design. It includes a 2× AGP interface that is also PCI-compatible, and, unlike its 1× AGP predecessor, the T2R4 supports AGP sideband addressing, pipelined transfers, and AGP execute-mode texturing. Figure 3 shows the T2R4's unique 128-bit VLIW display-list processor, also found in the T2R. This processor is controlled by the T2R4's driver code, reducing CPU loading for data-transfer tasks.

The T2R4 supports a 128-bit array of up to 32M of 100-MHz SDRAM, SGRAM, or WRAM memory. The unique WRAM option, implemented as two interleaved 128-bit arrays, permits slightly higher effective bandwidth than SGRAM, since WRAMs provide a separate port for screen-refresh transfers. In this configuration, the T2R4 must be used with an external RAMDAC, but this provides the key to the T2R's success in vertical-market applications such as air traffic control. A board-level OEM can combine large amounts of memory with a fast discrete RAMDAC to support resolutions up to 4,096 × 4,096 pixels, well beyond the capabilities of any other PC graphics chip.

For mainstream applications, customers will typically use SGRAM memory or less-expensive SDRAM. Number Nine says the choice of SDRAM typically reduces performance by only about 2% while permitting a significant reduction in the parts cost of a T2R4-based graphics card.

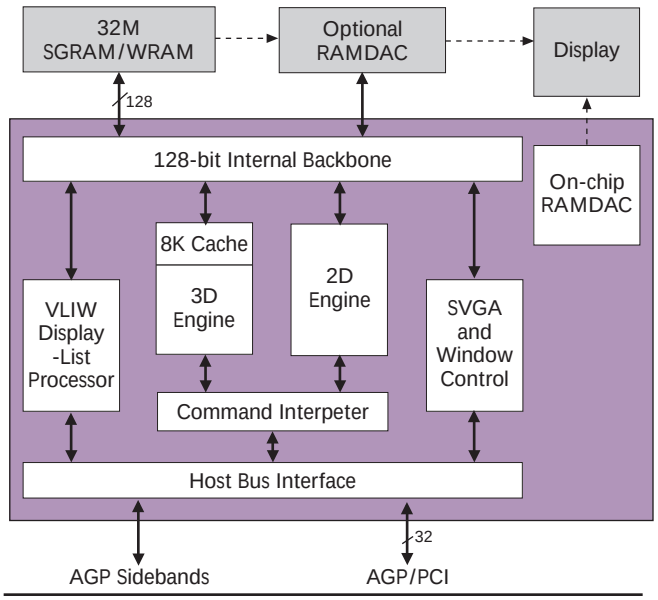

Figure 3. Number Nine's Ticket To Ride 4 is much like its predecessor but offers improved quality and faster operation.

Pricing and Availability

3Dlabs' Permedia 3 is due to ship in the fourth quarter, with pricing set at $45 in quantities of 100,000 units per month. The chip will be packaged in a 456-pin BGA package. More information is available from *www.3dlabs.com*.

The RIVA TNT is due in the third quarter from Nvidia (*www.nvidia.com*). Nvidia's chip will also cost $45, but in more moderate quantities of 10,000 units, and it will come in a 452-pin BGA package.

Number Nine asks $65 in quantity for the Ticket To Ride 4, which is also scheduled for third-quarter production. The chip is packaged in a 388-pin BGA. Number Nine (*www.nine.com*) sells chips to motherboard makers and vertical-market customers only, preferring to sell its own cards to mainstream OEMs and end users.

While the T2R was only a middling-fast 3D accelerator, the T2R4's faster AGP interface and other improved features should make it much faster than its predecessor. Number Nine has not released samples of the chip for testing but claims to have seen scores in excess of 1,110 3D WinMarks on a 400-MHz Pentium II system running DirectX 6. This is somewhat less than what we expect to see from Savage3D and Voodoo Banshee and well below the anticipated scores of the Permedia 3 and RIVA TNT.

While the company has not provided details, Number Nine says the T2R4 will include a digital interface for flat-panel displays at speeds above 135 MHz. Some vertical-market applications use PCs with multiple flat-panel monitors, making them an attractive target for Number Nine.

Professional Choices Proliferate

These three 3D chips are likely to be the most popular choices for professional graphics applications. Number Nine's Ticket To Ride 4 will be the preferred choice for customers who need support for unusually high resolution monitors along with excellent 2D and 3D performance.

For more mainstream buyers using commodity displays, both 3Dlabs' Permedia 3 and Nvidia's RIVA TNT will provide even better performance and more advanced features. The Permedia 3 will be more popular among CAD-oriented customers because of 3Dlabs' greater experience with OpenGL acceleration, and the P3 is better positioned to take advantage of the trend toward flat-panel displays. The RIVA TNT, with its unmatched 3D performance, should be a hit with game developers as well as game players. M