Technical Brief

Quadro vs. GeForce GPUs

Features and Benefits

Table of Contents

List of Figures

List of Tables

Purpose

NVIDIA® Graphics Processing Units (GPUs) are recognized as leading the industry. However, a question frequently asked is, "What are the major differences between the consumer-level NVIDIA GeForce™ family[1] and the workstation-class NVIDIA Quadro™ family[2]?"

This technical brief addresses that question by explaining the features and benefits of each product family and demonstrating how they relate to end-user applications for computer-aided design (CAD) and digital content creation (DCC). It covers hardware differences and application of features, such as the acceleration of the antialiased points and lines. Also addressed are the differences in the driver features—for example, the optimized OpenGL® memory management. Finally, it describes application features and enhancements provided with workstation drivers—POWERdraft, MAXtreme, and QuadroView—and shows how they benefit commonly used applications.

[1] The GeForce family of GPUs currently includes GeForce2 MX, GeForce2 MX200, GeForce2 MX400, GeForce3 Ti 200, GeForce3 Ti 500, GeForce4 Ti 4600, GeForce4 Ti 4400, GeForce4 Ti 4200, GeForce4 MX 460, GeForce4 MX 440, and GeForce4 MX 420.

[2] The Quadro family currently includes Quadro2 EX, Quadro2 MXR, Quadro2 Pro, Quadro DCC, Quadro4 550 XGL, Quadro4 700 XGL, Quadro4 750 XGL and Quadro4 900 XGL.

Why Choose a Workstation?

The term *professional workstation* implies many things to many people. However, it usually translates to expectations of high quality, excellent reliability, responsive support, and high performance. Not to mention leading-edge technology—although not at the expense of quality and reliability.

These expectations exist because there is a specific goal in mind by workstation users—a goal that is ultimately critical to success. The goal may be designing a revolutionary gadget that makes millions, or it may be creating key animated scenes in the next blockbuster film. Each goal has a level of investment and expectation of success. The quality, reliability, support, and performance that define a workstation ensure this success.

When NVIDIA introduced the GPU in 1998, it created a discontinuity in graphics price performance and turned the traditional workstation marketplace upside down. Although NVIDIA offers the workstation-branded Quadro and the consumer-branded GeForce GPUs, many consumers question the benefit of a professional workstation over a consumer PC. The workstation/consumer distinction was clouded by that fact that, prior to the price-performance discontinuity, a workstation cost $10K to $20K, whereas a PC cost under $5K. With the introduction of the NVIDIA GPUs, these costs dropped dramatically and it became more difficult to distinguish the two classes of systems on price alone.

This technical brief explains in detail the range of features and benefits offered by the Quadro family of workstation GPUs, and places them into the context of the professional user. It highlights the hardware features and benefits of the workstation GPU family over the consumer GPU family, and provides details on application support, hardware and driver features, and value-added applications targeted at specific workstation markets.

Overview of Workstation Products

NVIDIA workstation-class GPUs define the standard for professional 3D performance. To cater to the needs of different users, NVIDIA offers several versions that tailor to the specific requirements of those applications.

Table 1 summarizes the workstation products and provides a brief overview of their performance and features:

Table 1 Features of Workstation GPU Family

	Quadro2 EX	Quadro2 MXR	Quadro DCC	Quadro2 Pro	Quadro4 550 XGL	Quadro4 700 XGL / 750 XGL	Quadro4 900 XGL
Target Audience	Entry level	Entry level	Feature	Mid-Range	Entry level	Mid-Range	High End
Graphics Core	256 bit	256 bit	256 bit	256 bit	256 bit	256 bit	256 bit
Memory Interface	128-bit SDR	128-bit SDR	128-bit DDR	128-bit DDR	128-bit DDR	128-bit DDR	128-bit DDR
Memory Bandwidth	2.7GB/s	2.9GB/s	6.4GB/s	6.4GB/s	6.4GB/s	8.8GB/s	10.4GB/s
Memory Capacity	32MB	32MB	64MB	64MB	64MB	64MB/ 128MB	128MB
Triangle Rate	21 million tris/s	25 million tris/s	25 million tris/s	31 million tris/s	34 million tris/s	54 million tris/s	60 million tris/s
Fill Rate	350 million pixels/s	400 million pixels/s	800 million pixels/s	1 billion pixels/s	1 billion pixels/s	1.1 billion pixels/s	1.2 billion pixels/s

Each GPU in the workstation family offers more features than the GPUs in the consumer family. The next section describes these features in detail.

Workstation Features

Antialiased Points and Lines

Many workstation applications, particularly in the CAD market, offer the option of using antialiased points and lines. With this option turned on, component edges can be viewed as precisely as possible without encountering the aliasing artifacts that are associated with lines displayed on a rasterized display.

In Figure 1, a series of close-up screen shots show antialiasing turned on and off for the major CAD applications:

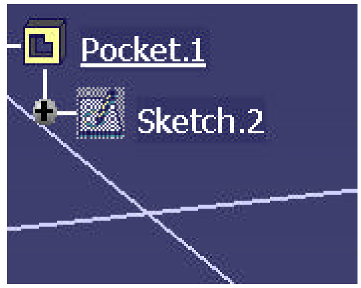
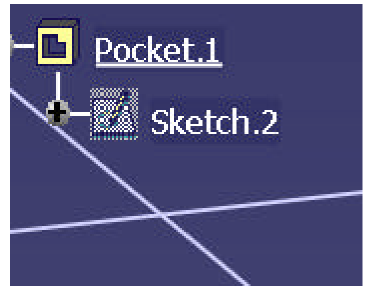
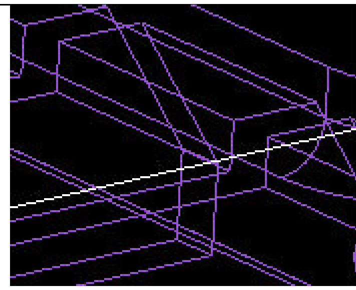
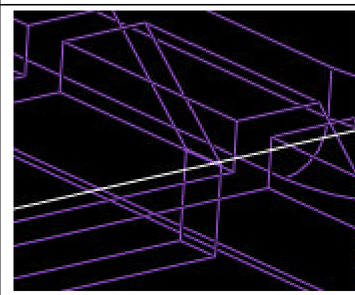
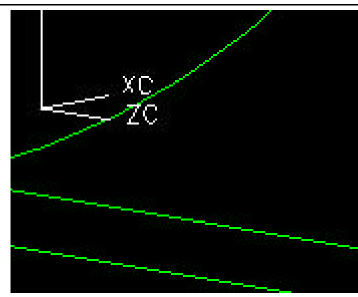
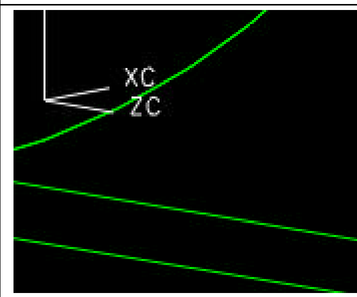
	Non-Antialiased	Line Antialiased
CATIA V5 Turn on Line Antialiasing: Tools/Options/Display/Visualization		
ProENGINEER 2000i2 Turn on Line Antialiasing: View/Model Display		
Uigraphics Version 17 Turn on Line Antialiasing: Preferences/Visualization/Visual		

Figure 1. Points and Lines Enabled in Major CAD Applications

To address this feature in professional workstation applications, the Quadro GPU family supports antialiased lines in hardware. The result? When antialiased points and lines are used on the Quadro family of GPUs, performance is noticeably higher than on the GeForce family of GPUs.

To quantify the productivity advantages of antialiased points and lines, Figure 2 shows the relative performance of a Quadro2 EX compared with a GeForce2 MX for antialiased lines of various sizes. Both GPUs have comparable core and memory clock rates and use a 128-bit SDR memory interface:

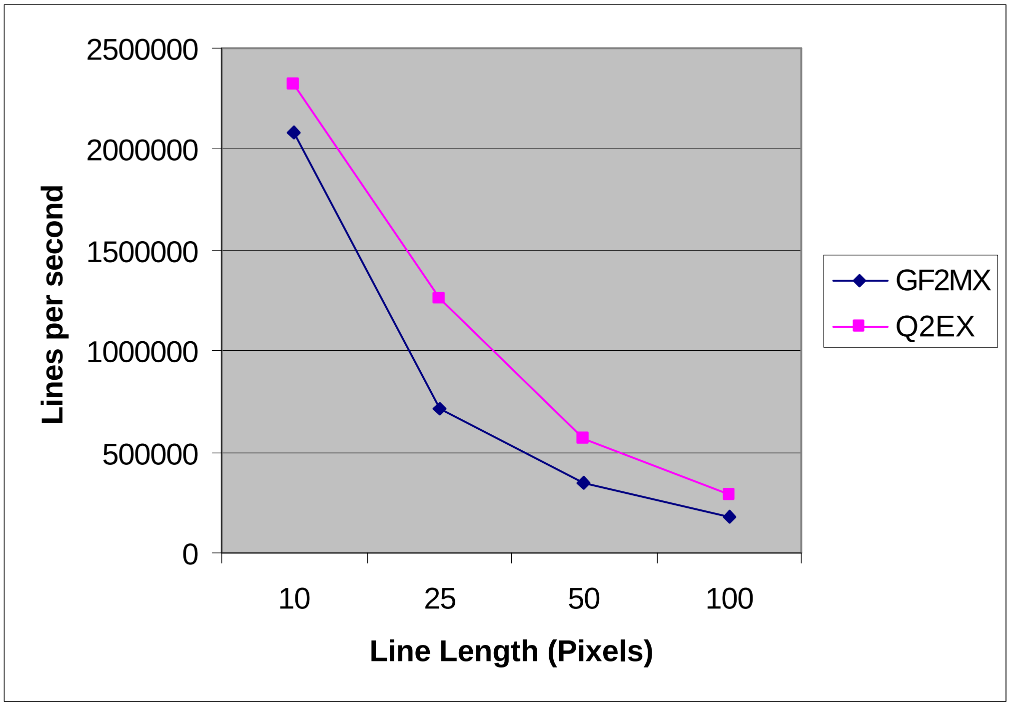


Figure 2. Antialiased Line Performance

The performance advantage of the Quadro GPUs is clear—improved performance when using applications that take advantage of antialiased points and lines. This lets professionals work with improved visual quality, while not sacrificing performance and interactivity. For a CAD designer working in wire-frame—which is a significant amount of a user workflow—high-quality lines make the difference between a successful design and an exercise in frustration.

The increase in productivity afforded by the quality and performance of antialiased lines is a clear advantage of Quadro workstation GPUs.

Logic Operations

Another hardware feature difference between NVIDIA's workstation and consumer GPUs is support for OpenGL Logic Operations. (Refer to Figure 3 for an overview of the OpenGL pipeline.) Logic operations are the final stage in the pipeline and are applied to incoming fragments and affect how, and if, they are written into the frame buffer.

For a full explanation of the stages of the OpenGL pipeline, refer to *The OpenGL Programming Guide.*

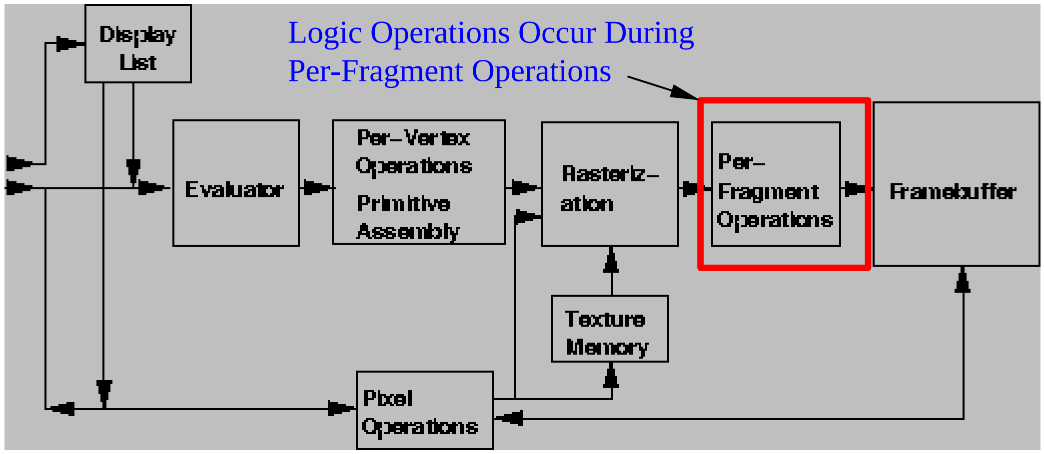


Figure 3. The OpenGL Pipeline

Logic operations are often used by workstation applications in mechanical computer-aided design (MCAD) and DCC markets. They're used to draw on top of a 3D scene to make specific features visible without significantly changing or complicating the existing drawing functions or adversely affecting performance.

Example of CAD Application

A good example is 3D CAD packages that highlight features or components when the cursor is moved over a model or assembly.

Figure 4 shows an example using CATIA V5, where the highlighted feature is orange:

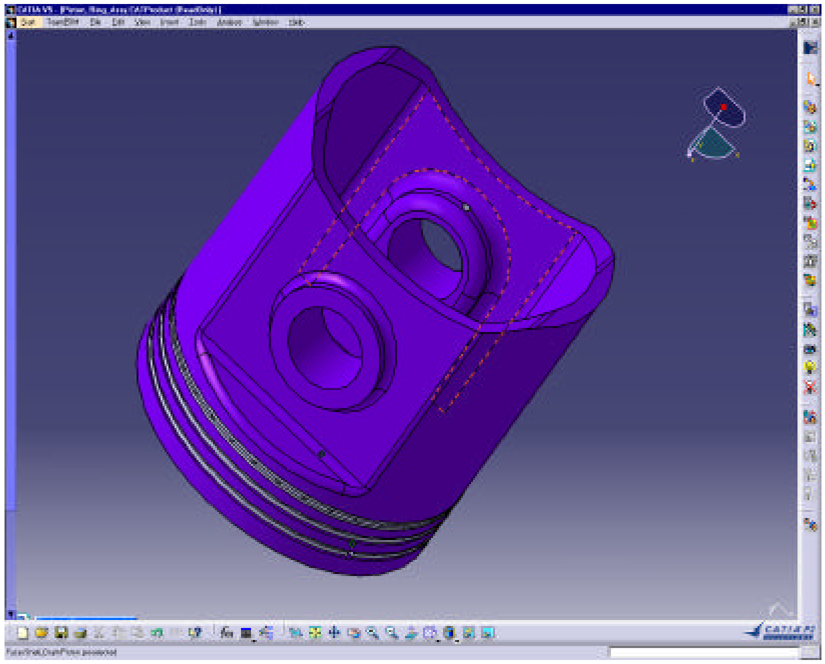


Figure 4. CATIA V5 Screen Shot Showing Highlighted Feature

The feature highlighted by the cursor (Figure 4) is drawn in an orange outline so it can be dynamically identified by the user before a selection or pick is made.

Example of DCC Application

A similar example in a DCC application is demonstrated where the XOR logic operation is used to draw sophisticated cursors, such as those in the paint operation of Alias/Wavefront's Maya application. Refer to "Overlay Plane Support," which shows a screen shot of Maya with the paint cursor used. The XOR logic operation draws the cursor on top of the 3D scene for applications that do not support overplay planes.

Table 2 demonstrates the performance advantage of the Quadro GPU family over the GeForce GPU family for hardware logic operations. It is a comparison of pixel fill rates for the Quadro2 EX and GeForce2 MX when 100-pixel, lit-shaded triangles are drawn. The 100-pixel triangles were used to minimize the contribution of transform-and lighting-calculations on the fill-rate results:

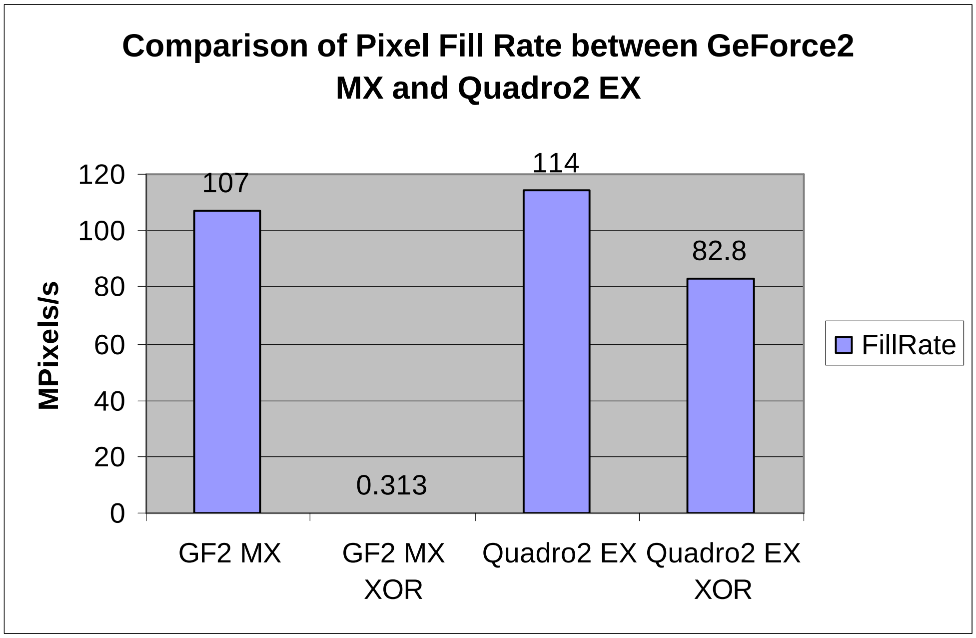


Table 2. Comparison of Pixel Fill Rates

If the XOR logic operation is enabled, the performance drop of the Quadro2 EX is minimal when compared to that of GeForce2 MX. In professional applications where logic operations are used, this equates to significant improvement in performance—a definite productivity benefit.

Clip Regions

During a typical workflow, workstation applications pop up many windows for menus or alternative views of components or scenes. These applications often occupy the full screen, so the result is many overlapping windows. At first, this may be of little consequence to the user; however, depending on how they are handled by the graphics hardware, overlapping windows may noticeably affect visual quality and graphics performance.

NVIDIA's GPU architecture manages the transfer of data between a window and the overall frame buffer by clip regions. When a window has no overlapping windows, the entire contents of the color buffer can be transferred to the frame buffer in a single, continuous rectangular region. However, if other windows overlap the window, the transfer of data from the color buffer to the frame buffer must be broken into a series of smaller, discontinuous rectangular regions. These rectangular regions are referred to as "clip regions."

Figure 5, a screen shot from ProENGINEER, details the window arrangements and highlights how they affect the number of clip regions required for transferring data from the color buffer to the frame buffer:

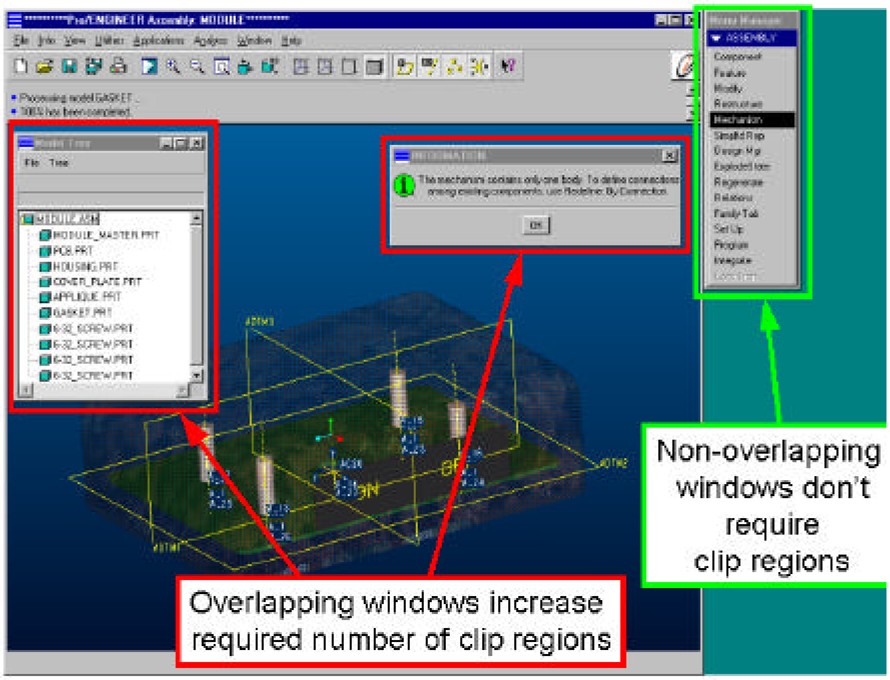


Figure 5. Window Overlap in ProENGINEER and Its Effect on the Number of Clip Regions

To provide the best overall graphics performance, the transfer of data using clip regions is hardware-accelerated. It's not possible to support all hardware-accelerated clip regions, however, so when the overlapping windows require more clip regions than are accelerated by hardware, a default software path is used. As expected, when a software path is used for clip regions, the speed of the transfer between the color buffer and frame buffer is affected and this in turn impacts overall graphics performance.

Most consumer applications and games don't create many pop-up windows, so the GeForce family of GPUs supports only one clip region, whereas the Quadro family support up to eight clip regions.

Figure 6 shows how hardware accelerated clip regions affect overall graphics performance. Figure 7 is a comparison of peak (1-pixel) lit, smooth-shaded, depth-tested triangle performance between the Quadro and the GeForce solutions for two scenarios: one with overlapping windows and one without overlapping windows. Since peak triangle performance represents the maximum capability of the GPU to process and draw triangles, any detrimental effect on this result arising from overlapping windows demonstrates the impact of fewer hardware-supported clip regions.

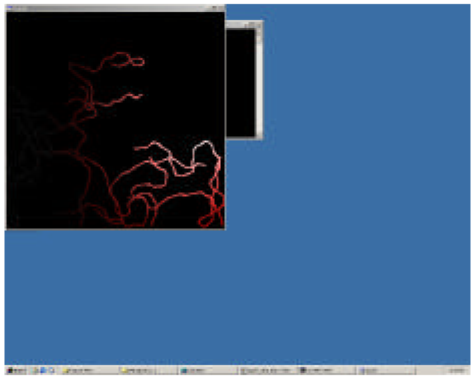
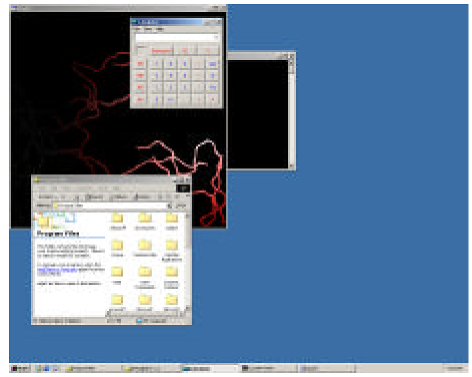
	
No Overlapping Windows	Overlapping Windows

Figure 6. Comparison of Peak (1-Pixel) Lit, Smooth-Shaded, Depth-Tested Triangle Performance

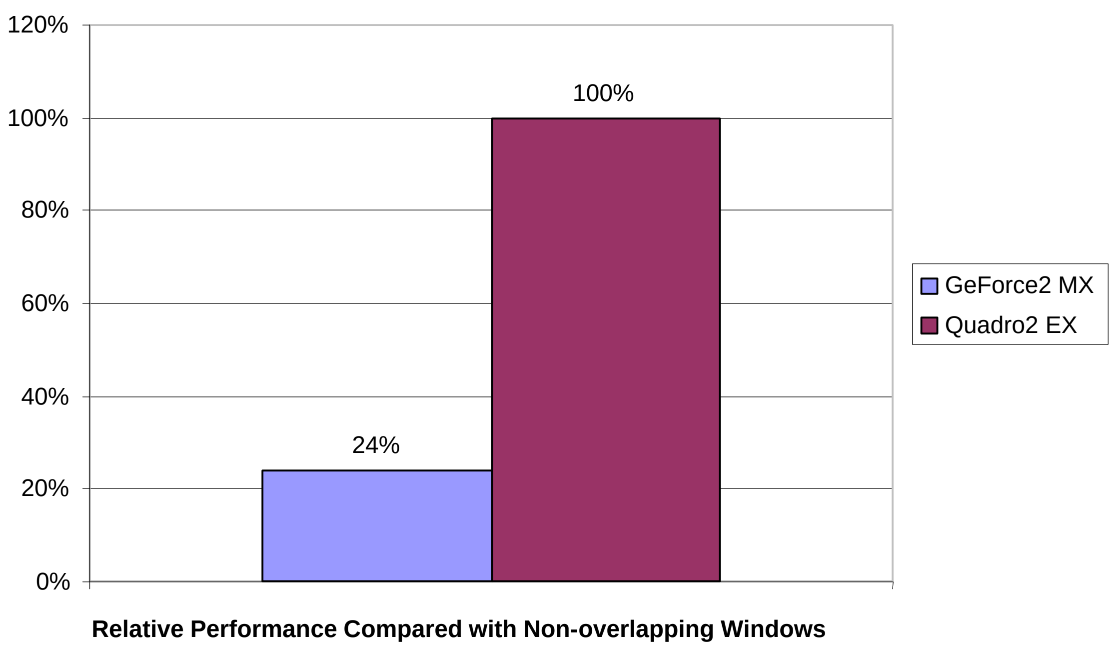


Figure 7. Effect of Overlapping Windows on Peak Triangle Performance

Hardware-Accelerated Clip Planes

In many situations, understanding the relationship between components in a complex 3D can be eased by using clip planes. Clip planes allow sections of the geometry to be cut away so the user can look inside solid objects. Looking inside objects is particularly useful for visualizing assemblies that comprise hundreds or thousands of components. For this reason, many professional CAD applications, including ProENGINEER, allow users to define clip planes.

The Quadro family of GPUs supports clip-plane acceleration in hardware—a significant performance improvement when they are used in professional applications. Tests 6 and 10 of the SPECopc Viewperf MedMCAD-01 Test define a clip plane, and are useful for quantifying the performance benefits of clip-plane support on the Quadro2 family. Figure 8 shows the relative performance of the Quadro2 EX and GeForce2 MX on the MedMCAD-01 benchmark:

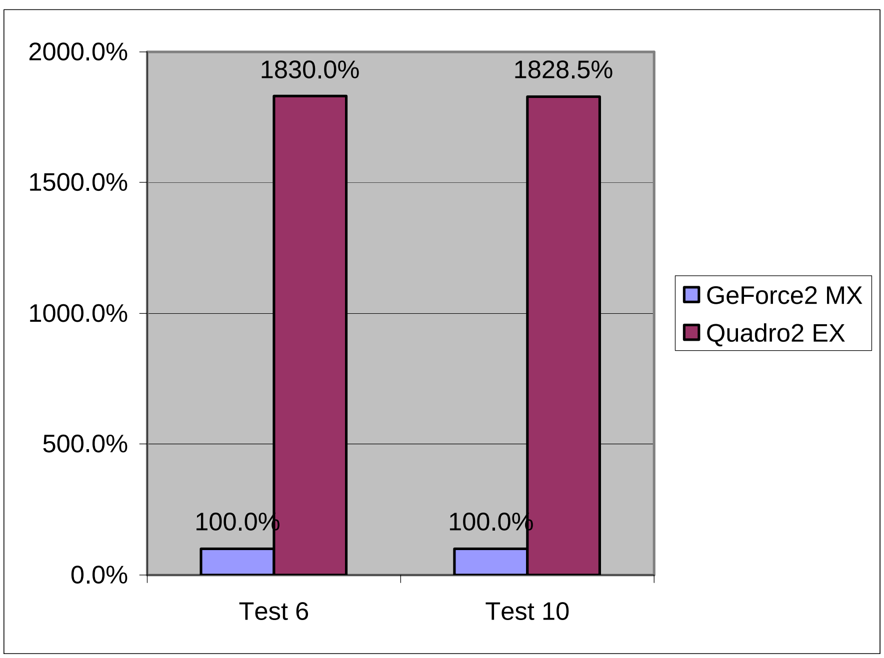


Figure 8. Clip Plane Performance (Using SPECopc MedMCAD-01 Tests)

The results of Test 6 and 10 in Figure 8 clearly show that the hardware acceleration of clip planes significantly benefits Quadro2 EX when compared with GeForce2 MX.

Quadro Memory Management Optimization

Another feature offered by the Quadro family of GPUs is Quadro memory management optimization, which efficiently allocates and shares memory resources between concurrent graphics windows and applications. In many situations, this feature directly affects application performance and so offers demonstrable benefits over the consumer-oriented GeForce GPU family.

Normal Memory Demands

NVIDIA's GPU architecture uses a common pool of dedicated high-speed graphics memory known as a Unified Memory Architecture (UMA). UMA stores various graphics buffers such as the frame buffer, textures, and data. Compared with competitive products—for example, 3D Labs' Wildcat II graphics that employ separate memory for frame buffer, textures, and display lists—NVIDIA's approach is clearly advantageous because it maximizes the use of hardware resources. It's also worth noting that, while the Wildcat II 5110 claims 64MB of dedicated texture memory, the memory is divided between two hardware pipelines. In reality, Wildcat II 5110 has only a maximum of 32MB of resident texture memory available to applications, which is significantly less than might appear at first sight.

The superiority of NVIDIA's UMA architecture is evident when display configurations do not fully consume frame buffer memory—which is typical in normal usage. Instead of the remaining frame buffer memory being wasted because it is unused, UMA allows it to be used for other buffers and textures.

Consider a typical 1280x1024 desktop with 32-bit color. A full-screen double-buffered application that uses 24-bit Z consumes approximately 1280x1024x2x (4bytes+3bytes), or 18MB of memory. Using 64MB of dedicated frame buffer memory, therefore, would equate to wasting real dollars because less than half is used. Even 64MB of frame buffer split between two separate graphics pipelines, equating to 32MB each, is overkill because only a little over half is used.

NVIDIA's UMA avoids waste by letting 'spare' memory be used for textures.

High Memory Demands

In certain circumstances, applications require much more memory, such as when they use stereo. Quad-buffered stereo dramatically increases memory requirements because it usually requires twice the memory to provide the additional buffers (refer to "Quad-Buffered Stereo"). If quad-buffered stereo were used in the previous example, it would double the memory requirements to ~36MB. 32MB of dedicated texture memory is insufficient to accommodate this; however, 64MB of dedicated frame buffer uses just over half, which again equates to a waste of money. NVIDIA's UMA is the best of both worlds: It doesn't waste expensive hardware to cater to occasional situation, but it accommodates normal demands for large amounts of resource.

Another feature that dramatically increases memory requirements is full-screen antialiasing, which is often used in visual simulation applications. Increased memory demands also occur when several graphics windows or applications run concurrently, which happens when using professional applications. The Quadro memory management optimization is very important to professional applications because it efficiently accommodates large demands, but does not waste resource or memory when they are not needed.

Insufficient memory resources are not an issue for the Quadro2 Pro GPU, which has 64MB of graphics memory. However, it can be for the entry-level Quadro2 EX and mainstream Quadro2 MXR workstation GPUs, which both use 32MB of graphics memory. For these solutions the Quadro memory management optimizations ensure that all available graphics memory is used efficiently, preventing potential performance degradations or loss of functionality because of exhausting graphics memory. This is important for professional applications in both the CAD and DCC markets that use several graphics windows simultaneously, as well as define and use many textures.

To demonstrate the performance advantages of the Quadro memory management optimization, three scenarios were designed that place varying demands on required graphics memory:

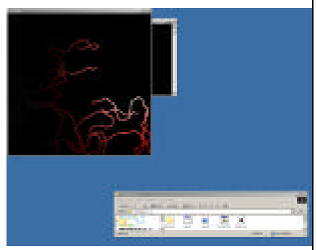
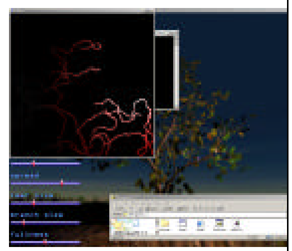
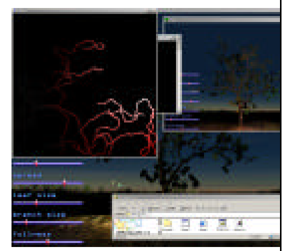
- The first scenario ran the GLperf application alone to measure both the peak textured triangle rate corresponding to 1-pixel, lit, smooth-shaded, depth-tested, textured (64x64 RGB tri-linear modulated).
- The second scenario ran GLperf concurrently with the NVIDIA tree demo maximized to full-screen resolution.
- The third scenario added a further instance of the tree demo run in a separate window.

The tree demos were deliberately paused before the test to prevent invalidating the results, which would occur if CPU and graphics memory resources were consumed. Likewise, all windows except the Glperf were pushed back on the window stack. This was done so that issues—such as those that arise from the number of hardware-accelerated clip regions supported through windows overlapping the Glperf window—would not inadvertently affect performance.

To increase memory requirements, each scenario was performed at screen resolutions of 1280x1024 and 1600x1200 using 32-bit color, and the effects on peak textured triangle rate were compared. By limiting the triangle size to 1 pixel and fixing the GLperf window to 600x600, any influences arising from fill-rate limitations were avoided.

Table 3 is a screen shot of each scenario and summarizes the results:

Table 3 Influence of Quadro Memory Management Optimization on Textured Triangle Rates

			
	Scenario 1 Single instance of GLperf application.	**Scenario 2** GLperf running concurrently with full-screen tree demo.	**Scenario 3** GLperf running concurrently with full-screen tree demo and additional tree demo.
GeForce2 MX 1280x1024	100%	100%	100%
Quadro2 EX 1280x1024	100%	100%	100%
GeForce2 MX 1600x1200	100%	32%	16%
Quadro2 EX 1600x1200	100%	100%	100%

Note: Percentage values correspond to the peak-textured triangle rate of 1 pixel, lit, smooth-shaded, depth-tested, textured (64x64 RGB tri-linear modulated) and normalized to the Quadro2 EX performance measured under Scenario 1 at screen resolution 1280x1024.

At a screen resolution of 1280x1024, the memory requirements of each scenario has minimal impact on performance. With the screen resolution set to 1600x1200, however, the increased requirements affect performance, and the peak performance of the GeForce2 MX begins to degrade. The Quadro memory management optimization clearly allows the Quadro2 EX to remain unaffected.

For professional DCC or CAD applications, which consume significant amounts of texture memory or open many separate 3D graphics windows, the Quadro Memory Optimization offers significant advantages and ensures optimal performance. These benefits are also evident in professional applications that offer quad-buffered stereo views or take advantage of full-screen antialiasing.

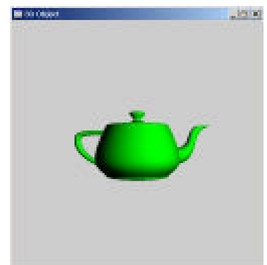
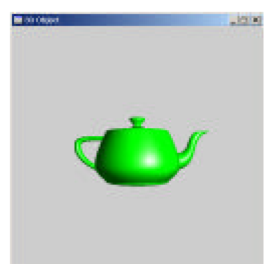
Two-Sided Lighting

Computer graphics use triangles or polygons to describe real-world objects. Three-dimensional vertices are often used to define the triangles or polygons and, depending on the realism of the scene, normal vectors may specify the orientation of the object surface at each vertex. To generate a realistic image, the vertices are transformed from the 3D coordinate system of the object into the 2D coordinate system of the screen.

In addition, the color at each vertex is determined from the lighting equations that model the effect of light in the real world. Lighting equations use three components (Table 4) to model how objects appear in the real world:

Lighting Components

Table 4 Components That Represent Lighting Effects in the Real World

Ambient	Ambient lighting doesn't depend on the angle of the object to the viewer or lights. Example: Objects that are visible on a cloudy day.	
Diffuse	Diffuse lighting illuminates objects depending on their orientation to a light source, but not depending on the angle at which they are viewed. Example: The sun shining behind the viewer. As an object is turned in front of the viewer, it often appears to change between its true color (when the largest side is perpendicular to the sun) to almost black (when the largest side is nearly parallel to the sun).	
Specular	Specular lighting illuminates objects depending on their orientation to the light source and to the viewer. Example: The glint on a car windshield or paintwork on a sunny day. As either the car or the viewer moves, the position of the glint moves.	
All three components are combined to generate a realistic image		

To maximize realism, the relative contributions from the ambient, diffuse, and specular components are adjusted. The default OpenGL contributions are 20 percent for the ambient, and 100 percent for both the diffuse and specular components.

As these default contributions show, the diffuse and specular components are usually high in comparison to the ambient component. Unfortunately, these relative proportions and the assumptions in the lighting calculations can cause some visual issues.

Visual Issues

When an object is rotated in three-dimensional space, the normals for an individual triangle or surface will, at some orientation, point away from the light source. The lighting equations use the dot product between the light vector and the surface normal to calculate the diffuse and specular components; in this situation, both the diffuse and specular components drop to zero. In the real world, this is analogous to holding a newspaper up to block the sun. When you do, it's impossible to read the newspaper.

The diffuse and specular lighting components are usually greater than the ambient component, so when they become zero the surface or triangle becomes dim, or even disappears, depending on the lighting settings. The effect is that during dynamic rotation of an object, parts of the object may appear and disappear—or at least become very dull—when viewed from different angles. Clearly, this isn't representative of the real world.

Solution for Visual Issues

The way to overcome this limitation is to use two-sided lighting. When two-sided lighting is enabled, instead of using the dot product of the normal and light vector to calculate the diffuse and specular components, the lighting calculation uses the magnitude of the dot product. This approach prevents the diffuse and specular components from dropping to zero when the surface normal points away from the light source.

As a result, these "backward-facing" triangles remain visible through all viewing angles. In many situations, as in CAD applications, objects are created as solids. This means that the backward-facing triangles are rarely seen because they only exist on the inside of an object.

In other situations, though, objects are not created as solids, and back-facing triangles are visible. In these situations, two-sided lighting is used to prevent the surface from disappearing when viewed from certain angles.

The effects of two-sided lighting are demonstrated in Figure 9:

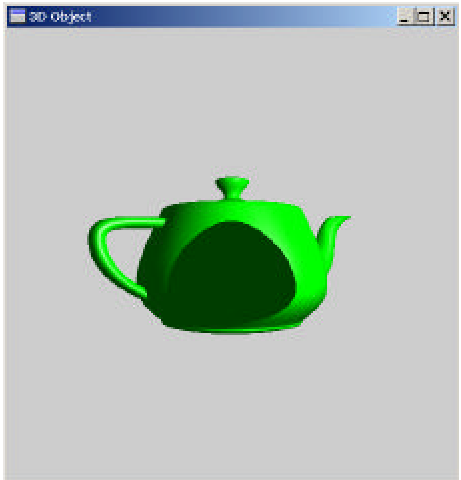
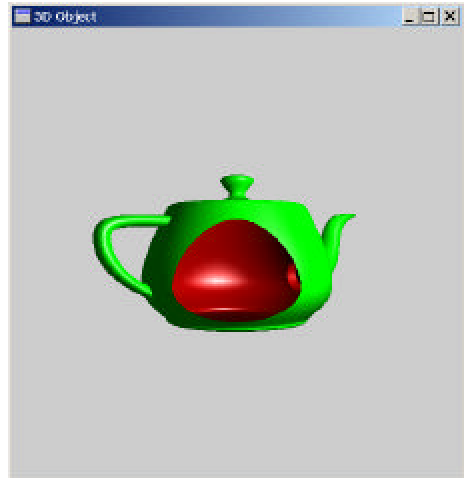
	
When two-sided lighting is not enabled, back-facing polygons (as shown by the inside of the teapot) are only illuminated by the ambient lighting.	When two-sided lighting is enabled, back-facing polygons are illuminated by all components (ambient, diffuse, and specular). Note: The inside of the teapot is deliberately red so that it can be distinguished from the outside.

Figure 9. The Effects of Two-Sided Lighting

The teapot on the left in Figure 8 shows a clipping plane that cuts the front away so that the inside is visible. Because the inside of the teapot exposes back-facing triangles, only the ambient component illuminates these triangles. As a result, the triangles are relatively dark.

The image on the right shows the effect of enabling two-sided lighting; clearly, the back-facing polygons show diffuse and specular components. (Note: The inside of the teapot is purposely colored red to distinguish it from the outside.)

One of the downsides to using two-sided lighting is (depending on GPU architecture) that more calculations may be performed at each vertex and so there may be degradation in overall graphics performance.

To quantify the potential performance implications of NVIDIA GPUs in competitive systems, Figure 10 shows results of subtests in the SPEC Pro/Designer (ProCDRS-03) benchmark that enable two-sided lighting. Both solutions were tested in an IBM Intellistation M Pro system with a P4 1.7GHz CPU. The NVIDIA Quadro2 Pro GPU and ATI Fire GL4 solutions are compared.

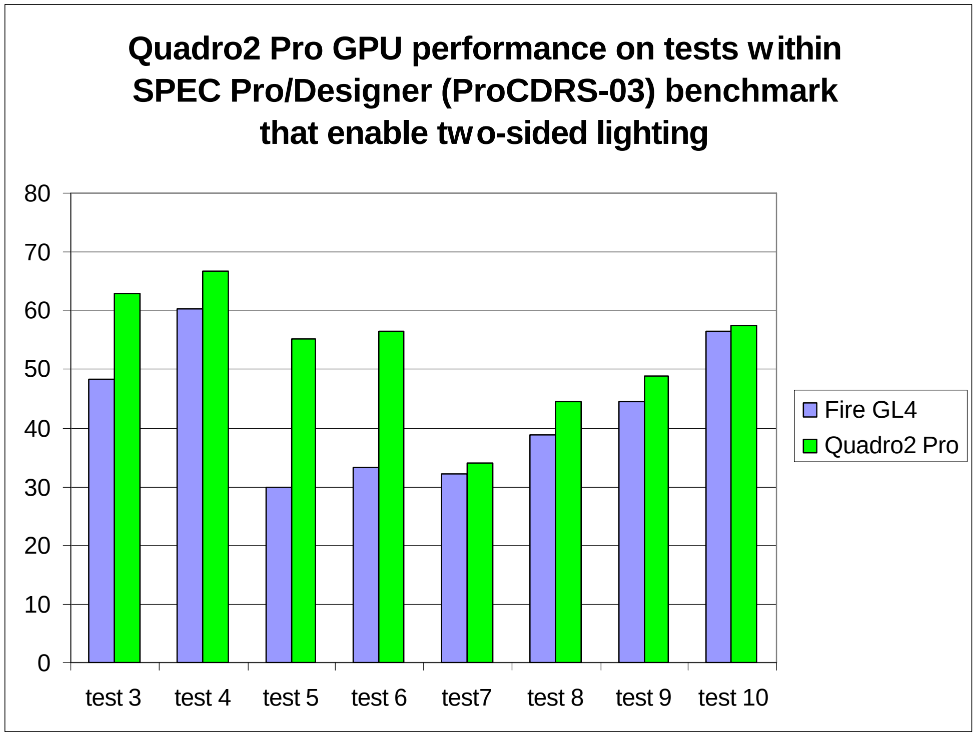


Figure 10. Relative Performance When Two-Sided Lighting Is Enabled

Figure 9 demonstrates that the architecture of the NVIDIA Quadro2 Pro GPU performs superiorly across all tests.

Overlay Plane Support

The user interfaces of many professional applications often requires elements to be interactively drawn on top of a 3D model or scene.

Cursor Issue

The most obvious example is the cursor, which is drawn in front of any 3D object or window. The cursor usually has specific dedicated hardware that allows its movement to be interactive and independent of other graphics elements on the screen.

The tradeoff for this, however, is that the cursor size is typically limited to around 32x32 pixels. A larger image invokes a software path, which noticeably affects performance and interactivity.

Pop-Up Menu Issue

Another example of a user interface element is a pop-up menu, which lets users select context-sensitive functions, depending on the current task. Unfortunately, when these menus pop up in front of an OpenGL window, they cause the contents of the window beneath to become "damaged." Since OpenGL windows typically store lots more information at each pixel than just the color—for example depth, alpha, and stencil information—damage from the pop-up windows can noticeably affect performance. That's because pop-up data is temporarily stored and recovered.

These user-interface elements usually need to be interactive as well as drawn on top of 3D models or scenes. A common example is a simple rectangle that can be stretched or "rubber-banded" over objects. However, these user-interface elements can't take advantage of dedicated cursor hardware, so drawing the elements in the main 3D graphics window can significantly complicate program architecture and affect performance.

Solution

While there are ways to overcome these issues, such as using the OpenGL XOR logic operation (see "Logic Operations"), most professional applications use overlay planes. Overlay planes let items be drawn on top of the main graphics window without damaging the contents of the windows beneath. Windows drawn in the overlay plane can contain text, graphics, and so on—the same as any normal window. However, the number of bits available to store color values is usually more restricted than in the main graphics window. Even so, the performance advantages and flexibility afforded by the use of overlays significantly outweighs the limitation in available color depth.

The way overlay planes typically work is to support a transparency bit, which when set, allows pixels underneath the overlayed window to show through. Creating pop-up menus in the overlay planes, therefore, prevents damage to the main graphics window and therefore improves performance. Likewise, clearing an overlayed window to the transparency bit and then drawing graphics within it allows user-interface items to be drawn over the main graphics window. Clearing and redrawing only the overlayed window is significantly faster than redrawing the main graphics window. This is how animated user-interface components can be drawn over 3D models or scenes.

A good example of this user interface component is the brush outline in AliasWavefront's Maya application. In Figure 11, which shows a screen shot that illustrates the brush outline feature, the red lines of the brush are drawn in overlay planes.

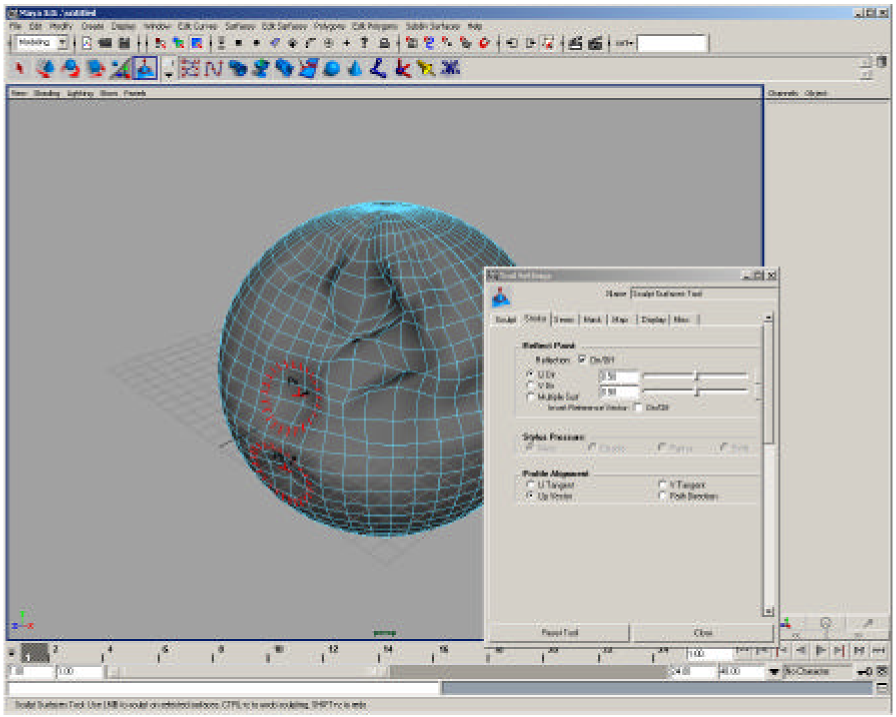


Figure 11. Brush Outline Feature in Maya That Uses Overlay Planes

Support for overlay planes is exclusive to the Quadro2 family of workstation GPUs and allows them to meet the needs of professional applications. Overlays are not supported on the GeForce family of GPUs.

It's also interesting to note that some X servers on Linux operating systems can be configured to place pop-up desktop components within the overlay planes as well. In this situation, hardware overlay plane provide a definite productivity aid in many aspects of user workflow.

Quad-Buffered Stereo

The Quadro GPU family supports quad-buffered stereo; the GeForce GPU family does not.

Many professional applications let users view models or scenes in three dimensions, using a stereoscopic display. The application generates separate images from the left and right eye perspective and both are alternately displayed. Special glasses with an LCD shutter in front of each eye are synchronized to the graphics card so that when the left eye image is displayed, the right eye shutter is closed. Similarly, when the right eye image is displayed, the left eye shutter is closed. This way, each eye receives the correct perspective and the object appears to have true depth extending in or out of the monitor. For optimum interactivity, it's important to maintain the highest screen refresh rate, because each eye is updated at half the monitor refresh rate.

The preferred way to implement stereo in professional applications is through OpenGL quad-buffered stereo. Quad-buffered stereo provides four buffers to the application—front-left, back-left, front-right, and back-right—that correspond to double-buffered left and right views. When it creates a graphics window, the application checks the hardware (via the OpenGL call glGetBooleanv) for stereo support. Likewise, to select the appropriate buffer—typically GL_BACK_LEFT or GL_BACK_RIGHT—the OpenGL call glDrawBuffer is called with the appropriate argument.

For more details on programming stereo applications, refer to [3] and [4], which provide useful background information, and references [5] and [6], which provide code examples. Figure 16 (in "Application Productivity Tools") shows a screen shot from 3D Studio Max that uses NVIDIA's MAXtreme plug-in driver to enable stereo support in the main viewing window.

Stereo support on the Quadro GPU family significantly benefits professional applications that demand stereo viewing capabilities.

Pentium 4 Optimizations

With the Pentium 4 microprocessor family, Intel introduced a series of architectural improvements that benefit performance. These improvements include the Streaming SIMD Extensions 2 (SSE2) instruction set—a further development of the MMX and SSE instruction sets—and the Intel NetBurst microarchitecture.

The SSE2 instruction set allows developers more flexibility and capability for improving the performance of application. This is especially true for applications that are inherently parallel and exhibit frequent, localized memory accesses. These accesses are particularly true for 3D graphics and multimedia applications, as well as for many professional workstation applications. The SSE2 instruction set also provides cache ability and memory-ordering instructions that can improve cache use and application performance.

NetBurst

Intel's NetBurst microarchitecture supports existing IA-32 applications while allowing operation at high clock rates, and provides performance scalability for higher clock rates in the future.

The key design considerations for NetBurst were:

- Create a deeply-pipelined architecture to enable high clock rates with different parts of the CPU running at different speeds.
- Optimize for frequently executed instructions so that on these instructions latencies are low and execution efficiency are high; thus, overall throughput is maximized.
- Minimize the impact of stall penalties by techniques such as parallel execution, buffering, speculation, and out-of-order execution.

The architecture comprises three main components: the in-order front end; the out-of-order execution core, and the in-order retirement unit (Figure 12):

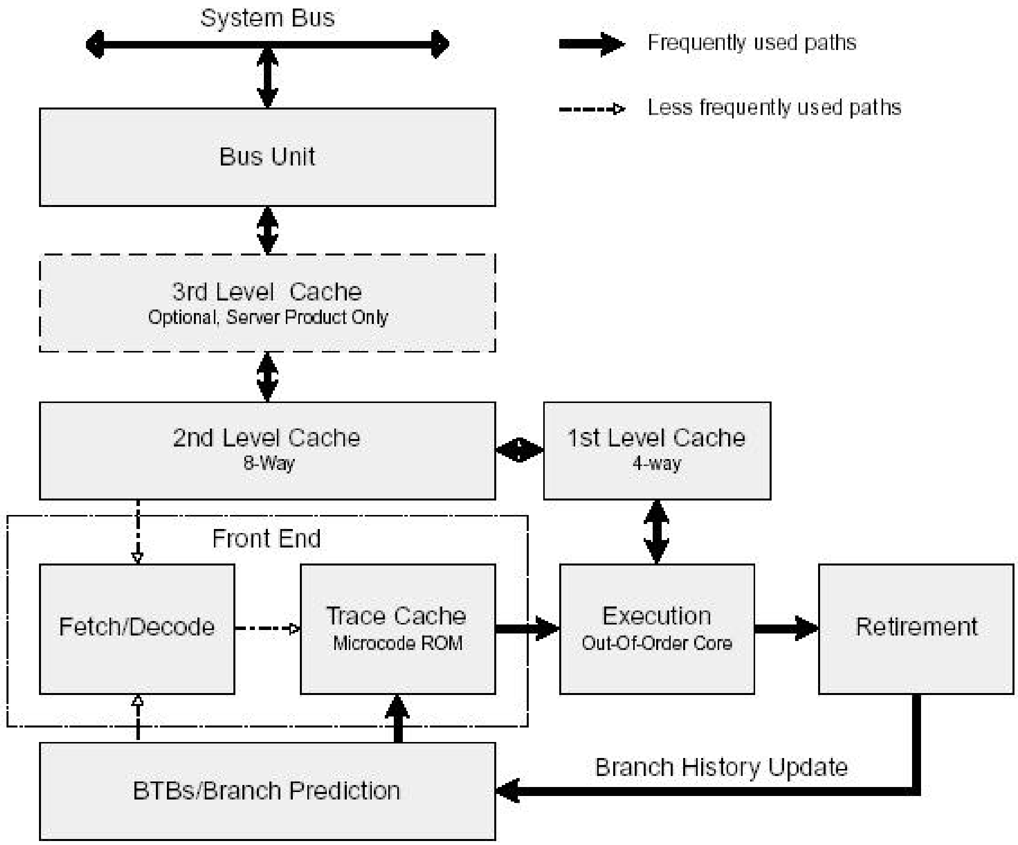


Figure 12. Intel's NetBurst Microarchitecture

For more information on the SSE2 features and benefits and the Pentium 4 architecture in general, see references [7], [8], [9] , [10] and [11].

3D graphics involve accessing significant amounts of data from memory, as well as performing many operations on the data. As a result, you would expect the SSE2 instruction set enhancements and the Intel NetBurst microarchitecture enhancements of the Pentium 4 to significantly improve performance. Figure 13 details the percentage improvements between the 12.41 drivers (which include Pentium 4 optimizations) and the 6.49 drivers (which were released prior to these optimizations):

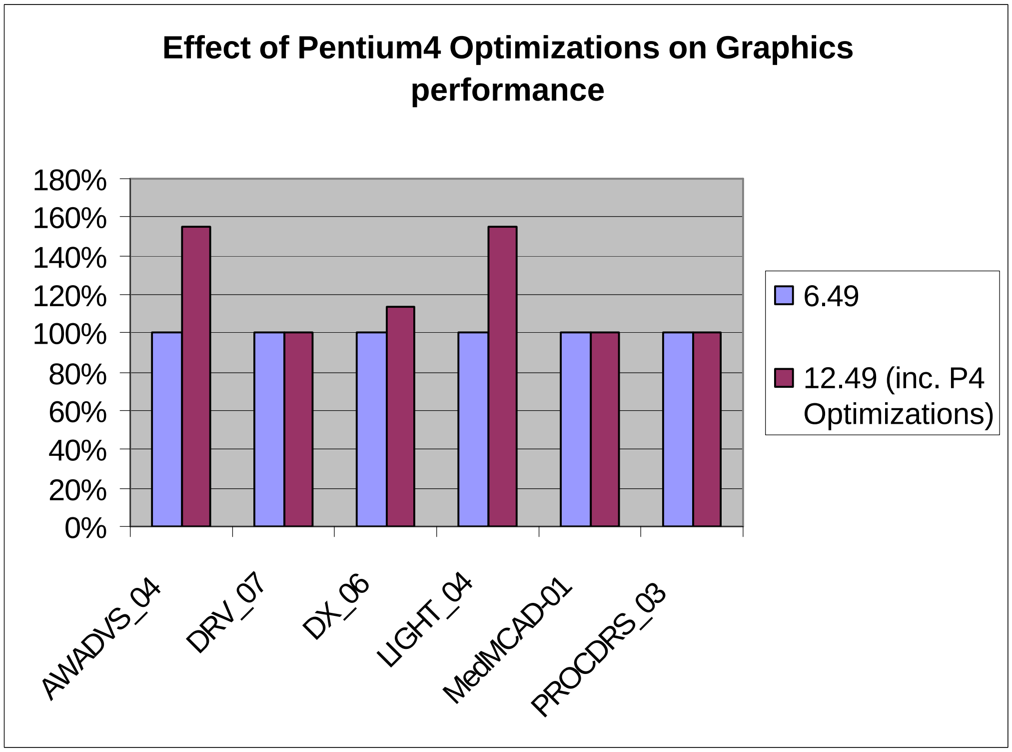


Figure 13. Effect of Pentium 4 Optimizations on Graphics Performance

Although the DRV_07, MedMCAD-01, and ProCDRS_03 tests don't seem to benefit greatly from the Pentium 4 optimizations, remember that these tests use Display Lists and Vertex arrays to send data to the graphics card instead of immediate mode.

NVIDIA graphics drivers are optimized so that the CPU plays a minimal part in the transfer of data for Display Lists and the Vertex Array. Thus, any improvement to CPU architecture has little influence on overall performance. It's not surprising, therefore, that the DRV_07, MedMCAD-01, and ProCDRS_03 tests show little improvement as a result of driver optimizations for the Pentium 4. This also applies to the DX_06 tests, a substantial number of which use display lists. The AWADVS_04 and LIGHT_04 tests, on the other hand, use immediate mode to transfer data to the graphics card and show performance improvements of over 50 percent through tuning to the P4 processor.

Although the Pentium 4 driver optimizations benefit both Quadro and GeForce GPU families, professional applications that use immediate mode graphics are likely to show similarly significant improvements in performance. These translate directly into productivity benefits that are attractive to professional workstation users.

Unified Driver Architecture

One of the most revolutionary and significant benefits offered with NVIDIA's professional workstation and consumer GPU families is the Unified Driver Architecture (UDA). The UDA allows one set of drivers to be used across the entire range of NVIDIA products—including consumer and workstation products. It would be incorrect to assume, however, that because one driver works with all NVIDIA GPU solutions, that all NVIDIA solutions are the same. This is not true.

Figure 14 shows a diagrammatic representation of the UDA and how binary compatibility is made possible:

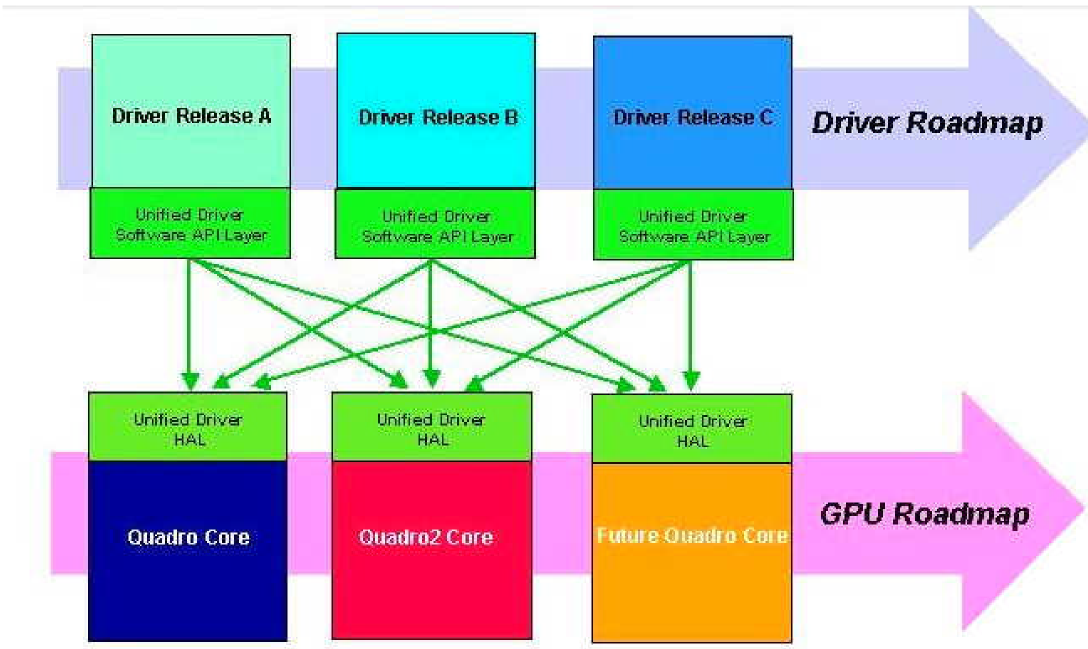


Figure 14. NVIDIA's Revolutionary Patented UDA

The interface between the driver and the hardware (Figure 14) has two components: the Unified Driver Software API Layer and the Hardware Abstraction Layer (HAL). Combined, they provide the following key advantages:

- The driver can assume a uniform interface for all hardware GPUs. This is achieved through a series of class structures that abstract the functions of each GPU. This approach benefits the end user—and prevents the driver from growing significantly in size through unnecessary duplicate code to support different GPU architectures.
- UDA provides extendibility for new features. This is very important for end users; without it, supporting multiple GPUs would be limited to providing the lowest common denominator in features across all GPUs. This wouldn't be much use to professional users.

NVIDIA's proven record of delivering industry-leading features with every new GPU, while maintaining support for previous generations, demonstrates that the features are not restricted to the lowest common denominator of GPUs.

UDA also offers several indirect benefits for end users. Not only is the majority of the driver code common between GPUs, it is also common between operating systems. In fact, around 98 percent of the code base is shared between Linux and Windows. A key benefit this offers the end user is quick access to feature enhancements and bug fixes across the entire product range. It also means that new operating systems, such as 64-bit Linux, can be adopted quickly and easily.

The advantages of UDA are a strategic advantage for end users: a single system-installation image that can be used across an entire range of products. Where production cycles are mission critical—at film and animation studios or engineering firms—this is a major advantage. The reduced system administration overhead and speed with which solutions can be deployed and updated affords unparalleled agility, ensuring that users can use leading-edge technology for competitive advantage without sacrificing quality.

Application Support and Optimization

NVIDIA Quadro GPUs provide several additional features and benefits in the area of professional application optimization and certification.

Application Optimization

In the workstation market place, applications provide the enabling technology that lets users unlock the potential of system capabilities. The integration and coupling between GPU performance and features with applications is the catalyst that makes this happen. To this end, NVIDIA works closely with all workstation application developer, including AliasWavefront, Adobe, Autodesk, Avid, Bentley, Dassault, Discreet, Multigen-Paradigm, Newtek, Nothing Real, Parametric Technology Corp., SDRC, Softimage, SoildEdge, SolidWorks, and Unigraphics.

By working closely with these and other software developers, NVIDIA ensures that applications take full advantage of all features provided by GPUs and that graphics drivers are optimized to the needs of the application. In most cases, these needs are specific to a particular application.

To accommodate this, the NVIDIA graphics control panel for Quadro workstation GPUs allows custom application-specific settings. These settings are accessed from the OpenGL control panel, which is accessed via the Additional Properties button from the Advanced panel in Display Properties. Figure 15 is a screen shot of the custom OpenGL Application settings panel.

This panel and the application-specific features and tuning are not available on the consumer GPU family:

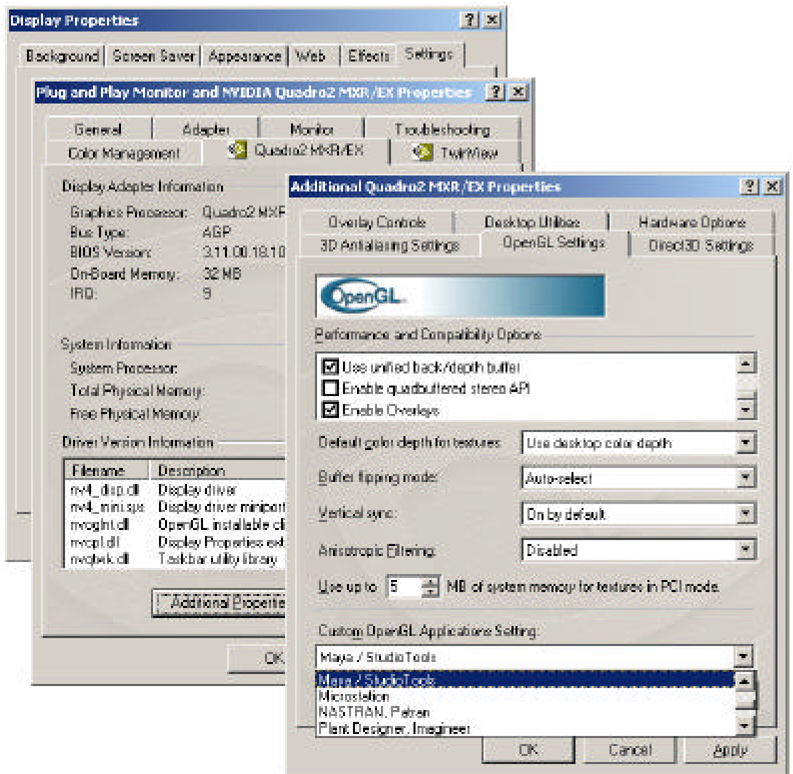


Figure 15. Custom OpenGL Application-Specific Settings

Certification

NVIDIA workstation graphics drivers undergo rigorous in-house quality and regression testing using many workstation applications. Table 5 lists the applications currently used for in-house quality and regression testing:

Table 5 Applications Used for In-House Quality and Regression Testing

Company	Application
Alias|Wavefront/SGI	Maya, Maya Linux, Maya Mac, Design Studio, Studio, AutoStudio
Actify	3D View
Adobe	Illustrator, Photoshop, Premiere, Live Motion (web gfx & animation), Go Live, In Design, Acrobat
Ansys	Ansys, Design Space
Autodesk/Discreet/Kinetix	3D Studio Max, Magma, 3D Studio Max, AutoCAD, 3D studio Viz, Character Studio, Lightscape, Mechanical Desktop, Architectural Desktop, Inventor, Combustion
Avid	Softimage 3D, Softimage 3D for Linux,

Company	Application
	Marquee, Sumatra XSI
Bentley	Microstation J
Caligari	truSpace
Crystal Graphics	3D Sensations, 3D Impact! Pro, PowerPlugs: Transitions
Dassault Systemes	Catia Version 5
Digital Immersion	Merlin VR, Merlin 3D
EAI/Sense 8	Indy 3D
Electricc Image	Universe
ERDAS	Stereo Analyst NT, Imagine 98, & NT
ESRI	ArcView GIS, ArcView 3D Analyst, ArcInfo
Hash	Animation Master 2000
HP / Co-Create	ME10, SolidDesigner
Hummingbird	Exceed, Exceed 3D
MetaCreations	Bryce 4
MSC. Working Knowledge	Working Model 3D
Multigen-Paradigm	CreatorPro, Vega
Nemo	NeMo Creation
Newtek	Lightwave 3D Win & Mac
Nothing Real	Shake
Parametric Technology Corp.	Pro/E, Pro/Designer, CDRS, 3D paint, Division Product View, Division Mockup, Pro/Intralink
Revit Technology Corp.	Revit
Right Hemisphere	Deep Paint 3D, Deep Paint 3D w/ texture weapons
Q E Radiant	Q 3 Radiant
SDRC	Ideas Master Series, Imageware Surfacer
Side Effects Software	Houdini for Windows, Houdini for Linux
Softcad International	SoftCAD 3D, SoftCAD.3D Lite, ArchiTECH.PC
SolidWorks	SolidWorks
Surfware	Surfcam 99, Surfcam 2000
Sven	Surface Suite Pro
Unigraphics Solutions Inc	SolidEdge, Unigraphics
Wolfram	Mathmatica Professional, Mathmatica 4.1.0 Professional

By testing new workstation drivers against numerous applications, NVIDIA is able to catch more bugs and regressions and deliver high-quality new driver releases.

Application Productivity Tools

NVIDIA workstation GPUs provide more than additional hardware features and application driver support. They also provide several application productivity tools that assist with the end user's workflow and productivity.

These productivity tools are available for free download from NVIDIA's Web site, see reference [12], and work only on Quadro workstation products:

- POWERdraft: Provides an optimized plug-in graphics driver for AutoCAD that significantly improves drawing performance—in some cases by over 200 percent. Also provides several functions and features that augment AutoCAD workflow.
- MAXtreme: Provides an optimized plug-in graphics driver for 3D Studio max that significantly improves drawing performance—in some cases by over 80 percent. Also provides several additional functions and features, such as stereo viewing capability.
- QuadroView: A standalone 3D viewer that automatically loads current models from AutoCAD when run concurrently, and from Inventor and vrml file formats. Also provides interactive viewing functions and features, such as stereo viewing.

POWERdraft and MAXtreme significantly improve productivity. Figure 16 shows the performance improvements afforded by the POWERdraft plug-in drivers, compared with the standard AutoCAD OpenGL driver. The test used to demonstrate the improvements was the CADALYST benchmark2001 ACAD200i. For further details on this test, see reference [13].

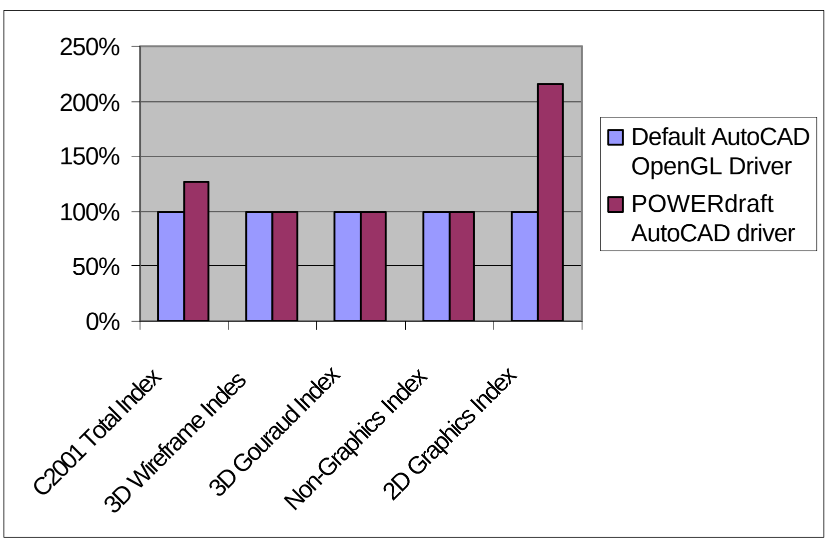


Figure 16.   POWERdraft AutoCAD Plug-In Graphics Driver on CADAYST Benchmark 2001

As Figure 16 shows, the performance improvements provided by the POWERdraft plug-in driver are significant—they improve performance by over 200 percent. One wouldn't expect improvement on the nongraphics and 3D operations because the POWERdraft plug-in driver targets 2D performance. Clearly, to aid 3D productivity, QuadroView provides significant features and benefits.

The performance improvements in 3D Studio max afforded by the MAXtreme plug-in driver (Figures 17) provide a comparison with the default 3D Studio max OpenGL driver on Discreet's 3D Studio max graphics card certification test suite for. For further details, see reference [14]. The results show significant productivity benefits on many 3D Studio max tests and, in some cases, the improvement is nearly 2x.

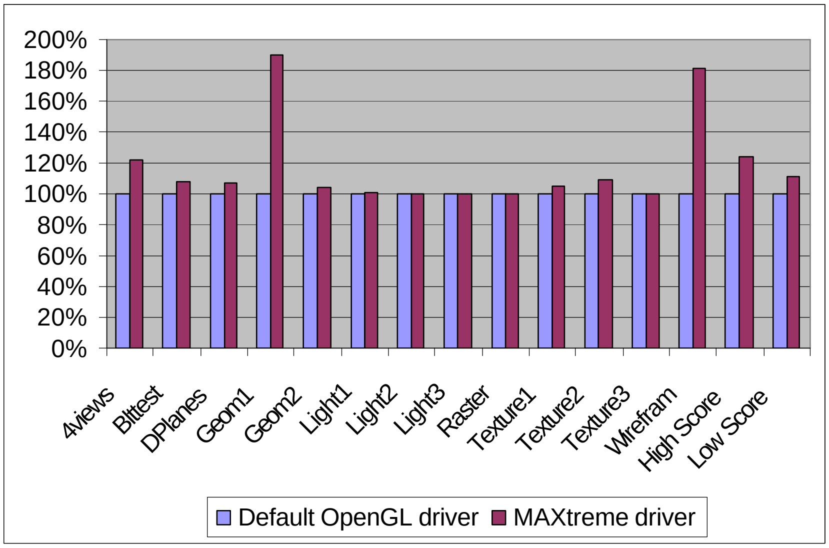


Figure 17. MAXtreme 3D Studio Plug-In Graphics Driver on the Discreet Certification Test

Along with the direct performance advantages, the application productivity tools provide a series of additional features and functions. Figures 18 and 19 are screen shots of POWERdraft and QuadroView.

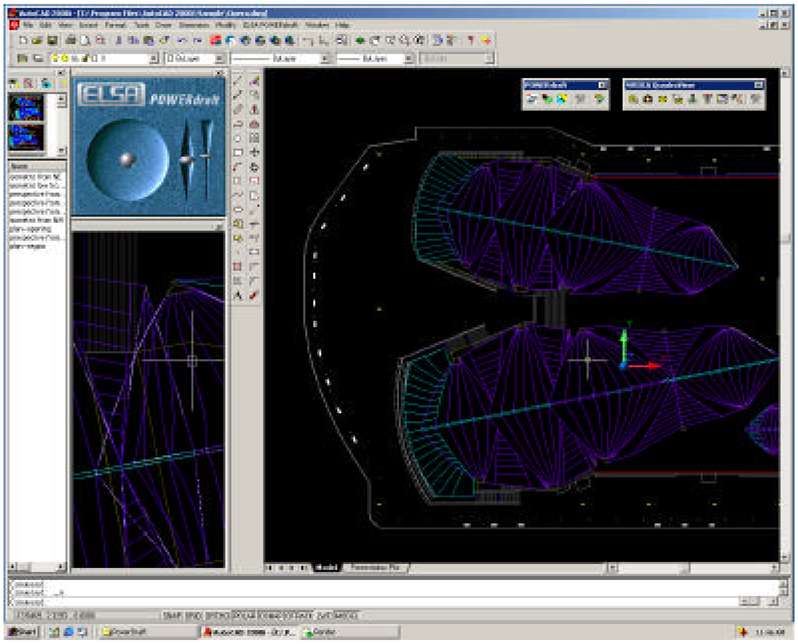


Figure 18. POWERdraft Used with AutoCAD

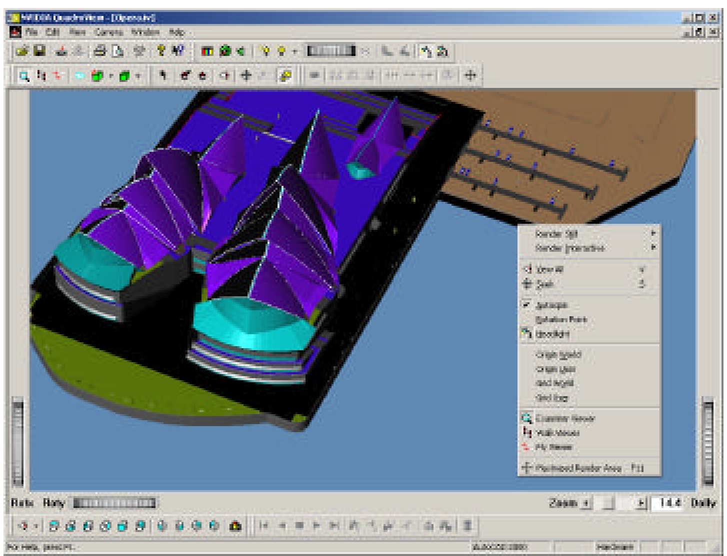


Figure 19. Model Display Automatically Loaded into QuadroView When Read into AutoCAD

As Figure 18 shows, POWERdraft provides several extra windows within the default AutoCAD window. These windows provide additional features that let users manipulate viewpoint, orientation, and zoom with several easy-to-use visual controls. They also provide a window that displays a close-up view around the current cursor location to provide greater detail in the area of interest. Also, a series of viewpoints can be defined and quickly selected through icons and text descriptions.

QuadroView lets users view models in stereo. To enable this, stereo must first be turned on in the OpenGL control panel, which is accessed via Additional Properties from the Windows Display Properties panel. When this is done, stereo can be enabled within QuadroView. The recommended mode is Raw OpenGL, which corresponds to quad-buffered stereo. When this is enabled for stereo and suitable glasses are connected (such as the Elsa Revelator or the Crystal Eyes Wired), the model can be viewed in three dimensions. There are also controls to affect eye spacing and parallax to ensure optimal viewing comfort. Figure 20 is a screen shot of QuadroView with stereo configured:

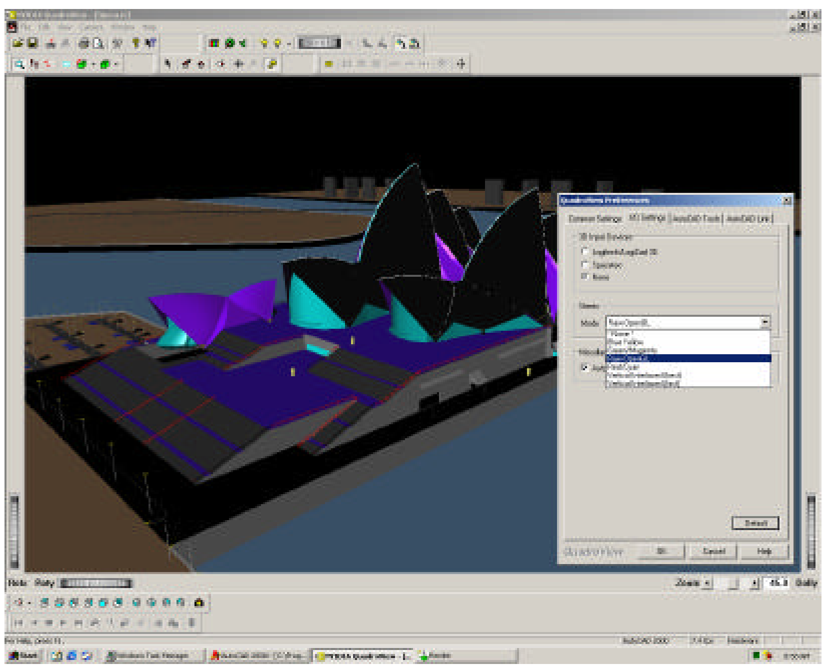
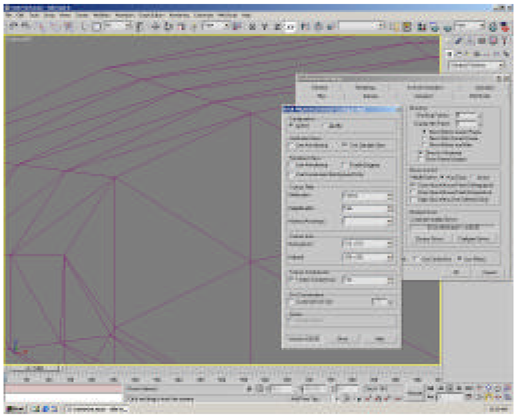


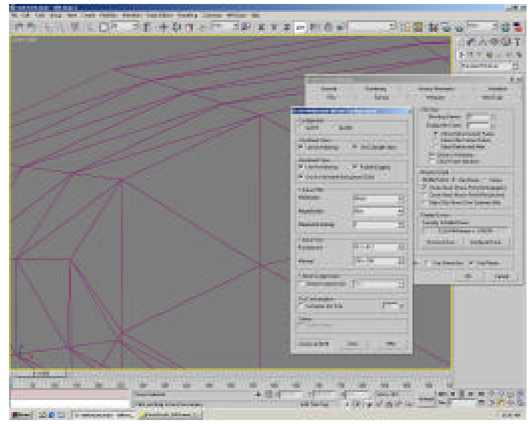
Figure 20. QuadroView Stereo Viewing Configuration

Like POWERdraft, MAXtreme offers 3D Studio max users many additional features. A key one is the ability to adjust graphics quality to optimize for either quality or performance.

Figure 21 shows side-by-side comparison of cases optimized for speed and quality:



3D Studio max optimized for speed



3D Studio max optimized for quality

Figure 21. Speed and Quality Optimizations of MAXtreme Plug-In Driver for 3D Studio Max

Along with the speed and quality optimizations, MAXtreme—like QuadroView—lets users view 3D Studio max scenes in stereo. Again, stereo must first be turned on via the OpenGL control panel and then enabled within the MAXtreme driver.

Figure 22 is a screen shot of 3D Studio max with stereo enabled in the MAXtreme driver. Like QuadroView, the eye separation and parallax can be adjusted to ensure optimal viewing.

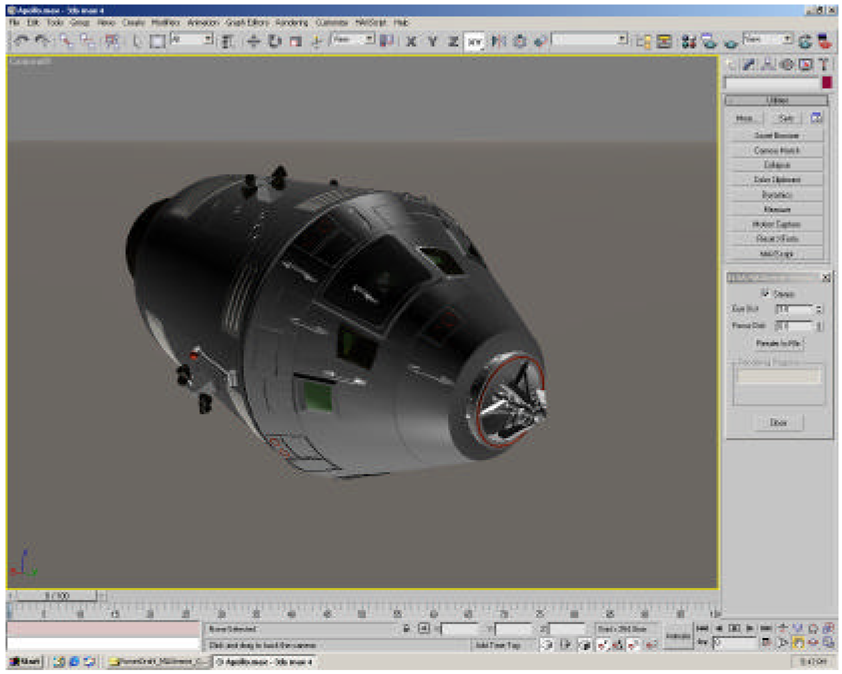


Figure 22. 3D Stereo Max Detailing Stereo Viewing with MAXtreme

For further information on POWERdraft, MAXtreme, and QuadroView, refer to the release notes installed with the tools, which provide a complete description of features and instructions on use.

Future Direction

As the workstation market evolves and new technologies introduce, the tools for users will become more sophisticated. This will create significant, but beneficial, discontinuities in workflows. A discontinuity was shown at SIGGRAPH 2001 in Los Angeles, where NVIDIA and Square demonstrated real-time playback of *Final Fantasy: The Spirits Within* on a single system powered by a Quadro DCC GPU. Each frame took over 90 minutes to render, so visualizing the same content in real time with suitable quality represents a major breakthrough.

Figures 23 and 24 show two frames from the *Final Fantasy: The Spirits Within* playback and demonstrate the visual quality achieved. In both cases, a wire-frame image is overlaid over each frame to show the construction of the content.

Figure 23. Frame #1 from *Final Fantasy: The Spirits Within*

Figure 24. Frame #2 from *Final Fantasy: The Spirits Within*

The sophisticated anisotropic lighting effects, particularly noticeable in Aki's hair, were created using the vertex and pixel shaders introduced with the Quadro DCC GPU. Likewise, the vertex shaders [15] and pixel shaders [16] were used to bumpmap Aki's and Sid's uniforms and skin.

In Figure 23, where Aki's hand casts a shadow on her uniform, the shadow buffer [17] support also allows real-time shadows, which increases realism. These effects produce realistic and lifelike results that, coupled with performance, let this level of realism be viewed interactively, move computer graphics to the next level.

The real-time demonstration of *Final Fantasy: The Spirits Within* shows how bump-mapping realistically represents materials on objects, like carpeting in a building or car seats. Likewise, other effects that capitalize on vertex and pixel shaders—such as anisotropic lighting—allow new levels of quality and realism that were previously unattainable.

As NVIDIA GPUs continue to improve, the benefits to the end-user workflow will progress in parallel.

Conclusion

NVIDIA leads the computer graphics industry, both in product delivery and in evolution of product features and performance. The workstation market space, however, has specific requirements that are driven by the needs of professional applications and mission criticality of user environments. This technical brief described the features that the Quadro workstation GPU family offers (over and above the consumer GeForce GPU family), which meet these workstation requirements.

While the benefits of hardware-oriented features—antialiased points and lines, logic operations, clip regions, hardware-accelerated clip planes, two-sided lighting, and overlays—are somewhat hidden from a end-user's workflow, this paper has shown how support for these features can translate directly into productivity benefits. The productivity benefits of software features such as workstation application support and the application productivity tools of PowerDraft, MAXtreme, and QuadroView, are obvious. However, this doesn't mean that their benefit is less meaningful.

Likewise, because NVIDIA's UDA removes the common administration headaches in production environments and provides unparalleled reliability and dependability in mission-critical situations, it would be easy to underestimate the significance of these benefits.

The effect of these benefits is compelling for professional workflows in production environments. However, when coupled with the price-performance of the Quadro GPU families, the advantages move to a new level. In examples like the real-time playback of the *Final Fantasy: The Spirits Within*, NVIDIA workstation GPUs enable step changes in workflow.

Because the competitive advantage and market leadership benefits from these step changes are often large, they can be difficult to quantify. Perhaps the best way they can be measured is by the market opportunity they create. As NVIDIA continues to deliver professional workstation products like the Quadro2 GPU family, it lays the foundation for these business opportunities, allowing those who capitalize on them to reap substantial financial rewards.

Appendix: References

OpenGL Programming Guide, 3rd Edition, Mason Woo, Jackie Neider, Tom Davis, Dave Shreiner. Addison-Wesley, ISBN 0-201-60458-2.

OpenGL Programming for the X Window System, Mark J. Kilgard. Addison-Wesley, ISBN 0-201-48359-9.

NVIDIA's Stereoscopic 3D Development Guide, http://partners.nvidia.com/developer.nsf

"StereoGraphics developers hand book", StereoGraphics Corporation, http://www.stereographics.com/support/developers/handbook.pdf

"3D Stereo Rendering Using OpenGL (and GLUT)", Paul Bourke, November 1999 http://astronomy.swin.edu.au/pbourke/opengl/stereogl/

"Calculating Stereo Pairs", Paul Bourke, July 1999. http://astronomy.swin.edu.au/pbourke/stereographics/stereorender/

Intel Pentium 4 Processor Optimization Reference manual,

http://developer.intel.com/design/Pentium4/papers/

IA-32 Intel Architecture Software Developer's Manual Volume 1: Basic Architecture, http://developer.intel.com/design/Pentium4/manuals/

IA-32 Intel Architecture Software Developer's Manual Volume 2: Instruction Set Reference Manual, http://developer.intel.com/design/Pentium4/manuals/

IA-32 Intel Architecture Software Developer's Manual Volume 3: System Programming Guide, http://developer.intel.com/design/Pentium4/manuals/

To download POWERdraft, MAXtreme and QuadroView go to http://www.nvidia.com/, look under download drivers, and then select either Win2K or Windows NT operating system.

Information on the CADALYST C2001 benchmark can be found under: http://www.cadalyst.com/reviews/cadbench/

Information on the 3D Studio max test suite can be found under:

http://www.discreet.com/support/max/index.html, and then look under the tested graphics cards. Descriptions of the tests themselves are located specifically at: http://www.discreet.com/support/max/videocards/r4_description.html

For information on vertex programs and how they can be used see: http://developer.nvidia.com/view.asp?IO=vertex_programs http://developer.nvidia.com/view.asp?IO=OpenGL_Vertex_Cheat

For more information on pixel shaders and how they can be used see:
http://developer.nvidia.com/view.asp?IO=dynamic_bump_reflection
http://developer.nvidia.com/view.asp?IO=bumpmappingwithregistercombiners

http://developer.nvidia.com/view.asp?IO=texture_shaders

http://developer.nvidia.com/view.asp?IO=bumpmappingwithregistercombiners

For more information on shadow volumes see:
http://developer.nvidia.com/view.asp?IO=cedec_shadowmap

http://developer.nvidia.com/view.asp?IO=shadow_mapping

NVIDIA Corporation
2701 San Tomas Expressway
Santa Clara, CA 95050
www.nvidia.com