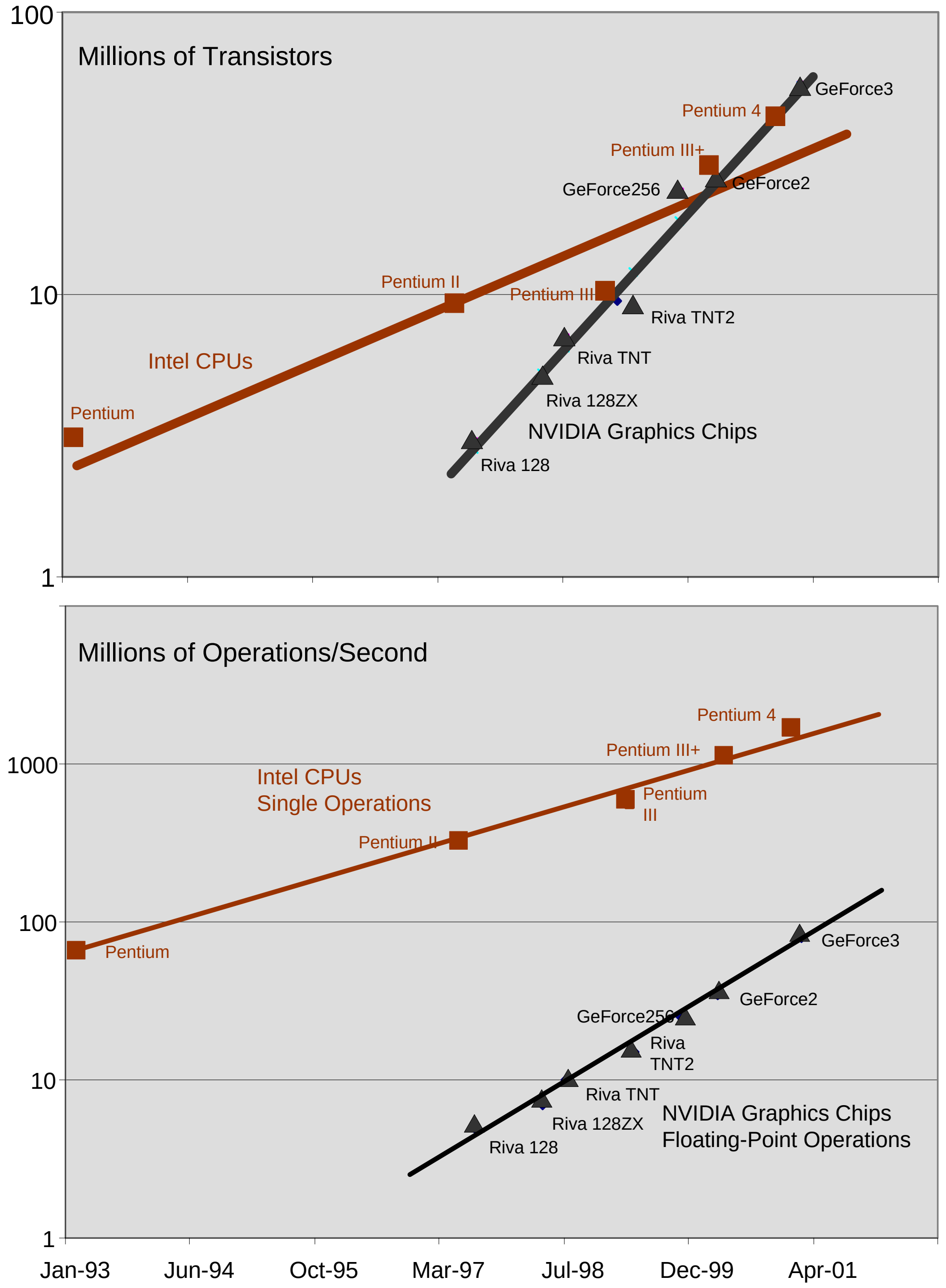
UCLA Anderson

NVIDIA Corporation, 2001

Exhibit 1—GPUs Advance More Rapidly Than CPUs



This case was prepared by Professor Richard P. Rumelt with the research assistance of Gregory Rose, UCLA Anderson M.B.A., 2004 and Oliver Costa, UCLA Anderson MBA Class of 2003. This case was prepared using information provided by NVIDIA Corporation, as well as public sources, to serve as the basis for class discussion rather than to illustrate either effective or ineffective management.

NVIDIA[1] was formed in April, 1993, by Jen-Hsun Huang, Curtis Priem, and Chris Malachowsky. Huang had been an engineer-manager at LSI Logic; Priem and Malachowsky had been CTO and VP for Hardware Engineering respectively, at Sun Microsystems. The founders believed that the time was ripe for the introduction of higher-quality 3D graphics chips in the PC market.

The early years were not auspicious. NVIDIA's first two products failed and start-up competitor 3Dfx Interactive came to dominate the 3D graphics market with a 1998 share of 85%. To make matters worse, rapid growth in 3D graphics processors brought industry giant Intel into the fray.

In 1995 over twenty firms designed and produced graphics chips for the PC. By the end of 2001 there were essentially only three leaders: NVIDIA, ATI, and Intel. Despite all the odds and obstacles, NVIDIA had emerged as the leading producer of discrete[2] graphics processors and one of the largest "fabless" semiconductor companies in the world. In 2001, it shipped about 47 million discrete graphics processors, 60% of the total market. According to Wired Magazine,[3] "Nvidia has risen to dominance by pushing the power of its high-end GPUs, letting the technology trickle down to cheaper price points, and hitting delivery windows. If that process sounds familiar, it is. In effect, Nvidia has become the Intel of graphics."

Development of the 3D PC Graphics Industry

Up through 1995, the development of high-performance graphics was largely aimed at expensive engineering workstations or video game consoles. This pattern changed in the mid-1990s due to three distinct events: (1) Microsoft's Windows 95 included multimedia features—audio and video capabilities—that triggered a surge of interest in multimedia, (2) 3D action games for the PC platform appeared, and (3) increasing integration in ICs, driven by Moore's Law, passed a critical threshold allowing a number of critical 3D graphics processes to be placed on a single chip.

The PC 3D Graphics Industry in 1995

In 1995, PC graphics was delivered by cards or boards plugged into the computer's bus. Graphics on mobile PC was delivered by (less sophisticated) chips integrated into the machine's mainboard. A graphics board consisted of at least three primary parts: the graphics chip, graphics memory, and RAMDAC. The graphics chip accepted information about what was to be displayed and converted it into information about the color of each point on the screen. The RAMDAC converted the information in the frame buffer into the analog signals necessary to drive a video display. More expensive and elaborate graphics cards might include audio functions, MPEG video compression coders and de-coders, and video capture functions.

[1] "Vidia" is an amalgam of "video" and "via," Latin words meaning "to see" and "the way" or "the road." The letter "N" is an abbreviation of the Latin "natus," meaning "to come into existence."

[2] Integrated graphics processors (IGPs) were integrated into the chipset accompanying the CPU. Discrete processors, by contrast, were higher-performing graphics units mounted on add-in boards, or designed into the motherboard by a systems integrator.

[3] Jeffrey M. O'Brien, "Nvidia," *Wired*, July 10, 2002. www.wired.com/wired/archive/10.07/Nvidia_pr.html.

The industry had two primary types of players: chip companies and board companies. Chip companies designed and made graphics processor chips. A board company (board-OEM) assembled the chips, memory, RAMDAC, and other components on add-in boards, wrote software drivers, and sold the boards. In 1995, the largest chip companies were S3, Cirrus Logic, and Trident. The largest board-OEMs were Diamond, Creative, ELSA, and STB. One important exception to this model was Toronto-based ATI, which designed its own chips and manufactured its own boards.

The three major chip companies in 1995 were Cirrus Logic, S3, and Trident with market shares of 35%, 20%, and 14% respectively. But competitive positions were volatile: just two years before, S3's share had been only 10%. In the next year, 1996, S3 passed Cirrus, coming to lead the industry with a 33% share, Cirrus' share falling to 19%.

Most consumers fastened on the brand names of the board-OEMs rather than the names of the chip makers. Diamond had the *Monster* and *Viper* series of add-in boards for the PC. ELSA produced the *GLADIAC* and *Synergy* lines. Each of the graphics cards within these lines had a graphics processor designed by one of the chip companies. Nevertheless, the board manufacturers were willing to shop around for the best performing chipset at the time, having no particular loyalty to any one chip company.

The Lure of Multimedia

Interest in multimedia skyrocketed in the mid-1990s as new devices, new media, and new operating systems appeared together with the burgeoning of the Internet. Windows 95 gave the PC platform the audio and video capabilities that Apple's Macintosh had already had for several years. Plunging prices on CD drives meant that more consumers could purchase and use CDs, exciting interest in multimedia encyclopedias and other yet to be discovered publications. To some, the promise of DVDs and Internet broadband in the future seemed to open up limitless opportunities for multimedia.

Coupled with the interest in multimedia was a fascination the issue of "convergence." Many analysts saw both the PC and the Internet as instruments of convergence, bringing publishing, video, and audio technologies together. Video game consoles were also projected to be centers of multimedia convergence, as they added more sophisticated audio, graphics, and video capabilities, added CD drives and Internet connectivity.

In this environment, NVIDIA offered its first product, the NV1, in 1995. The NV1 chip had been developed jointly with SGS-Thompson Microelectronics and combined 3D graphics, audio, and a port for Sega game pads. Jeff Fisher, NVIDIA's Executive VP of Sales, recalled that "the original dream was to be the Soundblaster of multimedia." The chip was sold to a number board-OEMs. The best known was Diamond Multimedia who branded the NV1 board as the *Diamond Edge 3D*.

The NV1's graphics were based on the concept of parametric surfaces: it was optimized for geometry described by four-sided quadratic (curved) elements. This was also the approach taken by Sega in their Saturn game console. Parametric surfaces allowed the representation of smooth curved surfaces with great accuracy. Sega expressed interest in having NVIDIA develop a chip for its forthcoming new game console (Dreamcast), so work was begun on an even more advanced chip, the NV2.

The NV1 received kudos for being such a technologically advanced chip, but its proprietary quadratic texture mapping proved to be its downfall. OpenGL, Microsoft's Direct3D, and 3Dfx's Glide were all based on polygons and developers were unwilling to invest in the special approaches needed for the NV1. Early in 1966, NVIDIA ceased sale of the NV1. Shortly thereafter, its development arrangement with Sega was cancelled, and work on the NV2 was halted.

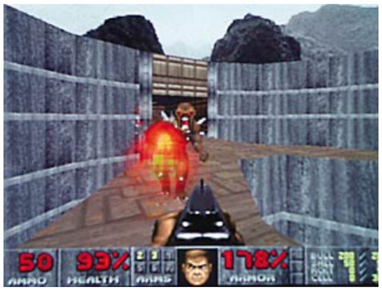
3D Action Games and the Rise of 3Dfx

NVIDIA's initial failure in the 3D graphics business was all the more frustrating because, by mid-1996, rival start-up 3Dfx[5] had pulled off a stunning success, taking leadership of the 3D graphics race with its Voodoo chip set. The "killer-app" powering 3Dfx's rise was neither multimedia nor console gaming. It was instead the emergence of 3D action games on the PC platform.

Id Software

PC-based 3D action games were the brainchild of John Carmack and John Romero, who had formed Id Software in 1991. Their hit games, *Wolfenstein 3D*, *Doom*, and *Quake* redefined action games, introduced clever new technologies for displaying 3D scenes on PCs, and moved the center of gravity of computer gaming innovation from the console industry to the PC platform. Each of these games was a "first-person shooter;" the player's view was through the eyes of a soldier as he races through rooms, battles Nazis and monsters, jumps obstacles, operates buttons and switches, peers around corners, and delivers withering fire from a variety of weapons. The dark themes and violence reflected a purposeful rejection of the child-like plots and characters in Nintendo's video games.

Screenshot of Doom



Carmack, only 20 when he wrote Doom in 1993, had already spent years struggling to get PC graphics to perform like arcade video games and to display 3D views. However, in a 3D game, the scene had to be redrawn each time the player's position changed, and PCs were too slow to do this quickly enough. Redrawing the scene every 5-10 seconds, for example, completely eliminated the sense of action and eye-hand coordination that were essential elements of action games. Carmack speeded things up by simplifying the 3D world to make it render faster and by fooling the eye. In Doom, the basic 3D objects were walls (trapezoids), each assigned one of a set of textures. The mapped textures, together with diminished lighting with distance, gave the appearance of a detailed scene even when there were only 8-10 surfaces in view. Carmack adopted the 3D scientist's binary space partition (BSP) method of storing object information, innovating by applying it to the whole 3D world (or level) being played.

Id distributed Doom, released in 1993, together with 9 levels of play, without cost, via the Internet. To play additional levels, one had to purchase the registered version of the game. With free distribution over the newly burgeoning Internet, the game was an overnight sensation. In another sharp break with practices in the console industry, Id also freely distributed a game

[5] 3Dfx changed its name to 3dfx in 1999.

editor, enabling enthusiasts to create wholly new levels for the game. Within a year, literally hundreds of user-written levels for the game were freely available on the Internet.

In 1996, Id released Quake, the breakthrough game for 3D graphics. In Quake, Carmack removed many of the simplifications of Doom, rendering complex 3D objects that could be rapidly viewed from any angle. The player in Quake could actually move in three dimensions and much more complex lighting and shading techniques were employed.

Perhaps most importantly, Quake had a version which was a multiplayer game with Internet connectivity. The Quake server, running on any PC, kept track of each distant player's keyboard and mouse actions and sent messages to the PCs of each player. A player's PC, running Quake locally, displayed the real-time results of their own and other players' actions. A large on-line community developed around Quake, with new levels constantly being created by user groups. Online gamers became amazing proficient at the game, constantly honing their skills in contests with other skilled players. After Quake, most PC games included online play as a feature.

Id's Quake was the pivotal game in pushing the development of 3D video hardware because the complexity of the Quake engine meant that the frame rate (screens per second) slowed drastically as the resolution, or image quality, was increased. For reasonable game play on a normal PC, users had to select 320x240 resolution, which had a grainy pixilated look. Gamers appreciated good image quality, but also demanded speed—online players saw a game's frame rate as a matter of life and death.

Screenshot of Quake

3Dfx

Seeking a significant improvement in game performance, Carmack began discussions with a few companies who wished to build 3D accelerator cards. 3Dfx Interactive, a 1994 startup, had plans for a very high-performance board, but Carmack objected to the $400 projected price. Instead, he chose to work with chip-designer Rendition in making their Verite chips work with Quake. The improvements obtained in look and game-play were real, but unspectacular.

Lacking a deal with Id, 3Dfx continued the development of its card, the Voodoo. It also developed a proprietary graphics API branded as Glide. 3Dfx's strategy was to tie a popular game to Glide and its Voodoo board, preventing easy imitation because Glide was proprietary. The first software firm to accept the gambit was CORE Design, building Tomb Raider to work with Glide. The real-time images of a 3D Lora Croft were the hit of the 1996 E3 trade show.

At the same time, Id found that there wasn't a good fit between the Verite chip's structure and the Quake engine. Seeking to stimulate some innovation, Carmack released GLQuake on the Internet, a *free* version of the game that provided its output in the OpenGL language. Any board that could interpret OpenGL with speed, would have an immediate market! 3Dfx was poised to exploit this move: the cost of RAM memory had plummeted in two years and its high-performance boards could be sold for $200. It wrote a "wrapper" program that translated OpenGL into Glide and, in the third quarter of 1996, began to sell Voodoo cards to support the free GLQuake. Using Voodoo, GLQuake ran at 640x480 with 16-bit color at 30 frames per second, giving better-than TV quality resolution. Sales took off like a rocket and the 3D accelerator business was born.

The 3Dfx Voodoo card's raster operations shined at creating blended anti-aliased scenes, without the jagged edges and pixel artifacts that other chips produced. 3Dfx was one of the new breed of "virtual" semiconductor firms. It was fabless, outsourcing manufacture of its chips to TMSC in Taiwan. It neither distributed nor retailed its products. Its chips were sold to board-OEMs who marketed the plug-in boards under their own brand names. 3Dfx's primary customer was Diamond Multimedia and its Voodoo board, *Monster 3D*, sold for $199.

Questioned about how 3Dfx managed to achieve such a sharp technological advance in 1996, Gary Tarolli, 3Dfx's Chief Technology Officer said[6] "we left out 2D! This simplified the problem and freed up lots of gates. And we used 2 chips." A Voodoo board was purely 3D, containing no 2D functionality. Consequently, it did not replace the user's 2D video board, but was a pure 3D add-on board. When initialized, it took control of the entire screen image. When released, it handed control back to the PC's regular video controller.

The API Issue

An API—application program interface—was a *standard* governing the communication between an application program and a software service, normally supplied by an operating system. A graphics API permitted an application program to ask for a set of shaded lighted polygons, defined in three-space, to be displayed and the API would sort out which covered which, how to project them onto a viewing plane, how lighting shifted with viewing angle, and how generate the required pixel-pixel color information.

In the 1990s, Silicon Graphics (SGI) had refined the idea of the graphics pipeline and developed a language (GL) for invoking pipeline operations. A public standard, OpenGL, was released in 1992. It then consisted of about 120 commands to draw points, lines, and polygons. OpenGL emerged as the API of choice for engineers and programmers working in Unix and Linux environments.

Microsoft Windows NT, released in mid-1995, supported OpenGL. To Microsoft's surprise, the growing market for 3D applications was not NT users. Instead, it was 3D gamers. Microsoft decided that it would develop its own API for consumers—one that combined 3D graphics with audio and video support. Reworking device drivers obtained via the acquisition of Rendermorphics, Microsoft announced a new 3D API called Direct3D. Direct3D was one of the services supplied by a variety of multimedia drivers, the complete bundle being named DirectX.

Many developers were not enthusiastic about DirectX. In 3D games, the CPU was fully loaded by the complex mathematics of transformation, lighting, and rendering. Developers feared that Microsoft's Direct3D would not be efficient enough to support gaming. In the quest to outperform competitors, firms would have to work around it. The debate was fully joined when John Carmack posted a critique of Direct3D. He had ported Quake to OpenGL a month earlier and had just abandoned an attempt to do the same for Direct3D. It was, he concluded, "a horribly broken API. It inflicts great pain and suffering on the programmers using it without returning any significant advantages. …There is not good technical reason for the existence of D3D.…" Carmack concluded that Id Software would use OpenGL in game development. He would, of course, continue to use DirectX drivers for input and audio.

[6] "Interview with 3Dfx CTO Gary Tarolli," sharkyextreme.com,.November 21, 1998.

Microsoft's Bill Gates replied directly, posting arguments for Direct3D. Other Microsoft spokespersons chimed in, noting that DirectX would have higher-quality drivers, certified by Microsoft, and that there was a clear path for its improvement and support over time.

NVIDIA's New Strategy

Faced with the failure of his company's first two products, Jen-Hsun[7] Huang, NVIDIA's President and CEO, reformulated the company's product policy. For a time, a technical advisory board was created consisting of Ed Catmull (Pixar), John Carmack (Id Software), Doug Kaye (Mondo Media, web graphics), and Pat Hanrahan (Stanford Computer Graphics Lab). New technical depth was added, with David Kirk hired to be Chief Scientist. Second-round venture capital was raised from first-round investors Sequoia Capital and Sierra Ventures.

The changes in direction were sharp. Instead of multimedia, the company would focus on desktop PCs. Instead of quadratic surfaces, the company would embrace triangles. With regard to APIs, the company would bet on the ascendance of DirectX and design its chips to work directly with the DirectX API. OpenGL would be supported via wrapper software.

Top management sketched out a roadmap that saw ever more powerful graphics chips. The increased power would come from three sources: speed, parallelism, and integration. Moore's Law would permit faster clock rates and more transistors per chip. The increased number of transistors would enable more integration and parallelism. Integration meant bringing more of the graphics pipeline onto the chip and parallelism meant more raster engines, and, eventually, even triangle engines, working simultaneously.

To accomplish the roadmap, the company's hardware engineers were organized into three development teams. While one group worked on the upcoming chip design, another group worked on the spring refresh of last year's design, and a third worked on next year's new design.

NVIDIA's top management paid particular attention to the sources of delays in the development process. Designing a chip was a logical problem and the work was done with software tools. However, hardware development was much slower and riskier than software development. Software could be tested quickly and debugged, often on a daily or weekly cycle. By contrast, once a chip was designed, masks had to be constructed and a preliminary fabrication run accomplished before the design could be tested electronically. That test, in turn, required appropriate interface electronics and driver software. But the driver software itself could not be fully tested until a chip was ready for it to drive.

In the semiconductor industry, the completion of chip design was called *tapeout* because it had been traditionally signaled by writing magnetic tapes containing the coded instructions for the construction of the necessary masks. Using contract fabricators, the wait from tapeout to first samples about one month. If bugs were found in the hardware, there could be lengthy delays as designs were altered, masks redefined and rebuilt, and a new fab run initiated. To address this issue, NVIDIA invested in emulation technology. Specifically, NVIDIA organized its chip design around the use of simulation and emulation techniques.

Simulation and emulation involved testing the logic of a chip, its actual electrical functioning, and having a sufficiently good "model" to allow the advance development of driver software. Design tools were the forte of founder Chris Malachowsky and he pushed the use of

[7] Pronounced Jen-Sun.

sophisticated tools for the formal verification of a chip's logic. In addition, the company began to invest in simulation of physical design—understanding the electrical characteristics of a particular physical layout. Finally, FPGAs (field-programmable gate arrays) were used for emulation tasks such as driver development. FPGAs were general-purpose ICs that could be "programmed" to mimic a chip or module. FPGA emulations ran hundreds or thousands of times slower than the real chip, but could be invaluable in avoiding costly re-spins. In addition, the simulations could be used to develop drivers before the chips were fabricated.

NVIDIA's founders all believed strongly in taking control of the creation of drivers—the chunks of software that connected the PC user's operating system to the actual graphics hardware. In the era of technologically simpler 2D graphics, board-OEMs performed most of the driver development. This pattern of operations had four important drawbacks. First, the complexity of 3D chips meant that driver development required a whole new range of skills and knowledge. If poor drivers were written, NVIDIA would be damaged. Second, board-OEMs could only start driver development after receiving working chips from the chip maker, creating an additional delay. Third, board-OEMs had mixed incentives about communicating driver problems to the more knowledgeable chip-designer for solution. Chip makers typically sold to several board makers. Thus, if NVIDIA were selling chips to two board-OEMs, any information on driver problems passed back to NVIDIA from one would also benefit the other. As a step towards dealing with driver problem, NVIDIA committed itself to the early development of drivers, through simulation. Forth, NVIDIA wanted to write a unified driver that would handle all its chips, as they developed over time. Independent driver development would be incompatible with such a policy.

The Riva 128

In a first test of the new strategy, Jen-Hsun Huang, NVIDIA's CEO, bet the company on the NV3, designed in early 1997. The NV3 would be the most complex graphics chip ever built. To be successful, the NV3 would not only have to compete with the Voodoo, but also with the new Rage Pro chip introduced by ATI. The NV3's design featured a 128-bit memory bus and was one of the first video graphics cards to use the new Intel-sponsored AGP high-speed memory port. It featured on-chip triangle setup and raster operations. Unlike the Voodoo, it was also a normal 2D video chip.

In developing the new chip, NVIDIA used a number of emulation tools. VP of Software Development Dwight Diercks recalled the results of a many hour simulation using new techniques:

> The graphics output looked great, but then someone noticed that there was an extra pixel between the edge of the image and the "C:\" prompt. The hardware guys said "Close, but your simulator isn't dead-on. There is an extra pixel there on the left." We stayed up for two days checking it out. It wasn't the simulator. It was a real hardware design flaw. If we hadn't caught it with simulation, it would have required a full spin to fix it. That really helped prove the value of this approach.

The NV3 was released in August 1997 as the *RIVA 128,* and it was sold to board-OEMs Diamond Multimedia, Canopus, Elsa, and STD. Diamond's RIVA 128 board was branded as the *Diamond Viper 330*.

RIVA 128 add-in boards retailed for $140 and up and provided a one-board solution with excellent 2D and 3D performance; they gained good acceptance from PC-OEMs. The RIVA 128 also outperformed the Voodoo chip set on games designed to run under Microsoft DirectX. Whereas the Voodoo used a wrapper program to translate DirectX into Glide, the RIVA 128 was designed from the ground up to work with DirectX. Conversely, of course, the Voodoo did much better than the RIVA 128 with games that had been optimized for Glide. And most Glide-only games would simply not even run on the RIVA 128.

TomsHardware.com, a popular hardware review site, gave the RIVA 128 high marks for its speed and ability to run OpenGL games at higher-than-Voodoo resolutions. It reserved its overall crown, however, for the 3Dfx Voodoo because it was critical of the RIVA 128's image quality. TomsHardware.com preferred the look of Voodoo's anti-aliasing methods.

The RIVA 128ZX was a spring refresh (upgrade) of the RIVA 128; it had twice as much video memory and a faster AGP port. NVIDIA's novel approach to drivers became apparent with the release of this chip. The new driver for the RIVA 128ZX was not only compatible with the older RIVA 128, it actually increased the older chip's performance.

ATI's contemporaneous Rage Pro, was caught in the middle of this competition. The Voodoo beat it on Glide games and the RIVA 128 beat it on Direct3D games. ATI did not supply OpenGL drivers for the chip until late 1999.

NVIDIA's Drive to Dominance

In the four years 1998-2001, NVIDIA introduced eight new graphics processor cores to the steady beat of a 6-month development cycle. The complexity of these processors increased dramatically with each new release. Whereas the RIVA 128 had 4 million transistors, the RIVA TNT, released one year later, had 7.5 million. One year after that, the new GeForce 256 had 22 million transistors, about as many as Intel's Pentium III chip. The GeForce3 Ti, released in fall 2001, weighed in at 63 million transistors. (See Exhibit 1 on page 1.)

With each new release, NVIDIA added features and increased speed. Each new, more powerful, chip made old games look better and run faster and, in addition, opened up new possibilities for developers. In the two-years following the release of RIVA 128, leadership in 3D graphics passed from 3Dfx to NVIDIA. In addition, the NVIDIA chip advances ran in tandem with advances in Microsoft's DirectX. NVIDIA worked with the Microsoft DirectX development team to build support into their chips for the current version of DirectX and for features being included in the coming version. The pace of development at both Microsoft and NVIDIA was roughly similar—each new chip was associated with a different version of DirectX.[8]

The RIVA TNT

One year after the RIVA 128, NVIDIA released the RIVA TNT. With 7.5 million transistors, it was as complex as a Pentium II and much more powerful with regard to numerical calculation—it boasted a capacity of 10 gigaflops (billion floating point operations per second).[9] It had double the fill rate of RIVA 128. A key feature was its support for 32-bit color. 3Dfx used

[8] RIVA 128 supported DirectX 3.0, RIVA 128ZX supported DirectX 5 (Windows 98), the TNT supported DirectX 6, the TNT2 supported DirectX 7, and the GeForce 256 supported DirectX 8.

[9] The chip's computing capacity came from its parallel architecture—a number of different engines worked simultaneously on different triangles and on different rasterization problems.

16-bit color and would take another year to move to 24-bit color. Although many gamers continued to like the speed of the Voodoo on Glide-specific games, TNT provided a smoother less ragged color image. The TNT's performance marked a competitive inflection point, for from that point onwards, NVIDIA would steadily erode 3Dfx's market position.

Beginning with the TNT, NVIDIA adopted a policy of speed binning[10] its high end chipsets. "We decided to segment the market into different speeds," explained Huang. "No one had done this before in graphics."[11] Binning let NVIDIA offer the same basic at different price points. This allowed NVIDIA to have a range of products without bearing the development cost of separate designs.

The TNT also marked the use of Taiwan Semiconductor Manufacturing Company (TSMC) as a sole-source foundry. TSMC was the largest and most sophisticated IC foundry in the world and NVIDIA was an important customer. To support manufacturing operations in Taiwan, NVIDIA contracted with specialist firms in Asia to package, test, inventory, and distribute the chips to board-OEMs.

The TNT was optimized for Microsoft DirectX 6, which was released one month before the chip. Despite industry doubts, the people on the Microsoft development team were gaming enthusiasts who understood the importance of efficiency and stable drivers. The new version was a significant improvement over past versions of DirectX. The existence of good drivers for DirectX, especially NVIDIA's, encouraged more and more game developers to commit to writing games for that API.

The RIVA TNT, and its "spring refresh," the TNT2, were the first chips to fall under NVIDIA's unified driver architecture (UDA). This patent-protected approach to building a software driver had been a vision of the founders, especially Curtis Priem. Using this architecture, all NVIDIA chips used the same driver software. The UDA was an object-oriented specification that associated a set of graphics elements and actions with certain objects. The basic idea was that a module in the driver software sent inquiries to the chip as to its type and what objects and actions were supported. When a graphics service was needed, a specialized module (resource manager) determined which classes to employ.

In the early spring of 1999, the company's IPO raised $45 million and established a market capitalization of $400 million. Within 60 days, the company's market capitalization had increased to $850 million. By December it touched $2 billion.

As the NVIDIA's TNT2 was taping out, 3Dfx announced it was acquiring STB, a key board-OEM customer for RIVA TNT chips and a planned distributor of the TNT2 to PC-OEMs. At the same time, management was reviewing its policies with regard to board-OEMs generally. Jeff Fisher, Executive VP of Sales, recalled "We wanted to take control of the OEM relationships, working directly with large commercial customers, like Dell. One of the problems we hit with

[10] Binning was as old as electronics. For example, 100 Ohms resistors came off the production line with actual resistances of 70-130 Ohms. Those that fell in the range 99-101 Ohms were labeled "100 Ohm ±1%." Those that fell in the ranges 90-99 and 101-110 were labeled "100 Ohm ±10%". Speed binning was common in the RAM business.

[11] Roberts, Bill. "Growing Pains," *Electronic Business*. Highlands Ranch: May 2002. Volume 28, Issue 5, p. 64.

board-OEMs was that they would tune the drivers to do better on certain benchmarks. In the process, bugs were introduced."

However, initial discussions with Dell revealed that strong board-maker brands, like Diamond's, were important. Discussions with Diamond were unproductive. Fisher said that "Diamond couldn't face the economic reality that they were not adding any value to the board. They wanted to earn 25% margins on commodity DRAM and printed circuit boards, while messing up drivers and slowing the product's roll-out, all in an attempt to differentiate themselves. ATI was also looking for 20 margin points on the boards they made. Our position was clear—we didn't want to make boards because there was so little value added. Making boards would have killed our gross margins."

Top management organized a major effort to sell Dell. High quality Dell-certified contract manufacturing companies were sought out and agreements established. The economic benefits of the UDA were outlined. A team traveled to Dell and won—Dell would offer boards with NVIDIA chips, manufactured by Celestica Hong Kong, Ltd. In the months and years to come, NVIDIA would increasingly rely on such contract electronics manufacturers (CEM) for board production and distribution. The CEMs were free to brand the board as they chose; most chose to emphasize the NVIDIA name.

GeForce 256

Seven months after the TNT2, NVIDIA released the GeForce 256. Where the TNT2 gave NVIDIA parity with competitors; the GeForce moved the 3D graphics industry into new territory. The new chip was fast and had four parallel pixel pipelines. But its key feature was that it integrated the transform and lighting stages of the pipeline onto the graphics chip. Up until the GeForce, the industry had called its products graphics *accelerators*. NVIDIA called the GeForce a GPU for *graphic processor unit*, calling attention to the fact that its complexity was on a par with the Intel and AMD CPUs.

GeForce 256 had nearly 23 million transistors, twice the complexity of the Pentium II. Its floating point calculation capacity was 50 gigaflops, the equivalent of the Cray T3D super computer. It delivered 15 million triangles per second and a fill rate of 480 million pixels per second. By the time the chip was released, NVIDIA had built the world's largest site for chip design verification and simulation.

The traditional benchmarks used in the industry did not measure the GeForce's new features. Still, using the traditional benchmarks, the GeForce 256 beat all other chips. Only under special conditions—low screen resolutions, slow CPU, and Glide optimized games—could the new Voodoo3 beat the GeForce. And even there, reviewers wondered whether being "faster" mattered when the GeForce's frame rate was already over 100 per second.

When the chip was released, few software titles could exploit the chip's advanced features. The transform and lighting engine, for example, could dramatically speed up a game written in OpenGL, but its real value was in rendering detailed 3D scenes composed of thousands or millions of polygons. But the OpenGL games that existed did not have high polygon counts. Any game designed to look great on the GeForce 256 would simply not run on a system without its hardware support for transform and lighting calculations. Consequently, the GeForce 256's initial value was that it ran existing games faster. Any OpenGL game was instantly accelerated and, as games for the new DirectX 7 appeared, they also were accelerated.

The GeForce 256 was optimized for DirectX 7, which appeared two months before the chip was released. NVIDIA software engineers, along with engineers from ATI and a few other graphics companies, had worked with Microsoft on the definitions of the version 7 API. The appearance of DirectX 7 closed the debate over DirectX versus OpenGL for games. Even Id Software moved to work with DirectX as well as OpenGL. With each new version, released about once per year, DirectX's features and usability had improved. OpenGL, by contrast, had not evolved as rapidly. One industry observer wrote[12]

> In order for software developers to make use of the cool new features in video cards, they have to be supported in a widely accepted API. OpenGL simply isn't changing fast enough to be worthwhile. DirectX changes once a year. Every spring, the DirectX team at Microsoft unleashes a new beta development kit for the next version of DirectX, and in late summer or early fall, it's available as a download for end users. OpenGL changes as often as its governing body deems it necessary. … Several game developers have cried for improved texture management functions in OpenGL for some time now. Where are they?

X-Box

A few months after the release of the GeForce 256, Microsoft announced that it would build a new game platform, dubbed the "X-Box," and that NVIDIA would supply the graphics processor and communications processor. The chip to be built for the X-Box would be a proprietary version of the planned NV25. The new game console would feature a DVD player, a 8 GB hard disk, HDTV support, a broadband Ethernet port, an Intel Pentium III CPU and an operating system based on the kernel of Windows 2000.

NVIDIA's X-Box win was universally hailed as a huge plus for the company. It was seen as flowing from the firm's technological leadership and it was forecast that it would yield $2 billion in revenues over five years. Furthermore, it was expected that the X-Box activity would open up further business in the consumer electronics and communications industries.

GeForce2 and GeForce3

Roughly six months after shipping the GeForce 256, NVIDIA released the GeForce2 GTS. The chip contained 25 million transistors and its performance astounded analysts and competitors. Not a simple refresh, the GeForce2 GTS was the first graphics processor to allow programmers to control the shading of each pixel, rather than just of a triangle. John Carmack enthused that "Per-pixel shading … looks incredibly good across an entire world. Everyone at Id is way psyched about developing new content with the GeForce2 GTS."[13]

The new chip did more than slam home the fact that NVIDIA led the PC graphics performance race. It broke with the pattern of the spring release being a "refresh" of the previous fall release. That summer, NVIDIA formalized its commitment to move to a full 6 month development cycle. CEO Jen-Hsun Huang said:[14]

> NVIDIA's core strategy is to deliver a breakthrough product every six months, doubling performance with each generation at a rate of essentially

[12] Jason Cross, "The Hard Wire Column," *Computer Games Magazine*, February 25, 2001.

[13] NVIDIA Press Release, April 26, 2000.

[14] NVIDIA Press Release, August 14, 2000.

> Moore's Law cubed. Our unique ability to innovate at this staggering pace is one of the key elements of our competitive advantage. Our consistency has made it possible for our partners to rely on our architecture and roadmap. GeForce2 Ultra exemplifies our commitment to help our partners build innovative and winning products.

It was rumored that ATI engineers, after seeing the specifications of the GeForce2, went back and re-designed their transform and lighting engine on the up-coming chip, the Radeon 256. On release, the transform and lighting engine on the Radeon was 20% faster than the GeForce2 GTS. Soon thereafter, NVIDIA distributed a new UDA driver, Detonator 3. The new driver gave the older GeForce2 GTS a performance boost of 20-30%, enough to overcome the performance advantage of ATI's Radeon.

In the spring of 2001, the eagerly-awaited NV20 appeared as the GeForce3. The GeForce3's dramatic new feature was programmability—the chip included programmable vertex and pixel shaders.[15] These new capabilities were coordinated with and supported by DirectX 8. Released five months previously, DirectX 8 provided API support for the pixel and vertex shaders. NVIDIA engineers had worked closely with Microsoft in developing the newest version of DirectX, with Microsoft actually licensing NVIDIA's shading technology for inclusion in DirectX. TomsHardware.com wrote:

> The amount of technology provided by NVIDIA's new GeForce3, the Xbox predecessor chip, is simply overwhelming. This processor doesn't 'just' have 57 million transistors. They have actually been USED! If it is the Vertex Shader, the Pixel Shader, the High Resolution Anti Aliasing, the Crossbar Memory Controller or the Z Occlusion Culling, this chip has been STUFFED with high-tech!

DirectX 8 and the GeForce3 (and ATI's programmable chip, Radeon 8500) opened up a new way to create details: the use of "programs" to control the position and lighting at each triangle vertex, and the lighting at each pixel. Such programs were challenging to create, but offered the author much more room for creativity, much higher image quality, and, consequently, more differentiation in the finished product.

In fall 2000, NVIDIA released the GeForce3 Ti (Titanium). With the GeForce3 Ti came another upgrade in the NVIDIA UDA—Detonator 4. The new driver supported the new features of the GeForce3 Ti, and it also substantially improved the performance of the older GeForce3 and GeForce2 chips. Many reviewers and users were effusive in their praise for this free boost in performance for older chips. Some observers also noted that ATI's brand new chip, The Radeon 8500, had been beating the GeForce3 in some benchmark trials, but the new Detonator 4 driver had closed the gap. How much more performance, they wondered, was latent in NVIDIA's older chips?

Competitors

This section provides summary information on the fortunes of 3Dfx, Intel, Cirrus Logic, S3, ATI, and Silicon Graphics.

[15] The per-pixel effects on the GeForce2 had not been programmable. See the Note on 3D Graphics Technology for details on vertex and pixel shaders.

Exhibit 4 provides data on the sales history of leading 3D chip makers.

3Dfx

The original Voodoo's weakness had been that it did not handle ordinary 2D video. In mid-1997, 3Dfx introduced its 2D/3D Rush chip-set, aimed at the PC-OEM market. The product was not very successful. The 2D chip was somewhat unstable and the 3D performance was 20% slower than Voodoo. It was also incompatible with a number of already-written Glide games. Hardcore gamers stuck with the Voodoo1. This stumble hurt 3Dfx's reputation with gamers, and OEMs. Nevertheless, in the summer of 1997, the company was able to raise about $40 million in an IPO. The funds were used for hardware development and marketing.

In March 1998, 3Dfx released its Voodoo2 chip, 17 months after the Voodoo1. Management attributed the longer than one-year wait to problems with the Rush chip and with a diversion of resources to a forthcoming chip, the *Banshee*. The Voodoo2 was three times as fast as the original Voodoo1, and could apply two textures simultaneously to a triangle. Additionally, the new chip set let speed-hungry gamers run two Voodoo2 boards in the same machine, doubling the speed. Still a 3D-only solution, the Voodoo2 was snapped up by hardcore gamers. Surprisingly, 30% of users buying Voodoo2 solutions opted for a $600 two-board setup. Exhibit 2 provides a timeline of key NVIDIA, ATI, and 3Dfx chip introductions. Exhibit 3 displays the specifications of 3Dfx and NVIDIA chips. In 1998, 3Dfx's chips were on the top-selling 3D graphics boards, garnering about one-third of the market the performance market (about 15% of graphics chips went into performance boards in 1998).

3Dfx's CEO, Greg Ballard, had been hired at the end of 1996, as Voodoo1's sales began to ramp up. His prior experience had been in a game-development company. During 1998, he became more and more concerned with the company's brand image and with breaking into OEM sales channels. Hardcore gamers knew about 3Dfx, but most buyers saw only the brand names of the board manufacturers (e.g., STD, Diamond) or the brand names of the specific boards (e.g., Blackmagic 3D, Monster 3D).

In the summer of 1998, 3Dfx redoubled its marketing efforts. It began direct consumer advertising of the "3Dfx" brand and initiated a "3Dfx Inside" campaign with selected OEMs like Packard Bell. The idea was twofold—to establish an Intel-like brand for a chip and to establish "3Dfx," rather than the hardcore gamer "Voodoo," as the company's brand image in PC graphics. Scott Sellers, one of 3Dfx's founders, noted[16]

> You just can't find many examples of companies, other than monopolies like Intel, that have extremely strong brand presence when they are not selling directly to the end consumer.… as our board partners got more and more successful with our products, it became more and more unclear to the end consumer what they should be buying.

3Dfx released the Banshee chip in September of 1998, finally giving the company a 2D/3D product. But the Banshee had lower performance compared to Voodoo2. Banshee had only one pixel pipeline, rather than two. 3Dfx sold over 1 million Banshee boards, but enjoyed little success in the target PC-OEM market. CEO Ballard said "Banshee is the single biggest mistake

[16] Quoted in Geoff Keighley, "Reality Comes Knocking: The Story of 3Dfx Interactive," *GameSpot.com*, 1999.

we've made. …I should have invested in a bigger physical back-end group so we could have done Voodoo2 and Banshee at the same time."[17]

In December, 1998, 3Dfx announced it was acquiring board-maker STB for $141 million. With this move, 3Dfx thrust itself into the highly competitive add-in board manufacturing business. Management declared that henceforth STB would be the company's sole and exclusive graphics board manufacturer and distributor, and that the boards would carry the 3Dfx brand. 3Dfx chips would no longer be sold to other board-OEMs. STB, NVIDIA's largest single customer, would no longer buy NVIDIA's chips. These policies were aimed at building the 3Dfx brand and at capturing STB's PC-OEM business. STB boards with NVIDIA TNT chips had scored significant design wins at Dell and other OEMs. 3Dfx hoped to convert these OEMs to alternate designs based on 3Dfx chips.

3Dfx's stock fell 21% the day the STB deal was announced. Nevertheless, analysts were mostly enthusiastic about 3Dfx's change in strategy. Robertson Stephens opined[18]

> With the graphics industry hurtling toward a leaner more vertical operating model, we believe this acquisition was necessary for 3Dfx. The PC market will likely mandate cost efficiency of the vertical markets. Moreover, the Darwinian advantage of the quick response time possible with vertical models, could further narrow the graphics field and enlarge opportunities for early movers.… As the trend to vertically integrate continues, there are a limited number of attractive partners for 3Dfx to choose from. With its aforementioned OEM relationships, we believe STB is potentially the strongest candidate for 3Dfx.

Some observers worried about a loss of focus on development, lower overall margins, and problems connected with STB's new plant coming on line in Juarez, Mexico.

3Dfx's acquisition of STB did not yield the hoped-for results. Sales declined as Diamond and other board-makers severed relationships with 3Dfx, and STB lost its design wins with OEMs. Brian Hook, the author of Glide, recalled that management was "way deluded into thinking that they could leverage STB's relationships with OEMs like Dell into getting inclusion in Dell's systems. Bzzzt. OEMs don't care about the board vendor; they care about the chip set. Choice of board vendor becomes an issue when it comes down to pricing and availability. "[19]

During the preceding year, 3Dfx engineers had been working on a next generation technology, code-named Rampage. But the Banshee effort had diverted key technical resources away from Rampage. Problems with design and manufacturing had extended that diversion by six months. By the time Banshee was finished, Rampage had to be reconceived—in the fast-moving world of graphics chips, the original technology concept was competitively obsolete.

Pushing Rampage back to the drawing board, 3Dfx shifted its design resources to upgrading Banshee. The new chipset, Voodoo3, was shipped in early-1999, one year after the Voodoo2 (7 months after Banshee). On release, it was the fastest 3D board on the market—just

[17] Quoted in Geoff Keighley, "Reality Comes Knocking: The Story of 3Dfx Interactive," *GameSpot.com*, 1999.

[18] "3Dfx Moves to New Strategic Model with STB Acquisition," *Robertson Stephens*, December 30, 1998, p. 1.

[19] Quoted in Geoffrey Keighley, "Reality Comes Knocking: The Story of 3Dfx Interactive," *GameSpot.com*, 1999.

about as fast as the dual Voodoo2 configuration.[20] For some hardcore gamers, bragging rights connected to frame rates were all important and Voodoo3 delivered 49 fps on the Quake 2 "Crusher" time test, compared with the TNT's 30 fps. However, reviewers and many gamers complained that the 3Dfx product was still rendering colors with only 16-bit accuracy, so that its image quality was dated.[21]

After the Voodoo3, 3Dfx divided its development efforts between two projects: Napalm and Rampage, both involving major new chip architectures, but there were problems in keeping the schedule. When NVIDIA introduced GeForce 256 and its heavy-weight follow-on, GeForce2, topping any 3Dfx card's performance. 3Dfx's stock went into tailspin. Voodoo4 (Napalm) was released in mid-2000. The 14-million transistor chip had been a huge challenge for the physical design team.[22] CFO David Zacarias explained that "It was so complex, it was breaking some emulation tools. It took a long time to verify."[23] CEO Ballard resigned in the fall of 1999.

3Dfx continued efforts on Rampage until, in the last months of 2000, it closed its doors, selling its patents, brands, and inventory to NVIDIA for $70 million in cash plus contingent stock consideration. Rampage was never released.

Intel

In early 1998, Intel entered the 3D graphics business, releasing the i740 chipset with the goal of capturing the high end graphics market. Expectations had built that i740 would beat existing hardware by a factor of two. NVIDIA was pushed into a loss during the second quarter of 1998 as customers waited to see if the Intel chip would prove to be a success. Investors were also cautious, expecting that Intel would dominate the business.

Intel's i740 did not provide the success envisioned. It outperformed the RIVA 128 on speed trials, but had no OpenGL driver to support key games. More critically, Intel did not rapidly upgrade, and was quickly left behind when the TNT appeared. In 1999, Intel announced that it was exiting the discrete graphics chipset business completely.

Jon Peddie,[25] explained that Intel had developed the i740 with the same processes and same approach that it used to develop its CPUs. These processes were largely driven by budgetary issues and manufacturing timelines. Intel's more methodical approach did not work in the extremely competitive 3D graphics industry. Intel, he argued, did not adapt to the graphics chip 6-12 month development cycle, rather than the microprocessor industry's 18-24 month cycle.

[20] There was only one AGP slot on PC motherboards, so AGP cards, like Voodoo3, were not designed to run in dual-card modes.

[21] The Voodoo3 chips delivered two 16-bit color signals which were "mixed" to smooth the image. 3Dfx called it "24-bit" color.

[22] The physical design team does "back-end" design: the placement of elements on silicon and design of the connection and routing systems for signals and power.

[23] Quoted in Robert Ristelhueber, "3Dfx burned by 'Napalm' graphics chip delay," *Silicon Strategies*, November 10, 1999.

[25] John Peddie Research was a technically oriented marketing and management consulting firm specializing in graphics and multimedia.

Intel continued to develop and produce integrated graphics processors (IGP). These processors were integrated with the CPU chipset on the motherboard of a PC. Intel garnered the lion's share of this market. The largest impact of Intel's entry fell on the marginal graphics players who had been competing for a share of the low-end market. From being an industry with 20-plus participants, Intel's IGP entry was largely credited with concentrating the industry down to three main players: Intel, ATI, and NVIDIA.

Cirrus Logic

In 1993, Cirrus Logic dominated the graphics accelerator business, supplying IBM, Apple, HP, Compaq, Sony, Toshiba, and others. It relied on internal wafer fabrication as well as outsourcing and reached a 60% market share. During the 1995 capacity crunch it had enough allocated capacity to ignore the shift to 64-bits, then lost share rapidly when enough capacity came online to supply S3 and others. By 1998, it had eliminated its PC graphics products.

S3

A 1989 startup, S3 adopted the fabless model and released its 64-bit 3D Virge in 1995. The next year, S3 surpassed Cirrus Logic to become the market-share leader in PC graphics. Cirrus Logic had been the leader, but had been slow to move to 64-bit chips. Two years later, S3 was in decline and volume leadership passed to ATI.

The company offered a novel 3D technological path with its Savage 3D chips (1998), but hardware bugs and poor drivers kept the card from being well received. In an attempt to change its fortunes, S3 acquired board-OEM Diamond Multimedia in mid-1999. The company slowly withdrew from the performance 3D graphics market and refocused on graphics chips for notebook manufacturers and IGPs for companies using VIA (non-Intel) chipsets.

ATI

ATI Technologies was established in 1985 and immediately released its Graphics Wonder board. The company continued to design and manufacture graphics boards as standards evolved, maintaining its position as a key player in the graphics chip and board industries. Unlike many chip design firms, ATI had been integrated into board manufacturing for many years.

When the 3D revolution struck in 1996, ATI lost the battle for performance leadership. Its 3D Rage cards were outclassed by products from 3Dfx and NVIDIA. Some of the issues had to do with hardware and some with problems in producing high-quality stable drivers in a timely fashion. Nevertheless, the performance segment of the market was, at that time, relatively small and ATI was, by 1998, the dominant supplier of 3D graphics chips, accounting for 25-27% of shipments. ATI maintained a strong position in the "value" (lower-priced) segment and moved to gain a dominant share in 3D graphics chips for notebook PCs.

In 1998, ATI pushed into the notebook market with its Rage Mobility chip. ATI made rapid progress, stripping market share from then notebook graphics leaders S3 and Neomagic. Struggling with supply problems, Neomagic was rapidly defeated. It withdrew and refocused on wireless multimedia.

Even though ATI gained share in mainstream graphics chips, management came to believe that, in time, Intel IGPs would dominate the low-cost sector of the market. CEO K.Y. Ho decided that ATI needed new leadership. In April 2000, ATI acquired 70-person ArtX for $453 million. ArtX had been formed in 1997 by a group of disgruntled SGI engineers who left to focus on PC graphics. Many in the group had worked at SGI with Nintendo on the N64 graphics processor.

Within a year ArtX entered into an agreement with Nintendo to develop a graphics processor for the new gaming system it planned, GameCube.

In ArtX, ATI obtained a game console-related business and a solid pocket of ex-SGI graphics specialists. Key among them was ArtX CEO, David Orton, who had previously managed SGI's Visual Systems Group. Orton was quickly given the COO job at ATI.

ATI's new objective was to expand into game consoles, wireless multimedia, and regain leadership in high-end PC graphics. Orton's first operational moves were to re-organize ATI's development work. In the past, ATI's development group had worked with a functional structure, different sub-groups having responsibilities for different aspects of design. Under that structure, engineers gained a great deal of experience at their specialty but also had to work across several design projects at the same time. Orton re-organized the group around project teams, each of which followed a chip design through to completion. In addition, he kept many of the ArtX employees on the west-coast, forming a separate team. The two teams were to leapfrog one another, hopefully reducing the company's development cycle time.

The new team's first chip was the Radeon 256, released in the summer of 2000, after the NVIDIA GeForce2 GTS. A true GPU, the 30-million transistor Radeon contained an integrated graphics pipeline. The company also released versions of the Radeon for the mobile market and two Radeon-branded IGP chips, one for desktops and one for notebooks. ATI binned the card, pairing the faster chips with a larger faster RAM block. The top-line card retailed for $399 and the slower card for $279. By contrast, the GeForce2 GTS was priced at $349 on launch.

The new chip featured three texture engines, giving it good performance on games using three textures per pixel. The transformation and lighting engine handled vertex skinning and keyframe interpolation, features that were absent in the GeForce2 GTS. Both features would have to wait for new games before their benefits would be evident. Vertex skinning allowed one set of polygons (the skin) to ride on, or adjust to, another set (the skeleton). Keyframe interpolation simplified the details of animating objects and expressions. A key technology introduced with the Radeon 256 was HyperZ. This was essentially a compression method that boosted the effective memory bandwidth of the board.

Despite the fact that ATI and NVIDIA were clearly competing on the basis of very advanced features, review sites continued to benchmark the chips speeds on standard games (Quake Arena) and tests (3D Winbench 2000). The Radeon did beat the GeForce2 on some preliminary benchmarks, but a new release of NVIDIA's Detonator driver reversed that lead. The new board was not the "fastest," but it did signal that ATI was back in the race for the number one position.

One year later, ATI introduced the Radeon 8500. This 60-million transistor GPU was rushed to its release just after NVIDIA's GeForce3 Ti. If the Radeon 8500 outperformed the latest NVIDIA chip, ATI would have 6-12 months of leadership. The chip's features were roughly similar to that of the GeForce3 Ti, but it had a slightly higher clock rate. The chip's specifications predicted that it could outperform the GeForce3 Ti in speed trials. Initial reviews, however, focused on driver problems. AnandTech.com wrote: "the Radeon 8500's `final' performance was a bit disappointing; we weren't expecting parity with the $199 GeForce3 Ti 200, we were expecting a GeForce3 Ti 500 killer. All of the specs pointed at a higher performing product, but in the end we are limited by what has been ATI's Achilles' heel: drivers."

New drivers appeared late in 2001. Review sites reported that with the new drivers the Radeon 8500 was competitive with NVIDIA's top product, beating it on a number of tests. Review site SharkeyExtreme.com wrote:

> Potentially the most noticeable are issues with long-distance 3D images on the RADEON 8500 … These ranged from very noticeable texture tearing or aliasing problems on stairs and building joints to entire blacked-out polygons where textures should have been. This type of problem is not present with NVIDIA cards … While the GeForce3 Ti 500 still gets the nod for all-out gaming systems, the RADEON 8500 is virtually unmatched as a total 2D, DVD and 3D total video solution.

Silicon Graphics

In its prime, SGI produced the most advanced graphics workstations that were used by graphic artists and engineers. The special effects in Jurassic Park were constructed on SGI workstations. In 1996, SGI acquired supercomputer maker Cray and chip maker MIPS. The expense of the move, and new competition on the workstation front, soon undermined the company's financial health and it entered into a turnaround mode that lasted several years. Dan Vivoli, NVIDIA's Executive VP of Marketing, said "SGI made all the wrong decisions. They turned left when they should have turned right."

In August of 1999, SGI was seeking ways to restructure and reached an agreement with NVIDIA: SGI would be able to use NVIDIA technology in its workstations, and would transfer about thirty of its 3D graphics engineers to NVIDIA. By 2001, SGI graphics workstations sported NVIDIA GPUs, and SGI alumni were guiding the fortunes of NVIDIA and ATI.

3D Graphics in 2001

The tech-boom of the 1990s ground to a halt in mid-2000. Calendar 2001 delivered the first year-over-year reduction in PC shipments, a 4.6% fall to 128 million units worldwide. At the very least, it was clear that the $37 billion personal computer industry had become mature. Unit shipments had grown at 15-25% per year during the last decade, but all indications were that the industry had become a mature cyclical with an expected future growth rate of 4% per year.

The fall in shipments boosted semiconductor inventories and most semiconductor makers experienced reductions in sales and sharp drops in profit. Intel's revenues fell from $34 billion in 2000 to $27 billion in 2001; its net income dropped precipitously from $10.5 to 2.2 billion. By contrast, in the same year NVIDIA's revenues rose 86% to $1.4 billion and net income rose 96% to $193 million.

The PC industry was the largest user of semiconductors, accounting for $36 billion in 2001, about one-quarter of all semiconductor revenue. Of this $36 billion, 62% was for CPUs, 14% for memory chips, 10% for CPU chipsets, and 5% for graphics processors.

Despite maturity, it was expected that price reductions in PCs would continue, driven by technological advances. Between 1996 and 2001, average selling prices (ASPs) had dropped from $2060 to $1450. CPU ASPs declined from $222 in 1997 to $170 in 2000 and then to $140 in 2001. By contrast, unit prices of graphics processors had been rising. Dan Vivoli, NVIDIA Executive VP of Marketing, said "Since 1997, Intel's price on a top-end CPU has come down by 50%. The price of a top-end NVIDIA GPU has risen by a factor of four or five. The difference is that Moore's Law is our 'friend.' While increasing transistor density threatens CPUs with becoming $5 items, 3D graphics needs all the power it can get. Each increase in power enables a big increase in the consumer's experience and enjoyment."

Categories

Graphics processors were of two basic types: integrated and discrete. Integrated graphics processors (IGPs) were generally simpler and less expensive and were built into the computer's motherboard, along with other parts of the *chipset*[26] supporting the CPU. About two-thirds of PCs shipped included an IGP; the remaining one-third contained a discrete (add-in) graphics processor. Shipments of discrete processors actually exceeded the number of PCs shipped by 20-30%. The extra discrete processors were put into PCs that had shipped with IGPs in order to upgrade their performance, were installed in older computers, and were installed as upgrades to older discrete processors.

The other main distinction in the market was between desktop and notebook computers. In a desktop computer, an IGP was installed on the motherboard whereas a discrete processor was placed on an add-in card which was plugged into the bus. Notebooks were much more integrated from the start, each manufacturer having their own internal mainboard design. Both IGPs and discrete processors had to be integrated by the manufacturer onto a notebook mainboard. Users could not upgrade the graphics in a notebook.

In 2001, 174 million graphics processors where shipped, 85% of them going into desktop PCs. About one-half of these were discrete processors. Of the 26.8 million graphics processors going into notebooks, about 73% were discrete processors. Manufacturer's revenue from the sale of all discrete processors amounted to $1.8 billion in 2001, with $1.1 billion from desktop discrete processors.

Producers

In 1996, analyst/consultant Jon Peddie had counted 16 suppliers of graphics processors; in 2001 he counted 11. The shakeout had not simply eliminated marginal players, it squeezed out some major players and dramatically re-ordered market shares. In 1966, the three leading producers of graphics producers had been S3 (40% share), Matrox (30%), and ATI (17%). In 2001, the big three were NVIDIA (30%), Intel (26%), and ATI (18%). Each of the three producers dominated a different segment of the market. NVIDIA's dominance lay in desktop discrete processors (63% share), and ATI's was in notebooks (46%). Intel dominated the IGP business (about 85% Share). The data in

[26] The chipset is a collection of microchips that support the CPU, governing the flow of data. A PC chipset normally includes a bus controller, a DMA controller, the AGP interface controller. It may also include a graphics controller.

Exhibit 5 summarizes the situation.

Discrete graphics processors were sold to both PC-OEMs and the public, with about 8-10% of production going into add-in cards distributed through retail and e-tail channels. The largest PC-OEMs were Dell (14%), Compaq (11%), HP (8%), and IBM (6%). Large PC-OEMs economized by specifying, whenever possible, a standard motherboard that included a chipset with integrated graphics. To "customize" the machine to a user, a faster CPU was plugged into the CPU socket, extra memory could be added, and a higher-performance graphics board could be plugged into the bus, disabling and replacing the services of the integrated graphics chip. PC-OEMs normally provided customers with a list of options for upgraded graphics capabilities. They did not seek exclusive discrete processor suppliers, and often carried both NVIDIA and ATI boards. They did, however, pay careful attention to customer demand, to cost, and to the service and support expenses associated with different discrete processor vendors.

Corporate purchasers accounted for about 60% of PC demand. Corporate purchasers' primary concerns were cost, reliability, and compatibility. They normally arranged either for a standard machine, or a set of machines, to be available for employees. Normally, machines for office support had simple IGP graphics. Professionals often had the option of specifying more advanced graphics options. Artists, engineers, and designers who worked in areas where graphics was essential were normally provided with very high-performance solutions.

Among individual consumers, some showed a clear interest in graphics performance and were aware of brand names in the industry. Others were unaware of graphics issues and simply purchased a PC, accepting whatever graphics solution the OEM had supplied.

The typical home user who purchased an expensive graphics card was a gamer, dubbed *enthusiasts* by the industry. Enthusiasts accounted for about 10-15% retail unit volume, but had an important influence on the buying habits of others. Exhibit 6 displays price, revenue share, and profit pool share for retail board segments.

NVIDIA in 2001

Product Line and Marketing

In 2001, NVIDIA sold 20 different graphics processors grouped into five major brands: GeForce, Quadro, Vanta, TNT2, and nForce. In addition, it made specialized graphics and communications chips for the Microsoft X-Box. The GeForce2 and Geforce3 were the company's high-end desktop chips. The GeForce2 was available in the original desktop version, a less expensive "MX" version, and a mobile version dubbed the GeForce2 Go. The TNT2 had been introduced in 1999, together with its low-end version, the Vanta. nForce was an integrated graphics chipset that could be installed on a motherboards using AMD processors. Exhibit 7 provides details about the product line. The company did not provide breakdowns of its business by product line, but Merrill Lynch[27] had estimated that 27% of 2001 revenues came from TNT2, 47% from GeForce2, 15% from GeForce3, and 8% from X-Box chips.

NVIDIA's pricing was aggressive. Its early market share gains in 1998-2000 were associated with accepting gross margins in the 32-38% range rather than the 45% margins that

[27] Joseph Osha and Matthew Chan, "NVIDIA Corporation: Getting Big, Fast," Merrill Lynch, November 19, 2001, p. 18.

had been traditional in the industry. By 2001, firms that could not compete on these terms had left the market.

NVIDIA did not engage in extensive advertising. Nevertheless, the company had found that mainstream computer users had a 30% unaided awareness of the NVIDIA brand. Management believed that it had achieved this by delivering exciting new advances in 3D technology on a regular basis. Dan Vivoli, NVIDIA's Executive VP of Marketing, explained the key role played by technology-aware enthusiasts:

> The opinion leaders in PC graphics are the enthusiasts. This group has held stable at about 2-3 million worldwide for a number of years. Enthusiasts are gamers, know how to upgrade their computers, often overclock their machines and add special cooling devices. An Electronic Arts survey showed that 70% of their gamer customers were willing upgrade their graphics for a better gaming experience.
>
> Enthusiasts often belong to gaming leagues. There is something like a rebel skateboard "thresher" culture in the enthusiast group. The top enthusiasts are famous within their circles, demonstrating amazing amounts of skill and knowledge. Enthusiasts want to have the newest best graphics chip and they aren't afraid to upgrade over and over again, if they can afford it. They get information easily on the Web and frequent the enthusiast sites, like GameSpot.com, and the tech sites, like TomsHardware.com, AnandTech.com, and HardOCP.com. On average, one enthusiast influences 10-25 other people each year.
>
> A larger key group is the *hobbyists*: 13-24 year-old "millennials" who are comfortable with technology. Hobbyists value a good graphics experience, but don't devote a lot of time or money to graphics. They use computers, video game consoles, cell phones, and know how to rip a CD into mp3 files. This is a growing group that tends to follow the lead of the enthusiasts.

Vivoli stressed that in marketing to PC-OEMs, their key interests were price, performance, and the expense of customer support. In mobile applications, power consumption was also important. Although NVIDIA marketed its products to key PC-OEMs, it relied on a worldwide network of board makers and motherboard manufacturers to actually sell into OEMs, systems integrators, and retailers.

To satisfy the needs of consumers and OEMs, NVIDIA provided chips with various levels of performance at different price points, kept assembly costs low by using skilled low-cost contractors to assemble boards, and used its Unified Driver Architecture (UDA) to cut software and drivers issues to a bare minimum.

Management viewed the UDA as important competitive strength. Dwight Diercks, VP of Software Engineering, explained:

> Every chip we have made is based on the UDA. It is very flexible: any chip will work with any driver. This makes it easy for us to support many products, even those manufactured years ago. Our web site recorded 65 million downloads of our UDA driver during the last 6 months!
>
> It would not be easy for one of our competitors to implement UDA. These companies did not start out this way and have a history of different chip

> architectures and drivers. To incorporate UDA principles they would have to virtually start over—a very costly proposition. Indeed, many of our competitors use more than one chip architecture. Both S3 and ATI, for example, use different chip architectures for different products. 3DLabs uses three, due to its acquisitions.
>
> For its driver, ATI posts a single large file with multiple binaries. Once the install manager detects what kind of chip is in use, it selects and loads the appropriate binary. This method is not integrated—a whole host of different drivers must be maintained and supported. Also, it does not work so well for corporate users who want to propagate a single disk image to multiple computers which may use different graphics cards.

NVIDIA's full line of graphics processors achieved its variety by purposeful design, speed binning, and the maintenance of older models. For example, when the GeForce2 GTS was released in April 2000, it appeared on cards retailing for about $300, and was mainly bought by hardcore gamers. Soon thereafter, NVIDIA released the GeForce2 MX, based on the same core as the GeForce 2 GTS, but with somewhat slower performance (a slower clock rate and 2, rather than 4, pixel rendering pipelines). On the other hand, the GeForce2 MX supported dual video monitors. The chip, aimed at the high-end mainstream, appeared on boards priced at about $200. To provide additional options, the GeForce 2 MX was speed binned into three groups, and given brands MX, MX 200, and MX 400. Finally, as newer models appeared, options multiplied further. CEO Huang said

> … our product life is much longer than our product cycles. How it works is a new high-end product has technology elements that soon move into midrange products, which then have a life of perhaps a year and a half. Then cost reductions push them into a low-end product that can last more than two years. For example, the RIVA TNT2 was introduced in the spring of 1999, and we will ship it well into the spring of 2002. And the GeForce2 moving behind it will ship into the spring of 2003. We replace the high end every six months. And each of these processors over its life will generate over a billion dollars of revenue. I don't know of any place other than Intel where that happens. [28]

NVIDIA's newest product was the XGPU delivered for the X-Box. Built on the same core as the GeForce3, the XGPU was expected to be a major contributor to revenues in the next few years. It was estimated that the X-Box might achieve a volume of 10 million units per year. Wall Street analysts forecast that NVIDIA's might realize $450-500 million in revenues from this business by 2003. Margins, however, were thought to be slim. Microsoft was selling the X-Box at a loss and analysts guessed that NVIDIA would not earn more than 20-22% on these revenues.[29]

Exhibit 9 provides financial statements for NVIDIA. Exhibit 10 provides similar statements for ATI for comparison.

[28] Quoted in Larry Waller, "NVIDIA Processors Dominating 3-D Graphics," *Electronics Journal*, March/April 2001, pp. 6-9.

[29] "The Trojan Horse that is X-Box," *Dundee Securities*, March 12, 2001. "NVIDIA," *Harris Partners*, January 20, 2003.

Manufacturing and Distribution

NVIDIA was one of the leading "fabless" semiconductor company in the world. Its chips were fabricated by Taiwan Semiconductor Manufacturing Company (TSMC), and independent contractors performed assembly, testing, and packaging operations. NVIDIA accounted for about one-fifth of TSMC's revenues. TSMC was the only company that NVIDIA used for foundry services. Di Ma, VP of Operations, noted that multiple sourcing was more common in the industry and that sole sourcing can be a challenging relationship:

> NVIDIA and TSMC are two strong companies and sometimes negations can be a zero sum game. On the other hand, the sole source relationship can have advantages. We get to know each other pretty well and it is very difficult for TSMC to say "No" to NVIDIA. After all, it is our only foundry partner and one of their biggest customers. Their inability to say "No" is an important factor in the relationship.

GPUs were the most complex semiconductor devices fabricated and there were few foundries capable of pushing Moore's Law as fast as TSMC had in its work with NVIDIA. The tooling and equipment required to push to the cutting edge was very expensive and foundries asked for, and got, a premium for such work. Marvin Burkett, NVIDIA's CFO noted that as one got closer to the "bleeding edge," the foundry received a greater portion of the potential surplus. "There are usually four technology choices available, directly measurable by the element width in microns. Old technology is very competitive and many foundries can do it. Then, you can use modern technology and give up, say 30% to the fab. Going further, you can use cutting edge technology. Few fabs can do it well and you will give, say, 70% to the fab. Finally, you can push the frontier to the bleeding edge and give up 100%." In 2001, old technology was 0.25µ (microns,) modern was 0.18µ, cutting edge was 0.15µ, and bleeding edge was 0.13µ.

In 2001, NVIDIA's top-end chips used 0.15µ processes and the company had been the first of TSMC's customers to access this process. Its one-year old GeForce2 chips were built using a 0.18µ process and the next major step (the NV30) would be 0.13µ. In 2001, TSMC's production mix had been about 20% cutting edge (0.15 µ), 25% modern (0.18µ), and the rest 0.25µ and older. Other foundries were less technically aggressive. At UMC, for example, only 5% of revenues came from cutting edge technology.

NVIDIA's "virtual" supply chain (Exhibit 8) started with TSMC and extended to other third parties. After chips were fabricated by TSMC, they were transferred to three firms for assembly, test, and packaging.[30] The finished chips were then shipped directly to board-OEMs and contract board manufacturers (CEMs) or delivered to stocking companies[31] in Singapore and Hong Kong, or to the company's Santa Clara warehouse. NVIDIA used an outside firm to manage a "supply hub" which ensured a smooth flow of chips to customers. The most critical aspects of the supply chain were the channels to the CEMs who manufactured add-in boards for PC-OEMs. Dell, for example, was primarily supplied by Celestica (Hong Kong), which also made other components for them. Dell managed a "pull" system out to Celestica and the hub had to accommodate that discipline.

[30] Advanced Semiconductor Engineering (Taiwan), ChipPAC (US), and Siliconware Precision Industries (Taiwan).

[31] JSI Shipping (Singapore) and Ten-Up International (Hong Kong).

Managing the complex system of fabricators, stocking companies, and suppliers required considerable skill, but running the virtual system gave NVIDIA a great deal of flexibility. One example how the system worked was provided by the GeForce3 Ti chip. This chip used the same architecture as the GeForce3 introduced 6 months earlier, but it had higher performance. The performance jump was achieved by (1) pressing TSMC to use a faster process technology, still staying within the basic 0.15 micron process, (2) working with CEMs to redesign the board, giving it an additional layer and a revamped power distribution pattern, and (3) working with its memory chip suppliers to adjust their speed binning cut points so as to permit (nominal) 3.8 ns memory chips to be clocked at 250 MHz, up from the 230 MHz clock rate on the GeForce3 memory chips.

Research and Development

Being fabless meant that NVIDIA was fundamentally a development company. R&D at NVIDIA included development and selection of ideas and concepts for implementation in hardware, the electrical and physical design of the hardware, an extensive design verification and testing program, and the development of software drivers and software to support the design and development process. About 600 of its 1000 employees worked in research and development.

An active intellectual community, centered in several universities and companies, engaged in a constant discussion and debate concerning how to best advance the field. Out of this ferment, NVIDIA had to select a few ideas for further development and implementation in hardware. The constraints were support by DirectX, potential support from game developers, and the current state of semiconductor fabrication technology.

The basic design cycle at NVIDIA was geared to two new important chip releases each year. This pattern was the heartbeat of the company and was attributed to Jen-Hsun Huang's early insight that PC-OEMs operated with such a cycle, driven by the schedules of government and schools. The expectation was that the fall release would be a major step forward and the spring release would be an upgrade. NVIDIA did not always fully adhere to this pattern—its spring release of the GeForce2 GTS in 2000 had surprised the industry in being an important new chip, not simply a refresh of the GeForce 256.

The whole development process for a major new chip took 16-22 months from start to volume production. A less complex refresh project might cut that time in half. The were five basic steps that had to be accomplished in the development process:

1. Logical design and verification took about 12 months on a new chip, 1-2 on a refresh. In this task, a team of 100 or more designers working on the logical design of the new chip. This logic expressed how the chip would implement the planned features and was expressed in millions of line of code—high-level languages for describing a chip's logic. This code was then tested on emulators which examined the internal consistency of the logic and simulated the actual functioning of the chip.

2. About two months before the completion of logical design and verification, physical design began. Physical design was the specification of the actual gates on silicon, their locations, and interconnects. It resulted in the code describing the photolithographic masks used to actually fabricate the chip. Physical design took about four months.

3. After design was complete, the chip taped out and went to the foundry for fabrication. After about one month, NVIDIA received back a batch of test chips—first silicon. These were carefully tested. Errors required rework and delay.
4. After the chip qualified, board designs had to be refined and qualified and the mass production assembly, test, and packaging operations tuned and verified. These steps took about 2 months.
5. The development of software drivers was begun about 6 months before tapeout and continued through to the shipment date. Software was delivered with new boards, but was also widely distributed via the Internet.

Simulation

A key aspect of NVIDIA's development process was the avoidance of delay, especially "full silicon spins." If the first silicon back from the foundry showed a problem, there might be serious delays. A minor problem with connections might require new masks for the top few layers—a metal spin-- taking 5-14 extra days. A major logic error would require redesign and going through tapeout again for all masks--a silicon spin. A silicon spin entailed a delay of 6-12 weeks.

One study of silicon spins in complex communications ICs showed that only 10% of designs worked at first silicon; 31% required an extra silicon spin, and another 31% required 2 respins. NVIDIA had never needed a full silicon spin; its average new core required 2 metal-only spins (Exhibit 11—NVIDIA Graphics Processors) and its average refresh required none.

NVIDIA Simulation Facility

The company's ability to create designs without logical errors played a key role in its rapid development cycle. In turn, its competence at design was linked to the skill of its technical staff and its large investment in design verification tools (emulation and simulation). Joe Sura, VP of Information Systems, said that

> NVIDIA has the largest semiconductor simulation activity on the planet. In 2001 we had about 2000 rack x86 CPUs, about 2000 Linux CPUs, and about 250 Sparc/Solaris racked servers. The totality had about 1 terabytes of RAM and 25 terabytes of disk storage.
>
> The more you simulate the better. NVIDIA first used simulation on the Riva 128. Over time, our methods have evolved. We have learned how to do this. I have a saying, "Everything you test works. Everything you don't test, doesn't." Everybody tests things; we try to test everything.

Research and development expense in 2001 was $154 million, 11.2% of revenue. In the previous year it had been $86 and $47 million, or 11.7% of revenue. The amount spent on R&D depended on the number of new designs created each year and on the complexity of each design. Complexity increased with the number of transistors, the clock speed, and the logic (versus memory) density of the chip. For GPUs, NVIDIA management estimated that developing a new chip cost $1 per transistor.

Software

The development of drivers for new chips was the responsibility of Dwight Diercks, VP of Software Engineering. Diercks managed a group of about 180 software engineers. Within the group, engineers specialized in DirectX, OpenGL, Macintosh, systems and bios, networking, video decoders, DVD-TV, and production software for OEM customers.

The biggest challenge the software engineers faced was the need for rapid development. Direcks explained

> Back in the old days [1996], engineers in graphics had about 10-12 months for driver development. Now, in large part due to NVIDIA's tight cycle, there is much less time. There are about 90 days from tapeout to having 1 million chips ready to ship. Without working drivers, that silicon is worthless. About 28 days after tapeout we get back about 30 chips, of which 5 or 6 work. We put them on boards and start to test the drivers that have been prepared. By day 45, we have 15-30 boards—finally enough to really start work (there are over 100 engineers who want access!) At the end of the process, we need about 10 days for Microsoft to provide a WHQL certification. So there are only about 35 days for software development.
>
> To handle this, we start driver development about 6 months before tapeout. We develop using emulation—large racks of FPGAs. We spend 4-6 months developing against emulators so we are ready when the first chips appear.

Technological Directions

A number of NVIDIA's top managers saw the company's overall long-term technological goal as creating "simulated reality." In the near-term, this meant real-time images that looked as good as the best animated motion pictures. But unlike a movie, these real-time images could be explored and instantly modified by the user. People watched a movie. By contrast, a user has the experience of moving around in a 3D scene, taking any feasible viewpoint, and, becoming a virtual participant in the action. Ultimately, one engineer explained, the goal was "the Holodeck on starship Enterprise."

NVIDIA management believed that the GPU was the key to the future of consumer computing. In the X-Box, Microsoft paid NVIDIA more for the GPU than it paid Intel for the Pentium III CPU. They argued that the returns to greater computational power was much higher for the GPU than the CPU. Chief Scientist David Kirk explained:

> Today we are millions of times short of reality. But that is what gives this business its potential—there is a virtually limitless demand for computational power in 3D graphics. Given the architecture of the PC, there is only so much you can do with a more powerful CPU. But it is easy to use up 1 teraflop of graphics computing power. The GPU is going to be where center of technology and value added in consumer computing.

Looking to the future, there seemed a lot of room left for continual improvement along the path already charted out. Kirk said

> At NVIDIA, our technical path has been to crawl up the graphics pipeline. At each stage we put more operations onto specialized silicon, where performance improved by a factor of ten compared to the CPU. It was exciting—as we integrated we got higher speeds with fewer chips. The *Riva128* added hardware setup to the graphics chip. The *TNT* added hardware transformation. With *GeForce* we got all of SGI's steps on one $100 chip, running faster than the *RealityEngine*, which had cost $1 million in 1992. *GeForce* was the first GPU and became the gold standard for the industry.

NVIDIA's march up the graphics pipeline had been facilitated by the pipeline's functional modularity. The two key units of modularity were the triangle and the pixel. The "Utah" paradigm divided all 3D objects into triangles and computed the light on each in a simplified way—it lit each triangle based on its position, but did not generate complex lighting based on reflections or diffractions from all the surfaces in the scene. That simplification made it possible to have "triangle engines" working in parallel, each transforming and lighting a set of triangles. By adding more engines, all working in parallel, the process was made faster and more triangles could be processed. Similarly, the problem of breaking triangles into pixels and lighting each pixel could be handled with parallel pixel rendering engines. Because the paradigm did not have pixels interacting with one another, each engine could work on some subset of the pixels to be displayed.

The GeForce3 had taken the first big step beyond the classic SGI pipeline, but stayed within the overall paradigm. The chip introduced programmability: triangle engines and pixel rendering engines could have aspects of their behavior controlled by small programs written by developers. Like the old features, these new capabilities were subject to the power of parallel processing. Programmability did introduce some potential problems. First, in the past it had taken 12-18 months for game developers to learn to use the new features on a chip. Using pixel and vertex shaders, however, introduced a whole new level of complexity. This complexity might slow the development of new games. Second, many game developers did not have the algorithmic skills to use the capability to its fullest.

Future Directions

The following quotations were taken from Larry Waller, "NVIDIA Processors Dominating 3-D Graphics," *Electronics Journal*, March/April 2001, pp. 6-9.

Jen-Hsun Huang (CEO):

> Our competitive advantage is the combination of our massive engineering resource coupled with our passion and expertise in 3-D graphics. We have arguably the largest pool of 3-D intellectual capital in the world. And because our projects are constantly breaking new ground in the field, we've become the company to work for 3-D engineers. Our ability to attract world class technical talent is our competitive advantage. …
>
> Over the next several years, 3-D technology will change dramatically. We now have access to millions of transistors, necessary to cost effectively implement the complex algorithms for synthesizing super realistic images. Jurassic Park dinosaurs will not just be in movies. They will come to life in your computer. Because of the complexity and demand for advanced GPUs, the industry will also change dramatically. GPUs will eventually be incorporated on

all computing platforms, from laptops, Internet appliances, and eventually handhelds. And you will see the workstation industry stop building proprietary graphics workstations, turning to companies like NVIDIA with the resources to invest in these massively complex GPUs. …

The market for 3-D is massive. As the market footprint for our products continues to increase, we see competition from several fronts. Intel and ATI are formidable competitors on the PC front. Sony will be our primary challenge in consumer electronics. Our partnership with Microsoft on Xbox will put NVIDIA on a world stage as we take on Sony in the highly visible game market. With respect to Intel, we are terrific partners and competitors simultaneously. In the performance segment, we partner to advance the platform. In the value segment, our upcoming integrated GPUs, derived from the Xbox architecture, will compete against Intel's core logic.

Exhibit 2—Timeline of Graphics Processor Introductions

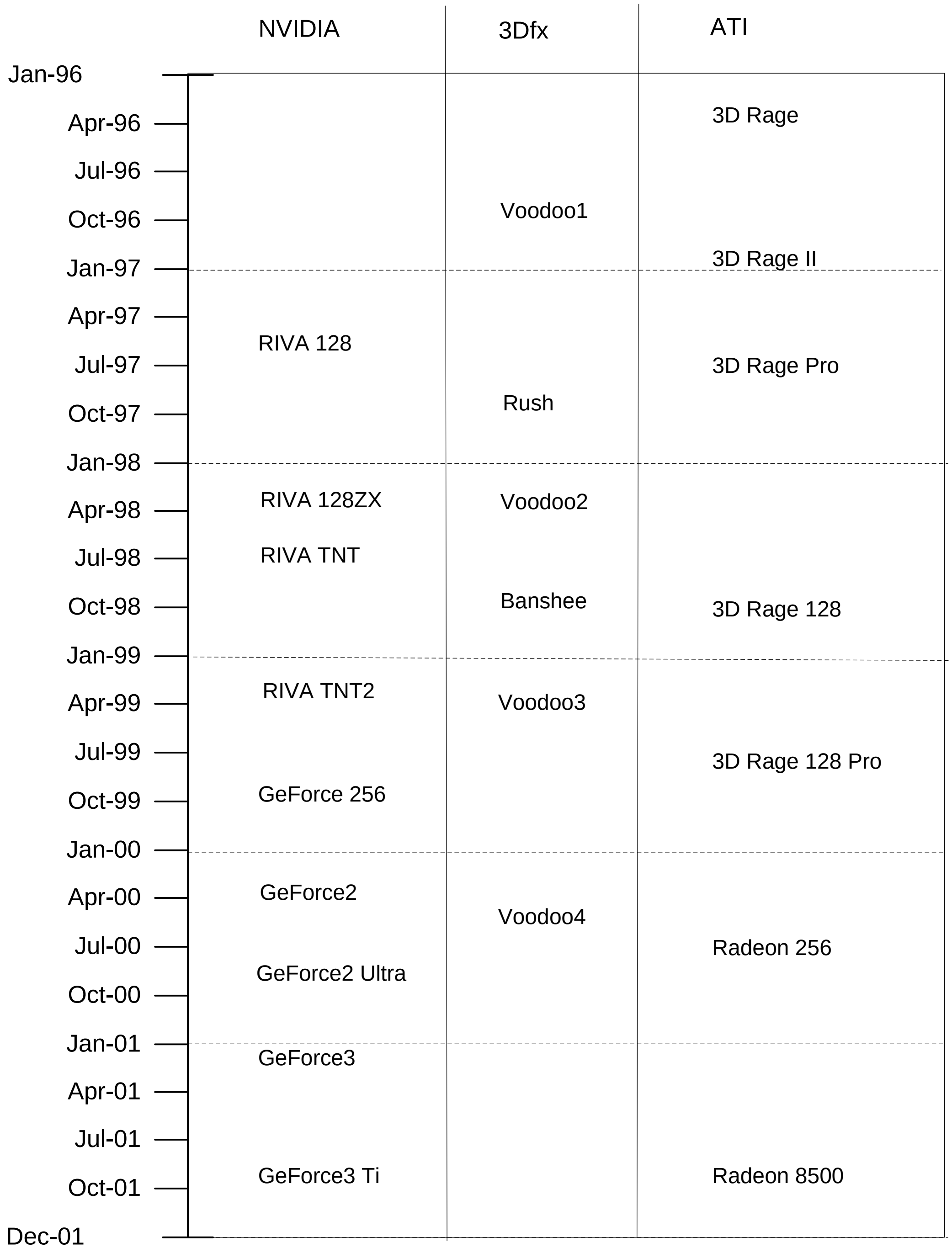


Exhibit 3—NVIDIA and 3Dfx Graphics Processors 1996-2000

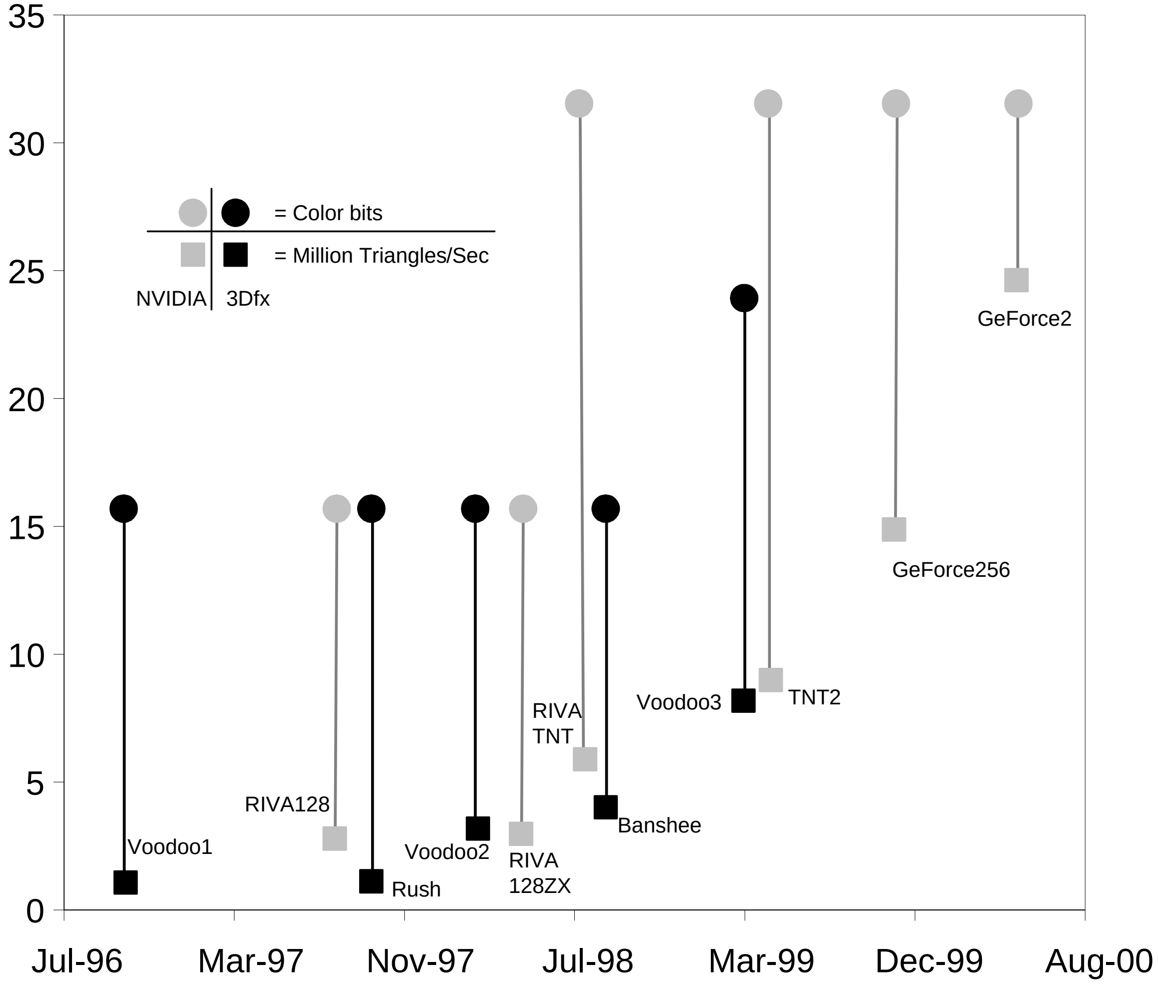


Exhibit 4—3D Chip Sales Histories

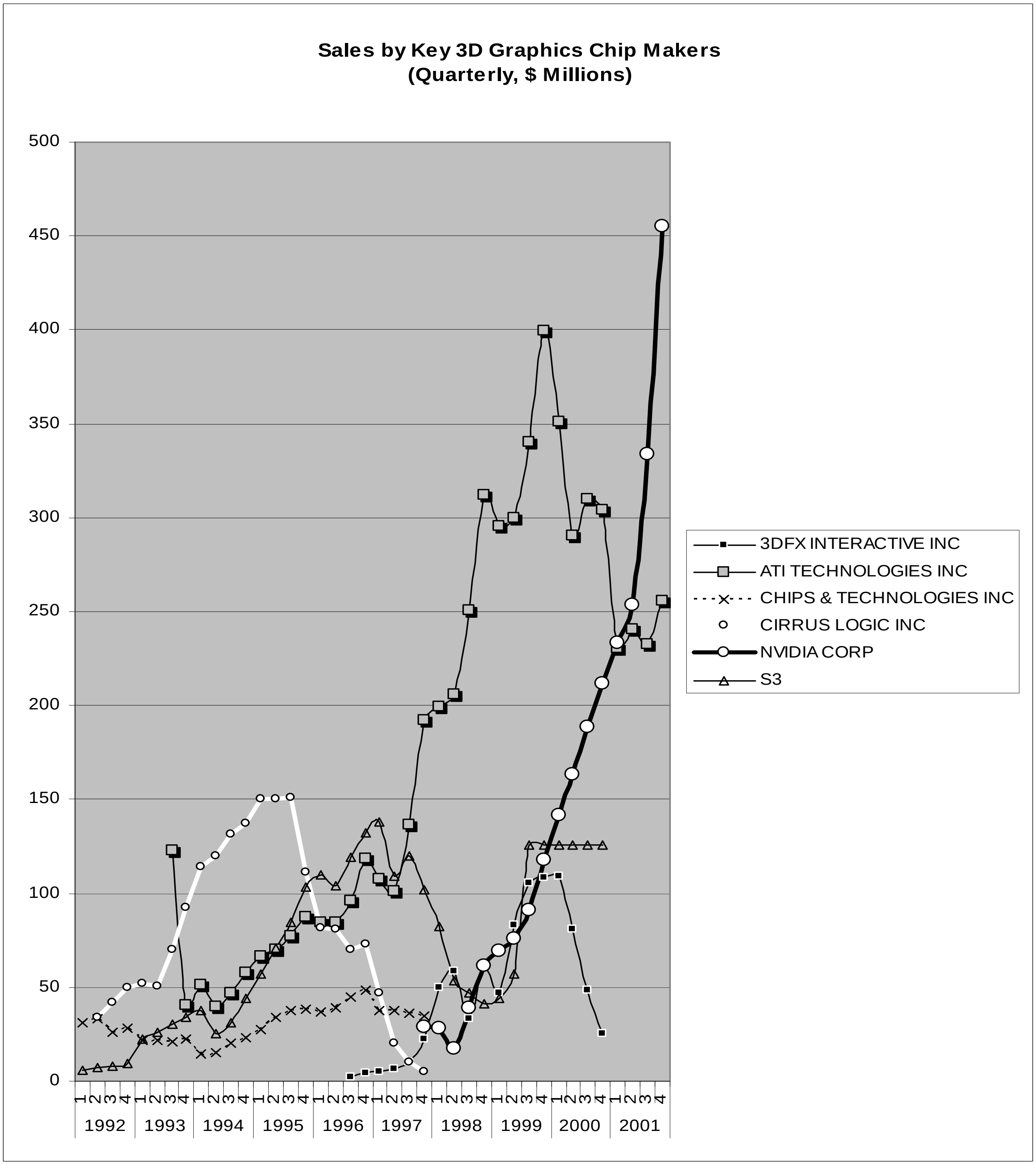


Exhibit 5—Graphics Processor Shipments by Producer and Type, 2001.

	ATI	*Intel*	*NVIDIA*	*Other*	*Total*
	Millions of Units Shipped				
Desktop	19	40	51	37	147
Desktop Discrete	18	0	47	10	74
Desktop IGP	1	40	4	27	73
Notebook	12	4	1	9	27
All Types	**31**	**46**	**52**	**45**	**174**
	Market Share of Producer by Type				
Desktop	13%	27%	35%	25%	**100%**
Desktop Discrete	24%	0%	63%	13%	**100%**
Desktop IGP	2%	55%	6%	38%	**100%**
Notebook	46%	16%	4%	33%	**100%**
All Types	18%	26%	30%	26%	**100%**
	Share of Producer Graphics Processor Shipments by Type				
Desktop	61%	88%	97%	83%	85%
Desktop Discrete	57%	0%	90%	22%	43%
Desktop IGP	4%	88%	8%	61%	42%
Notebook	40%	10%	2%	20%	15%
All Types	**100%**	**100%**	**100%**	**100%**	**100%**

Source: Various investment bankers' research reports, Wall Street Journal, case-writer estimates.

Exhibit 6—Retail Market Segments, 3D Graphics Add-In Boards

Segment	*Retail Price Range ($)*	*Average Retail Price*	*Segment Fraction of Unit Sales*	*Segment Fraction of Total Gross Profit*[32]
Basic	< 100	80	44%	18%
Low-end Mainstream	100-149	125	39%	39%
High-end Mainstream	150-249	175	9%	15%
Performance	250-350	275	7%	23%
Extreme Performance	> 350	380	1%	5%

Note: “Enthusiasts” were present in the Performance and Extreme Performance segments as well as part of the High-End Mainstream segment.

Source: Analysts’ reports and casewriter estimates.

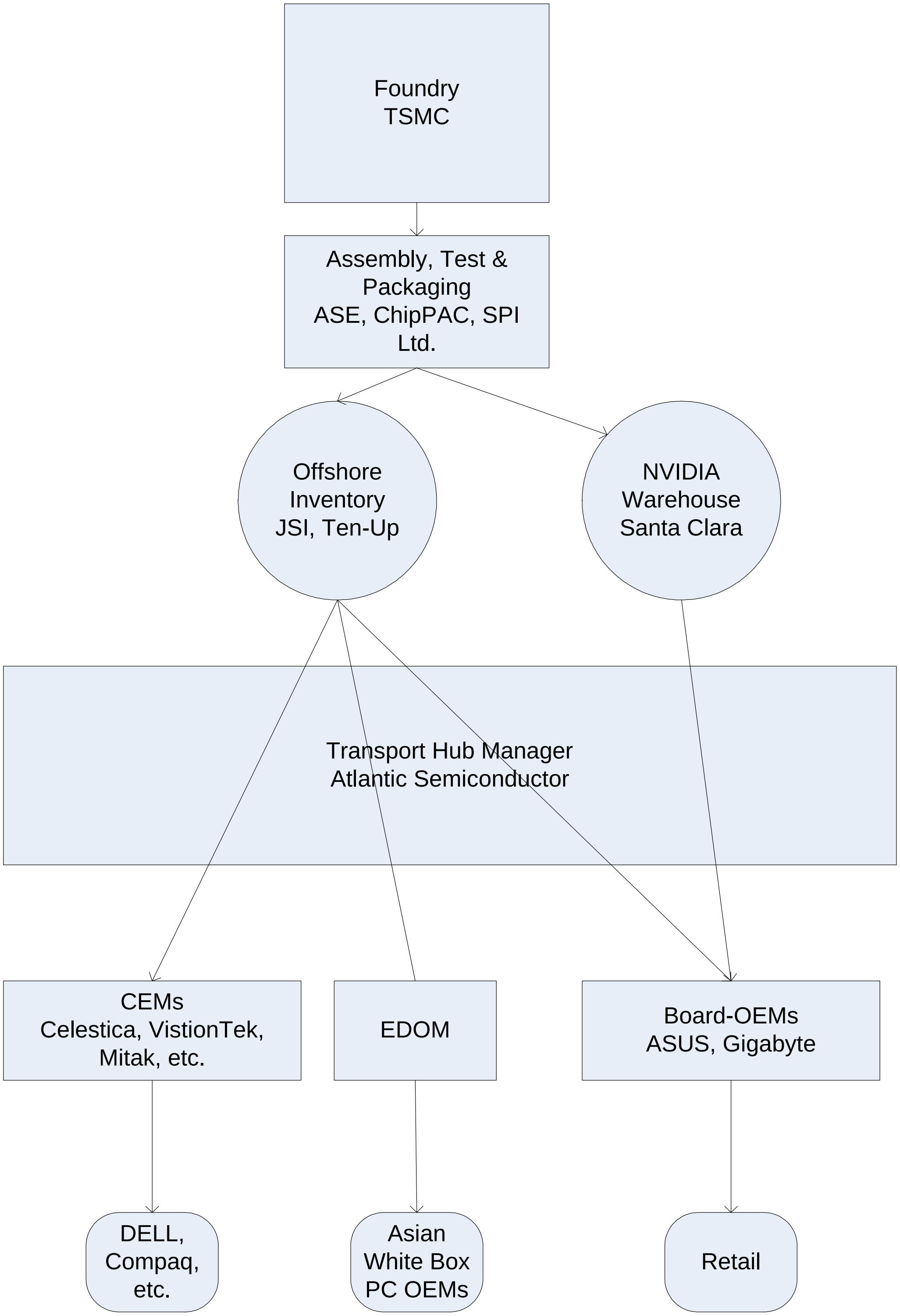
[32] Segment Gross Profit is measured for chip producers. Average gross profit, summed across all segments, was 35% in 2001.

Exhibit 7—NVIDIA Product Line in 2001

Model	*Target Market*	*Transistors (million)*	*Process Technology (microns)*	*Unit Price Range*
GeForce Graphics Processor Family				
GeForce3 Ti 500	Ultimate PC Enthusiast	57	0.15	**??**
GeForce3 Ti 200	Performance Enthusiast	57	0.15	**??**
GeForce3	Ultimate PC Enthusiast	57	0.15	**??**
GeForce2 Ti	Performance	25	0.18	**??**
GeForce2 Ultra	Performance	25	0.18	**??**
GeForce2 GTS	Performance Midrange	25	0.18	**??**
GeForce2 MX	Midrange	19	0.18	**??**
GeForce Graphics Processor Family				
GeForce2 Go	Performance	19	0.18	**??**
GeForce2 Go 200	Thin & Light High End	19	0.18	**??**
GeForce2 Go 100	Mainstream	19	0.18	**??**
NVIDIA TNT2 Graphics Processor Family				
TNT2 M64	Value	15	0.22	**??**
Vanta	Low Cost Value	15	0.22	**??**
Vanta LT	Low Cost	15	0.22	**??**
NVIDIA Quadro Workstation GPU Family				
Quadro2 Pro	Midrange 3D	25	0.18	**??**
Quadro2 MXR/EX	Entry 3D	19	0.18	**??**
Quadro2 Go	Laptop Workstation	19	0.18	**??**
NVIDIA nForce Platform Processors				
nForce 420	Mainstream	25/12	0.15	**??**
nForce 220	Value	25/12	0.15	**??**
nForce 415	Midrange/High-End	25/12	0.15	**??**

Source: Company Reports

Exhibit 8—Virtual Supply Chain



Source: Company reports and interviews.

Exhibit 9—NVIDIA Financial Statements

	Fiscal Year Ending			
	1/2002	*1/2001*	*1/2000*	*1/1999*
Income Statement Items				
Total revenue	**1,369,471**	**735,264**	**374,505**	**158,237**
Cost of revenue	850,233	462,385	232,662	109,746
Gross profit	**519,238**	**272,879**	**141,843**	**48,491**
Operating expenses:				
Research and development	153,920	86,047	46,914	25,073
Sales, general and administrative	97,185	58,697	36,312	18,902
Amortization of goodwill	12,684			
Acquisition related charges	10,030			
Discontinued use of property	3,687			
Total operating expenses	277,506	144,744	83,226	43,975
Operating income (loss)	**241,732**	**128,135**	**58,617**	**4,516**
Interest and other income	11,017	16,673	1,754	(29)
Income (loss) before income tax	**252,749**	**144,808**	**60,371**	**4,487**
Income tax expense	75,825	46,339	19,412	357
Net income (loss)	**176,924**	**98,469**	**40,959**	**4,130**
Balance Sheet Items	2001	2000	1999	1998
Cash, cash equivalents	791,377	674,275	61,560	50,257
Inventory	213,877	90,380	37631	28623
Total assets	**1,503,174**	**1,016,902**	**203,085**	**113,332**
Capital lease obligations, less current	5,861	378	962	1,995
Deferred revenue	70,193	200,000		
Long-term debt	300,000	300,000	500	
Total stockholders equity	**763,819**	**407,107**	**127,424**	**64,209**

Source: Company Reports

Exhibit 10—ATI Financial Statements

	Fiscal Year Ending			*Quarter Ending*	
	8/01	*8/00*	*8/99*	*11/01*	*11/00*
Income Statement Items					
Total Revenue	**1037.8**	**1283.1**	**1186.4**	**250**	**342.2**
Cost of Revenue	797.1	1045.5	780.6	170	259
Gross Profit	**240.7**	**237.6**	**405.8**	**80.1**	**83.2**
Operating Expenses					
Sales and marketing	75.5	86.5	64.1	20.3	21
Research and development	149.5	130.6	98.6	40.1	37.3
Administrative	37.3	31.9	33.1	8.8	8.6
Amortization	114.5	56.7	52.1	0	34.1
Operating income	**(136.0)**	**(68.2)**	**157.8**	**10.9**	**(18.8)**
Interest and other income	64.1	15.2	4.1	1.7	(0.4)
Interest expense	1.2	0.1	0.4	-	0.6
Earnings before tax	(73.1)	(53.0)	161.5	12.7	(19.8)
Income tax provision	(18.9)	16.3	54.3	1.8	(3.6)
Net income	**(54.2)**	**(69.3)**	**107.2**	**10.9**	**(16.2)**
Balance Sheet Items					
Cash and equivalents	171.5	74.8	85.5	196.9	76.4
Accounts receivable	134.9	180.4	178.5	178.3	198.6
Inventories	99	239.2	190.9	91.5	257.3
Total Assets	**848.9**	**1015**	**567.5**	**870.2**	**1025.7**
Current Liabilities	152.1	240.6	186.2	184.1	303
Equity	**696.8**	**774.4**	**381.3**	**686.1**	**722.7**

Exhibit 11—NVIDIA Graphics Processors

Product	*Core*	*Release Date*	*Process (microns)*	*Transistors (millions)*	*GFLOPS*	*TPS*	*DirectX Version*	*Technical Staff LD/PD*	*Core Speed (mhz)*	*Production Mass & Days*	*Fill Rate (mpxs)*	*Memory Speed (mhz)*	*Memory Bandwidth (ghz)*
	NV1	1995	0.5	0.75			none	1/1					
Riva 128	NV3	8/1997	0.35	3	5	3	3	1.6/3	??	??	??	??	??
Riva 128ZX	NV3	5/1998	0.25	5	7	3	5	??	??	??	??	??	??
Riva TNT	NV4	8/1998	0.25	7	10	6	6	1.7/4	??	??	??	??	??
TNT2	NV5	5/1999	0.22	9	15	9	6.1	1.5/4	??	A02/98		??	??
GeForce 256	NV10	11/1999	0.22	23	25	15	7	2.5/5	120	A03/104	480	150	4.8
GeForce2 GTS	NV15	4/2000	0.18	25	35	25	7	1.5/4	200	A03/102	1600	166	2.66
GeForce2 MX	NV11	6/2000	0.18	20	??	??	7	??	175	A01/110	700	166	2.66
GeForce2 Ultra	NV16	8/2000	0.18	25	45	31	7	??	250		2000	200	6.4
GeForce3	NV20	2/2001	0.15	55	80	30	8	3.5/6	250	A03/118	1600	230	7.36
GeForce3 Ti	NV20	10/2001	0.15	55	??	??	8	??	240		1920	250	8
XGPU	XBox	11/2001	0.15	60	80	??		1.5/7	??	A03/120	??		
In process	NV25	Spring 2002	0.15	63			8						

Note: TPS = triangles/second (millions); Technical Staff ratio is logical design/physical design. Production mask sets are lettered (A, B, etc) and numbered. A02 indicates the second variation of the first full set. A B set indicates a full respin. The "days" figure indicates the number of days from tapeout to full production.

Source: Akeley and Hanrahan Real-Time Graphics Architecture Lectures (Stanford). Chris Malachowski lecture "When 10M Gates Just Isn't Enough." Analysts' reports.