Nvidia Launches Multimedia Accelerator

First Chip to Combine Audio, Video, and 3D Graphics Acceleration

by Linley Gwennap

With its first product announcement, Nvidia has created not just a new device but a new product category: multimedia accelerator chips. The aptly named NV1 combines an innovative 3D graphics accelerator with a 350-MOPS DSP for multichannel audio synthesis. In addition, it provides standard PC capabilities such as GUI acceleration and Sound Blaster audio, eliminating the need for other graphics or audio chips in the system. The $55 chip can be installed on the motherboard or supplied on an add-in card that retails for about $200.

While maintaining compatibility with defacto PC standards, the Nvidia design delivers a significant boost in multimedia capability. The graphics accelerator generates realistic 3D images, comparable to emerging high-end game machines like the Sega Saturn *(see **090704.PDF**)*, at up to 60 frames per second. The audio DSP can generate up to 32 concurrent channels of CD-quality sound, mixing audio channels even if they were created using different sample rates. Either of these functions requires expensive add-in cards today.

The 40-person company has allied with giant SGS-Thomson for manufacturing and marketing capability. Although Nvidia will sell the part as well, it will mainly concentrate on design, and it already has plans for several iterations of its initial product.

3D Accelerator Based on Curves

Most 3D graphics applications are based on simple polygons. A polygon consists of three or more vertices, which are points in space, connected by straight lines. These simple polygons define a flat surface with the vertices in a plane; thus, they can be represented by small data structures and manipulated straightforwardly.

Unfortunately, the real world is round. A curved surface must be approximated by a number of interlocking polygons. If the number is small, the faceted surface appears distorted. A larger number of polygons creates the illusion of smooth curvature, but this illusion may require hundreds of polygons for a simple sphere. Manipulating a "sphere" that consists of so many polygons requires substantial CPU and memory bandwidth.

Nvidia's chip uses a version of a mathematical algorithm known as NURBS (nonuniform rational B-splines). Put simply, it represents objects as a set of curves rather than lines. The basic unit is a complex polygon in which the area between the vertices can have an arbitrary curvature. But many real-world objects can be represented by a small number of these curved polygons rather than a larger number of traditional polygons.

Given a set of vertices and curvatures, the NV1 will generate the desired curved surface. This computation frees the CPU from such effort, allowing the processor to manipulate a curved surface as easily as a simple polygon. Many objects can be represented with far fewer curves than with flat polygons, freeing CPU resources to add lighting or other special effects, or to increase the frame rate. Alternatively, for objects that are poorly represented using flat polygons, the NV1 can provide more realistic representations with the same CPU effort.

Graphics Engine Designed for Games

The Nvidia design is also unusual in its emphasis on frame rate rather than precision. Traditional 3D graphics has been used for mechanical design and similar tasks where spatial location must be maintained very precisely. Nvidia instead focuses on games and other multimedia applications, where image is everything. Internal calculations are made to an accuracy of ±1 pixel, but no more. Discarding visually irrelevant data allows these calculations to proceed more quickly. The chip also dispenses with Z-buffering, a mainstay of CAD engines, avoiding the cost of the extra buffer memory.

The chip uses its 3D graphics pipeline for 2D acceleration as well. It accelerates common graphics functions such as raster operations and line drawing while providing video functions such as pixel expansion and texture mapping. By combining these effects, the graphics engine can map a moving video image onto a curved shape. Video of an actor's face, for example, can be wrapped around a sphere to create a realistic 3D image of a character speaking.

Nvidia declined to provide benchmark measurements of its chip in either 2D or 3D modes. As a GUI accelerator, the chip is claimed to outperform all others in 16-bit color configurations, although the company admits that, with 8-bit color, performance is merely comparable to that of high-end GUI chips. For 3D graphics, benchmarks that measure polygons per second are not adequate, as Nvidia's polygons are far more complex than other vendors'. The chip produces images that are comparable visually to those from a Sega Saturn or Sony Playstation.

Powerful Audio Engine

The NV1 is designed to connect directly to either VL-Bus or PCI, but it performs best with a PCI connec-

tion. Unlike most graphics and audio chips, it can be a master on the PCI bus as well as a slave. The chip has its own DMA engine, allowing it to autonomously load data from memory.

This feature is handy for wave-table audio, which uses prerecorded sound samples to synthesize new sounds. Using DMA, the Nvidia chip can load a number of samples from memory, feeding its internal DSP. The CPU simply sets up the DMA chain and lets the NV1 go to work, freeing the CPU to handle other processing. This method also stores the wave samples in main memory or frame-buffer memory, eliminating the cost of the dedicated storage used in other audio products. The company will license its Nvision wave samples and will provide DirectSound drivers for Windows 95.

For compatibility with current audio applications, Nvidia provides Sound Blaster emulation. The NV1 watches the PCI bus for Sound Blaster register accesses, then executes them using its audio DSP. This configuration delivers basic audio capability to the end user while providing extended capabilities to newer software.

The DSP's audio algorithms are hardcoded into simple state machines. This decision significantly reduced the size of the DSP by eliminating costly instruction ROM. But by carving the algorithms in silicon, the company is gambling that there will not be significant changes during the lifetime of the chip.

Nvidia says the DSP can handle 32 simultaneous channels of 16-bit sound. It supports concurrent audio input and output at sample rates up to 48 kHz. The chip can also create effects such as tremolo, vibrato, and phase-shifting (for 3D sound). The audio engine connects to a standard codec to drive stereo sound channels.

The powerful DSP seems overengineered by the standards of today's applications, but Nvidia wants to deliver a product that will continue to outperform NSP-based solutions as CPUs get faster. With Native Audio, a 100-MHz Pentium can support audio mixing of eight channels or concurrent audio at 8 kHz without significantly degrading CPU performance *(see **090603.PDF**)*. Nvidia expects its DSP to provide audio functions beyond the capabilities of faster Pentiums and even P6 processors.

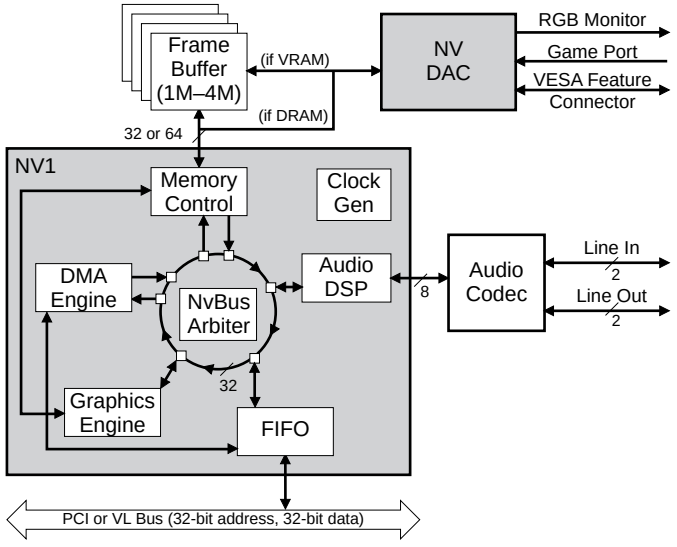

Figure 1. The NV1 contains complete audio and video subsystems and connects directly to PCI or VL-Bus.

Internal Ring Bus Allows Concurrency

As Figure 1 shows, the chip's internal units are connected in a unidirectional ring by a 32-bit bus. Each node can either send, receive, or stay off the bus, allowing data to pass transparently to the next node. A central arbiter controls the bus, accepting transaction requests and instructing each node. For example, to move data from the bus-interface FIFO to the memory controller, the graphics unit and DMA engine are set to pass-through, and the data flows directly from source to destination.

Nvidia considered a traditional bus in which all units have equal access to the bus. The ring structure, however, allows concurrent transfers in some situations. If the graphics engine is receiving data from the bus, for example, the DMA engine can simultaneously send data to the DSP using the other sections of the bus. Up to three transfers can occur at once, yielding a peak bandwidth of 600 Mbytes/s at 50 MHz.

The blocks are arranged so nodes that frequently communicate with each other are adjacent. The chip also has two side buses that directly connect the DMA engine to the bus and the graphics unit to the frame buffer, as these are frequently used paths.

The memory interface supports a frame buffer of up to 4M using either DRAM or VRAM. High-performance configurations can use a 64-bit-wide frame buffer, but a 32-bit width is supported for lower cost. A second custom chip, the NVDAC, contains a 135-MHz RAMDAC to drive a standard CRT. It also includes a game port connection. A joystick connected through this port uses the NV1's bus-mastering capability to interrupt the CPU when needed, reducing polling overhead and greatly improving response time in many applications.

Both Nvidia and SGS-Thomson sell the NVDAC in a set with the accelerator chip. Nvidia plans to combine the two chips into a single part in future products.

Partnership with SGS-Thomson

Instead of a traditional foundry relationship, Nvidia chose to grant SGS-Thomson marketing rights to its design. In a world of overbooked fabs, low-profit foundry wafers are at the bottom of the list, but wafers for internal products get higher priority. To keep from losing sales to its larger partner, the two-year-old startup concocted a clever market segmentation: SGS sells the STG2000, which connects to a DRAM frame buffer,

while Nvidia sells VRAM-based parts. The DRAM parts have a lower price but a higher volume; Nvidia skims the low-volume, high-profit segment of the market.

The initial product is built in a 0.5-micron three-layer-metal process. This advanced process helps cram 1.0 million transistors onto a single chip in a 208-pin PQFP. The first parts use a gate-array design to reduce time to market; these chips measure 121 mm^2. To reduce cost to a more reasonable level, Nvidia will move quickly to a semicustom design. Nvidia is also working on a second-generation design that uses a 0.35-micron process for greater integration and higher performance.

Software Support Is the Key

While the NV1 is a bit pricey, it is competitive with the price of a high-end GUI accelerator plus a sound chip, and the Nvidia design can match their capabilities and offer far more. For relatively little incremental cost, Nvidia can turn an off-the-shelf PC into a world-class game machine.

To create end-user demand, optimized software is required. Nvidia hopes to convince game writers to target Nvidia-enabled PCs by offering a unified audio and video platform. But to support the NURBS model, these programmers must rewrite their source code. Nvidia claims that this rework is relatively simple.

Many software vendors, however, are drawn to the new 3D game interfaces in Windows 95 and NT. These APIs allow vendors to write a single piece of code that will run on many 3D graphics chips. These standards, however, are for simple polygons, defeating the Nvidia enhancements.

The NV1 also leaves out a key video standard: MPEG-1. Nvidia could not fit a complete MPEG-1 decoder into its initial product, but this standard is gaining momentum in PCs. S3, for example, provides audio, video, and MPEG with its latest chip set *(see **0909MSB.PDF**)*.

Nvidia's first product has some impressive capabilities. For this product to flourish, the company must convince software vendors to rewrite their code to take advantage of these capabilities, even as Microsoft, along with other chip vendors, promotes a different standard. The NV1 appears far enough ahead of its competition to convince some software vendors to support it; so far, Domark, Interplay, Papyrus, and Velocity have signed up. Nvidia must strive to maintain its technology leadership as others introduce similar products. ♦

Price & Availability

Both the NV1 (VRAM version) and the STG2000 (DRAM version) are now sampling with production expected in 3Q95. Nvidia is selling the NV1 in a set with the NVDAC for $70 in OEM quantities. The STG2000 and NVDAC set, from SGS-Thomson, is priced at $55 in OEM quantities. Both prices include software drivers.

Contact Nvidia (Sunnyvale, Calif.) at 408.720.6100; fax 408.720.6111. Contact SGS-Thomson (Carrollton, Texas) at 214.466.7644; fax 214.466.6572.

