



Technical Brief

NVIDIA Quadro vs. GeForce GPUs
Features and Benefits

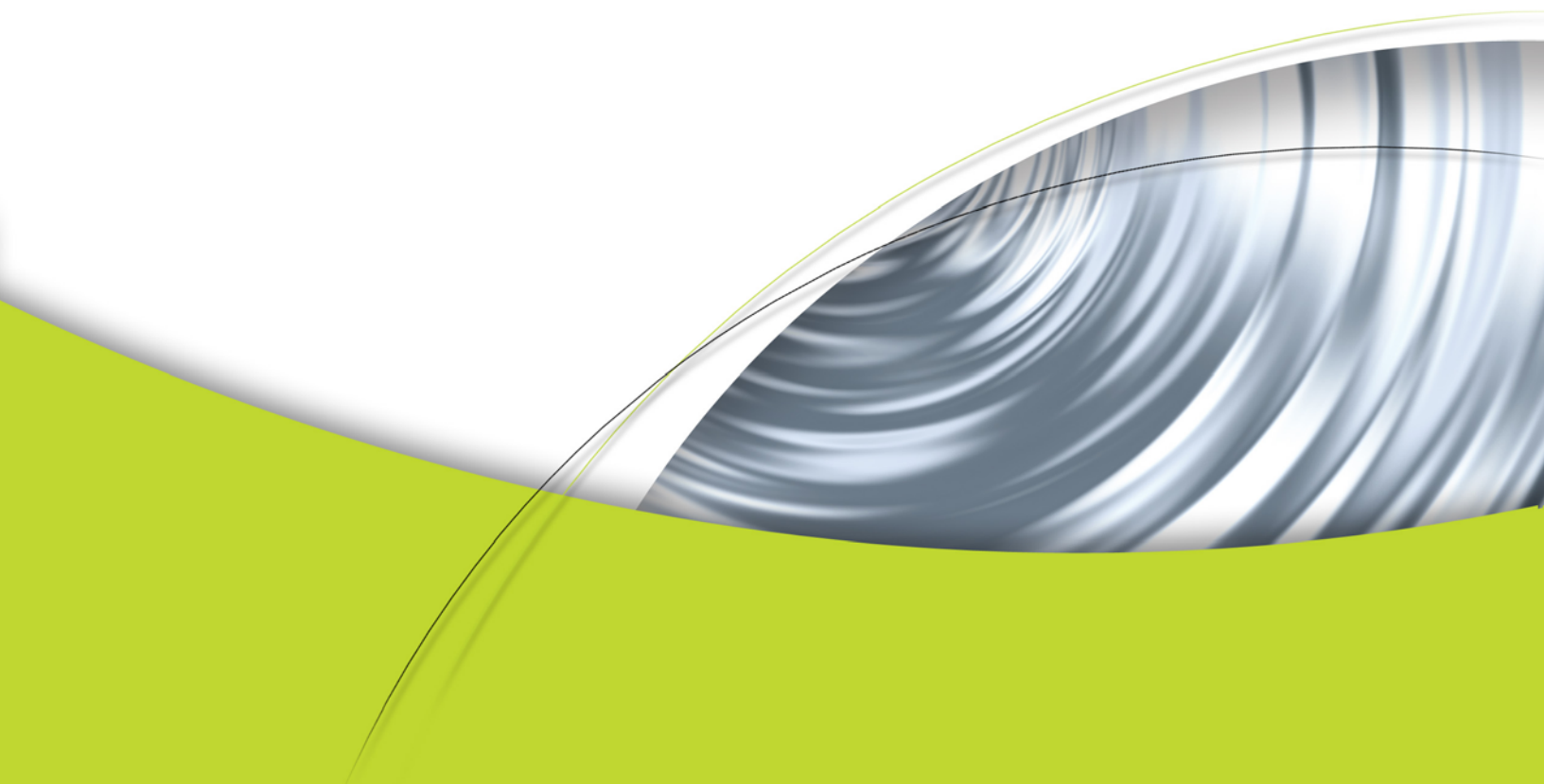


Table of Contents

Purpose.....	1
Why Choose a Workstation?.....	2
Overview of Workstation Products	3
Workstation Features	4
Antialiased Points and Lines.....	4
Logic Operations	5
Example of CAD Application	6
Example of DCC Application	6
Clip Regions.....	7
Hardware-Accelerated Clip Planes	8
NVIDIA Quadro Memory Management Optimization	9
Normal Memory Demands	9
High Memory Demands	9
Two-Sided Lighting.....	10
Lighting Components	11
Visual Issues	12
Solutions for Visual Issues	12
Overlay Plane Support	14
Cursor Issue.....	14
Pop-Up Menu Issue.....	14
Solution	14
Quad-Buffered Stereo	16
Pentium 4 Optimizations	16
NetBurst	17
Unified Driver Architecture	18
Application Support and Optimization	19
Application Optimization	19
Certification.....	20
Application Productivity Tools.....	21

Conclusion..... 26
References Appendix..... 27

List of Figures

Figure 1.	Points/Lines Enabled in Major CAD Applications	4
Figure 2.	The OpenGL Pipeline	5
Figure 3.	CATIA V5 Screenshot Showing Highlighted Feature.....	6
Figure 4.	Window Overlapping in Pro/ENGINEER and Effect on the Number of Clip Regions....	7
Figure 5.	Comparison of Peak Lit, Smooth-Shaded, Depth-Tested Triangle Performance.....	8
Figure 6.	Components That Represent Lighting Effects in the Real World.....	11
Figure 7.	The Effects of Two-Sided Lighting	13
Figure 8.	Brush Outline Feature in Maya That Uses Overlay Planes.....	15
Figure 9.	Intel's NetBurst Microarchitecture	17
Figure 10.	NVIDIA's Revolutionary Patented UDA	18
Figure 11.	POWERdraft Used with AutoCAD	22
Figure 12.	Model Display Automatically Loaded When Read into AutoCAD.....	23
Figure 13.	NVIDIA QuadroView Stereo Viewing Configuration.....	24
Figure 14.	Speed and Quality Optimizations of MAXtreme Plug-In Driver for 3ds max	24
Figure 15.	3ds max Detailing Stereo Viewing with MAXtreme.....	25

List of Tables

Table 1.	Features of Workstation GPU Family	3
Table 2.	Applications Used for In-House Quality and Regression Testing.....	20



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 and the workstation-class NVIDIA Quadro® family?”

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 multiple application resource management, and the acceleration of OpenGL features such as antialiased points and lines. Finally, it describes application features and enhancements provided with workstation drivers—POWERdraft, MAXtreme, and QuadroView—and shows how they benefit commonly used applications.

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 workstation users have specific goals in mind—goals that are ultimately critical to success. The goal may be designing a revolutionary car or spacecraft, 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 NVIDIA Quadro and the consumer-branded GeForce GPUs, many consumers were unclear about the benefits 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 tens of thousands of dollars, whereas a PC could cost less than \$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 document describes in detail the features and benefits offered by the NVIDIA Quadro brand 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 tailored to the specific requirements of those applications.

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

Product	Positioning	Memory	Graphics Precision	proe-02	ugs-03	3dsmax-02
NVIDIA Quadro FX 3000/3000G (with Genlock)	Extreme workstation graphics	256 MB/ 256 bit	128 bit floating point	41.6	44.4	26.1
NVIDIA Quadro FX 2000	High-end workstation graphics	128 MB/ 128 bit	128 bit floating point	40.8	42.7	26.0
NVIDIA Quadro FX 1000	Performance workstation graphics	128 MB/ 128 bit	128 bit floating point	33.9	33.2	23.3
NVIDIA Quadro4 980 XGL	Midrange workstation graphics	128 MB/ 128 bit	32 bit	23.3	23.6	20.0
NVIDIA Quadro4 750 XGL	Midrange workstation graphics	128 MB/ 128 bit	32 bit	19.9	18.0	17.9
NVIDIA Quadro FX 500	Entry workstation graphics	128 MB/ 128 bit	128 bit floating point	16.8	12.2	12.1
NVIDIA Quadro4 580 XGL	Entry low-profile workstation graphics	64 MB/ 128 bit	32 bit	15.2	12.4	11.4
NVIDIA Quadro4 380 XGL	Entry workstation graphics	64 MB/ 128 bit	32 bit	14.0	11.5	10.6
NVIDIA Quadro FX Go700	Mobile high-end workstation graphics	128 MB/ 128 bit	128 bit floating point	23.5	17.2	14.5

Table 1. Features of Workstation GPU Family

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 (sometimes called “wireframe”). 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.

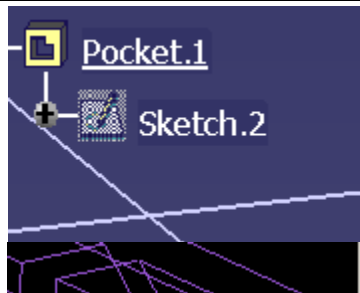
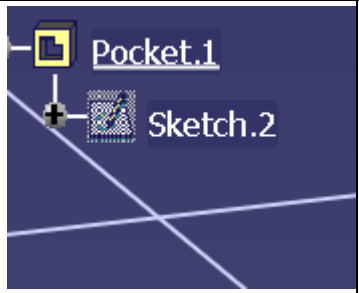
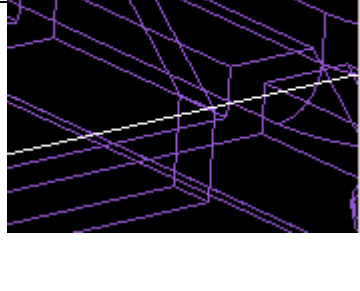
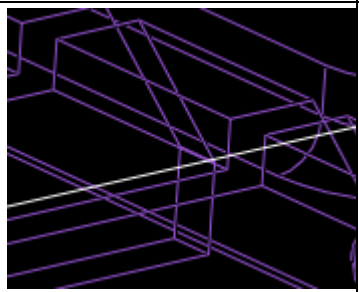
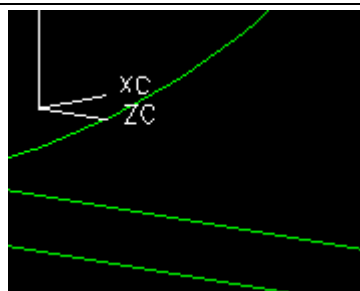
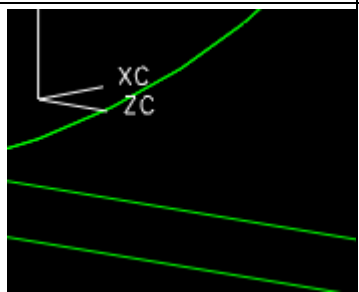
	Non-Antialiased	Line Antialiased
CATIA V5 Turn on Line Antialiasing: Tools/Options/Display/Visualization		
Pro/ENGINEER 2000i2 Turn on Line Antialiasing: View/Model Display		
EDS Unigraphics Version 17 Turn on Line Antialiasing: Preferences/Visualization/Visual		

Figure 1. Points/Lines Enabled in Major CAD Applications

In Figure 1, a series of close-up screenshots show line antialiasing turned on and off for major CAD applications.

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

The performance advantage of the NVIDIA 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 and not sacrifice performance and interactivity. For a CAD designer working in wireframe—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 NVIDIA Quadro workstation GPUs.

Logic Operations

Another hardware feature difference between NVIDIA's workstation and consumer GPUs is support for OpenGL Logic Operations (Figure 2). 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*.

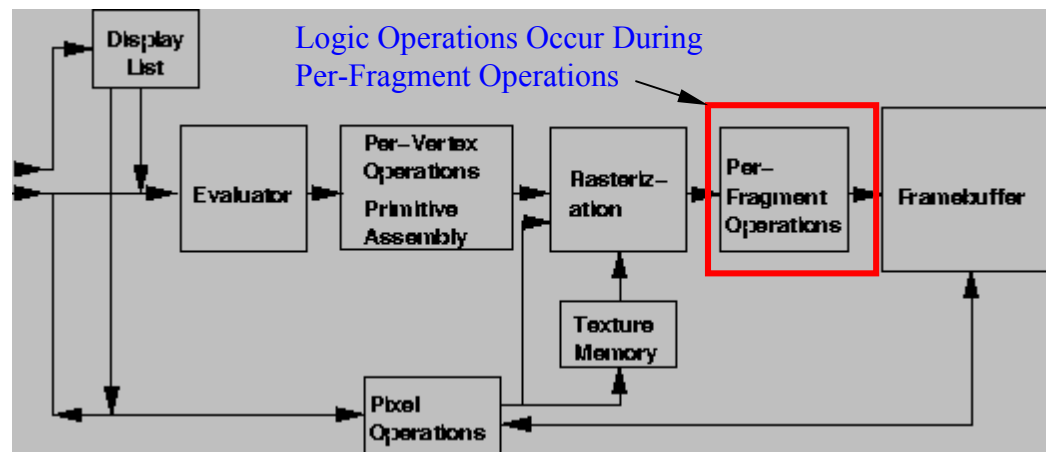


Figure 2. The OpenGL Pipeline

Logic operations are often used by workstation applications in mechanical computer-aided design (MCAD) and digital content creation (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

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

Figure 3 shows an example using Dassault's CATIA V5, where the highlighted feature is orange. The feature highlighted by the cursor is drawn in an orange outline so it can be dynamically identified by the user before a selection or pick is made.

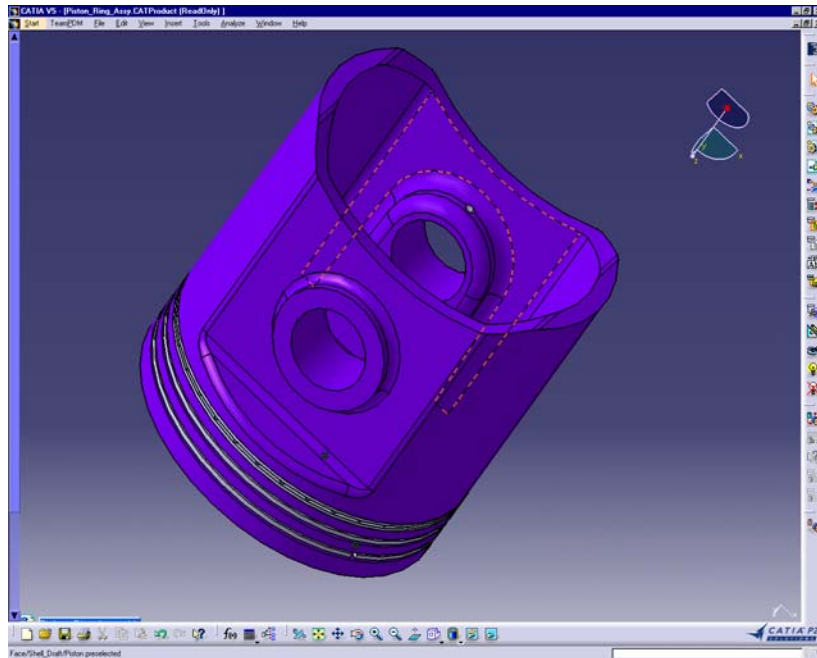


Figure 3. CATIA V5 Screenshot Showing Highlighted Feature

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' Maya application. Refer to [Overlay Plane Support](#), which shows a screenshot 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 overlay planes.

If the XOR logic operation is enabled, the performance drop of the NVIDIA Quadro is minimal when compared to that of GeForce. 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. Unlike consumer applications such as games, these applications often occupy the full screen, so the result is many overlapping windows. Depending on how they are handled by the graphics hardware, overlapping windows may noticeably affect visual quality and graphics performance.

NVIDIA's Quadro 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 4, a screenshot from PTC's Pro/ENGINEER, 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.

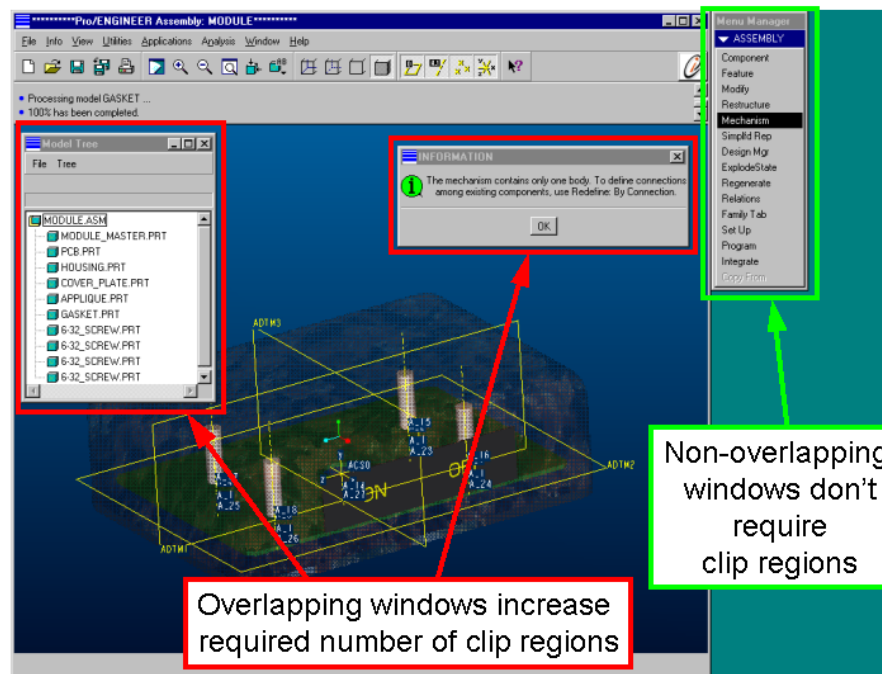


Figure 4. Window Overlapping in Pro/ENGINEER and Effect on the Number of Clip Regions

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 affects 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 NVIDIA Quadro family support up to eight clip regions.

Figure 5 shows how clip regions affect overall graphics performance. 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.

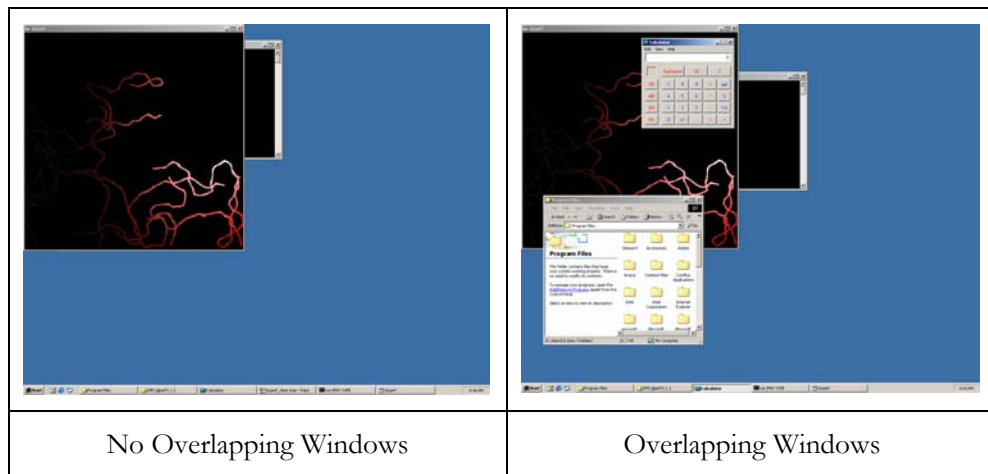


Figure 5. Comparison of Peak Lit, Smooth-Shaded, Depth-Tested Triangle Performance

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, professional CAD applications, including Pro/ENGINEER, often let users define clip planes.

The NVIDIA Quadro family of GPUs supports clip-plane acceleration in hardware—a significant performance improvement when it is used in professional applications.

NVIDIA Quadro Memory Management Optimization

Another feature offered by the NVIDIA Quadro family of GPUs is 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. 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.

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 around 36MB. Dedicated texture memory of 32MB 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 the occasional situation, but it accommodates normal demands for large amounts of resources.

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 NVIDIA Quadro memory management optimization is important for professional applications because it efficiently accommodates large demands, but does not waste resources or memory when they are not needed.

NVIDIA 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 NVIDIA Quadro memory management optimization, we designed three scenarios that place varying demands on required graphics memory:

- ❑ The first scenario ran the GLperf application alone to measure the peak textured triangle rate corresponding to 1-pixel, lit, smooth-shaded, depth-tested, textured (64×64 RGB trilinear modulated).
- ❑ The second scenario ran GLperf concurrently with the NVIDIA tree demo maximized to full-screen resolution.
- ❑ The third scenario added an additional 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 1280×1024 and 1600×1200 using 32-bit color, and the effects on the peak textured triangle rate were compared. By limiting the triangle size to 1 pixel and fixing the GLperf window to 600×600 , we avoided any influences arising from fill-rate limitations.

At a screen resolution of 1280×1024 , the memory requirement of each scenario had minimal impact on performance. With the screen resolution set to 1600×1200 , however, the increased requirements affected performance, and the peak performance of the GeForce began to degrade. The NVIDIA Quadro memory management optimization allowed the NVIDIA Quadro to remain unaffected.

For professional DCC or CAD applications, which consume significant amounts of texture memory or open many separate 3D graphics windows, the NVIDIA 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.

Lighting Components

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 (Figure 6) to model how objects appear in the real world.





<p>Ambient</p>	<p>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.</p>	
<p>Diffuse</p>	<p>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).</p>	
<p>Specular</p>	<p>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.</p>	
<p>All three components are combined to generate a realistic image.</p>		

Figure 6. Components That Represent Lighting Effects in the Real World

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.

Solutions for Visual Issues

The way to overcome this limitation is to use two-sided lighting. When two-sided lighting is enabled, the lighting calculation uses the magnitude of the dot product—instead of using the dot product of the normal and light vector—to calculate the diffuse and specular components. 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 7.

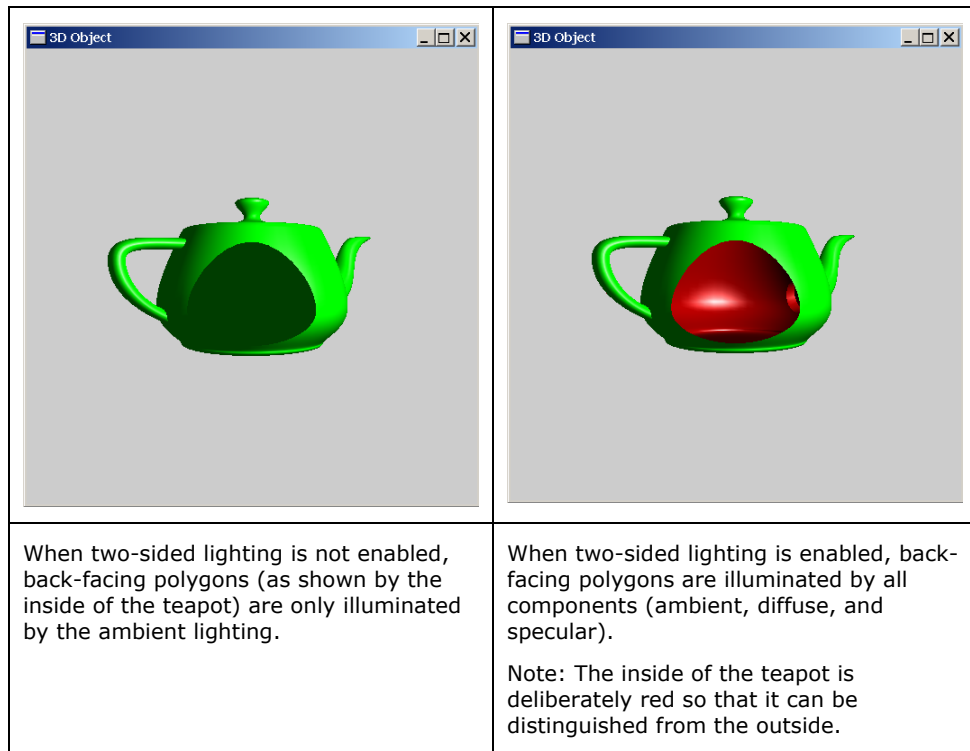


Figure 7. The Effects of Two-Sided Lighting

The teapot on the left in Figure 7 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 (depending on GPU architecture) is that more calculations may be performed at each vertex and so there may be degradation in overall graphics performance.

Overlay Plane Support

The user interface (UI) of many professional applications often requires that elements 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 movements 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 32×32 pixels. A larger image invokes a software path, which noticeably affects performance and interactivity.

Pop-Up Menu Issue

Another example of a UI 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 UI 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 UI 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 (refer to [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 overlaid window to show through. Creating pop-up menus in the overlay planes, therefore, prevents damage to the main graphics window and improves performance. Likewise, clearing an overlaid window to the transparency bit and then drawing graphics within it allows UI items to be drawn over the main graphics window. Clearing and redrawing only the overlaid window

is significantly faster than redrawing the main graphics window. This is how animated UI components can be drawn over 3D models or scenes.

A good example of this UI component is the brush outline in Alias' Maya application. In Figure 8, which shows a screenshot that illustrates the brush outline feature, the red lines of the brush are drawn in overlay planes.

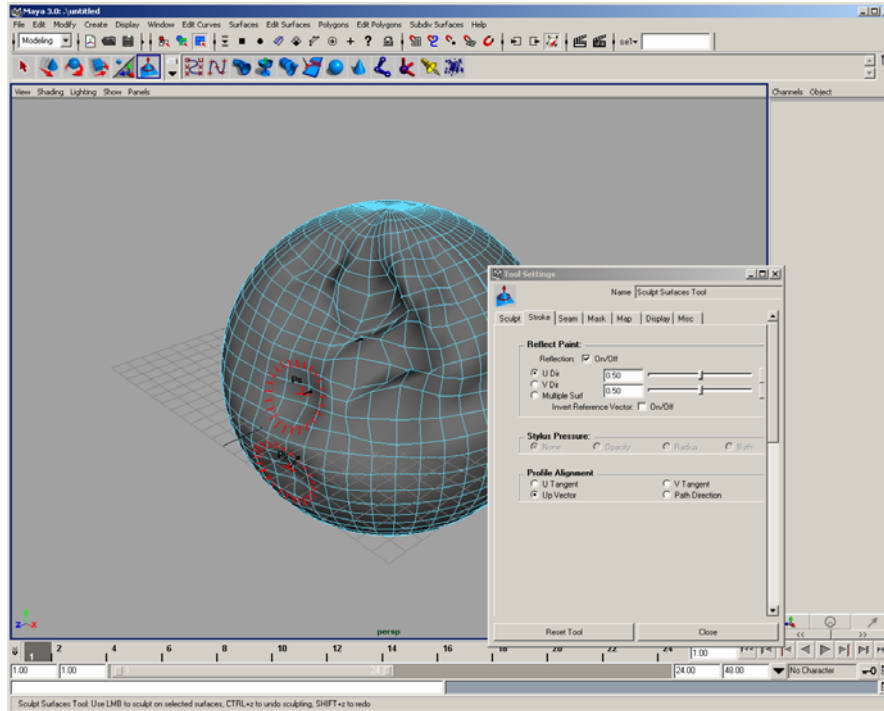


Figure 8. Brush Outline Feature in Maya That Uses Overlay Planes

Support for overlay planes is exclusive to the NVIDIA Quadro 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 planes are definitely productivity aids in many aspects of user workflow.

Quad-Buffered Stereo

The NVIDIA 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 the References Appendix. References 3 and 4 provide useful background information, and 5 and 6 provide code examples. Figure 16 (in [Application Productivity Tools](#)) shows a screenshot from Autodesk's 3D Studio MAX that uses NVIDIA's MAXtreme plug-in driver to enable stereo support in the main viewing window.

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

Pentium 4 Optimizations

With its 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.

These were the key design considerations for NetBurst:

- ❑ 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 9).

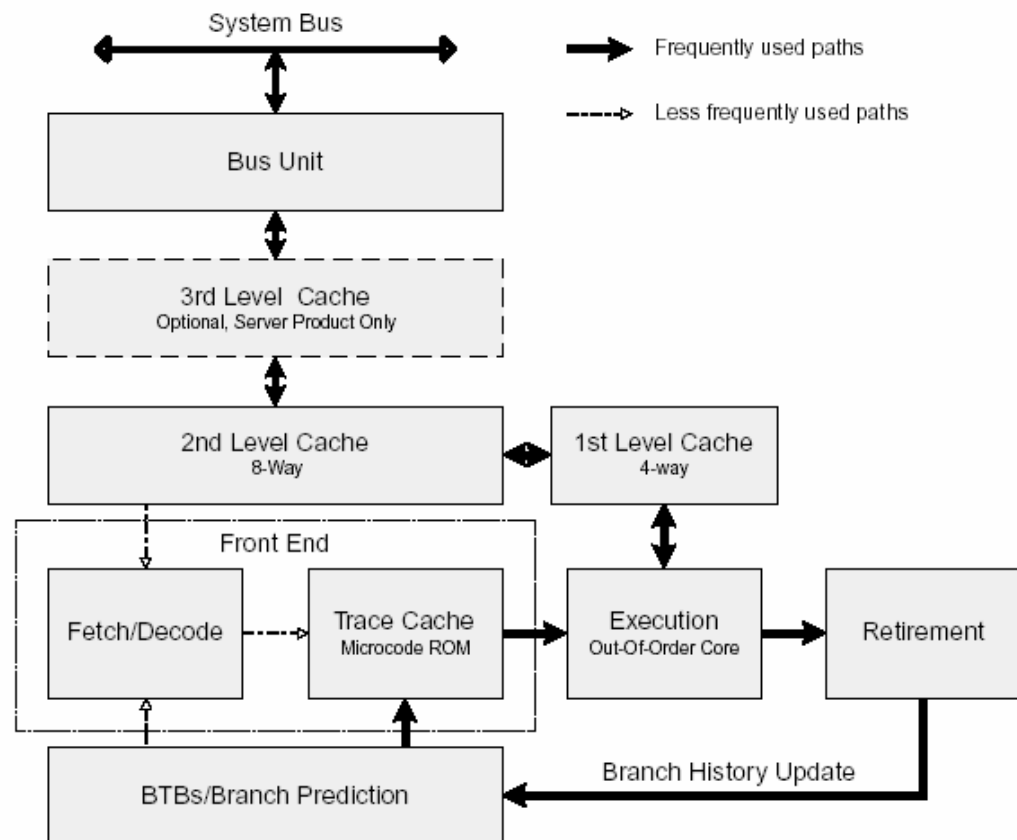


Figure 9. Intel's NetBurst Microarchitecture

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

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.

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. Although the Pentium 4 optimizations benefit both NVIDIA 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 NVIDIA Unified Driver Architecture (UDA). The UDA lets one set of drivers 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 10 shows a diagram of the UDA and how binary compatibility is made possible.

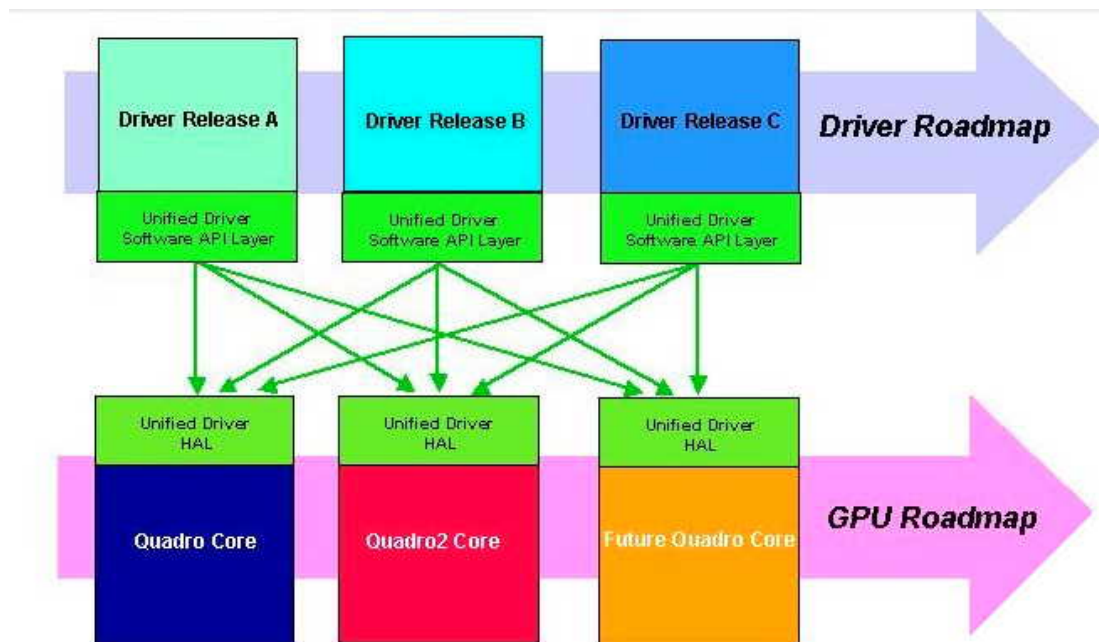


Figure 10. NVIDIA's Revolutionary Patented UDA

The interface between the driver and the hardware (Figure 10) 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 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 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. 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. This offers quick access to feature enhancements and bug fixes across the entire product range. It also means that new operating systems, such as the 64-bit Linux non-Windows operating system, can be adopted quickly and easily.

The advantages of UDA are a strategic advantage for 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 leading-edge technology for competitive advantage without sacrificing quality.

Application Support and Optimization

NVIDIA Quadro GPUs provide several additional features and benefits for 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 developers, including Alias, Adobe, Autodesk, Avid, Bentley, Dassault, Discreet, Multigen-Paradigm, Newtek, Nothing Real, Parametric Technology Corp. (PTC), SDRC, Softimage, SolidEdge, 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 NVIDIA Quadro workstation GPUs allows custom application-specific settings. These settings are accessed from the OpenGL Control Panel (Properties/Settings tab/Advanced/specific product name tab/OpenGL Settings/Custom OpenGL Applications Settings). This panel and the application-specific features and tuning are not available on the consumer GPU family:

Certification

NVIDIA workstation graphics drivers undergo rigorous in-house quality and regression testing using many workstation applications. Table 2 lists the applications currently used for in-house quality and regression testing.

Table 2. Applications Used for In-House Quality and Regression Testing

Company	Application
Adobe	Premiere
Alias	AutoStudio, DesignStudio, Maya, Studio Tools
ANSYS	ANSYS
Apple	Shake
Autodesk	Architectural Desktop, AutoCAD, Inventor, Lightscape, Mechanical Desktop, VIZ
AVEVA	PDMS
Avid	Avid
Bentley	MicroStation
Caligari	truSpace
CoCreate	OneSpace
Dassault	CATIA
Digital Immersion	Merlin 3D
Discreet	3ds max, character studio, combustion
EDS	I-deas NX Series, SolidEdge, Unigraphics, SDRC
Electric Image	Amorphium, EI
ERDAS	StereoAnalyst, VirtualGIS
ESRI	ArcGIS
Fluent	FLUENT
Hash	Animation Master
ICEMSurf	ICEMSurf
IronCAD	IronCAD
Kaydara	MOTIONBUILDER
Maxon	CINEMA 4D
MSC.Software	MSC.Nastran, MSC.Patran

Company	Application
MSC.Working Knowledge	Working Model 3D
Multigen-Paradigm	Creator, Vega
Newtek	LightWave 3D
Opticore	Opus Realizer. Opus Studio
Parametric Technology Corp. (PTC)	3Dpaint, Pro/ENGINEER, CDRS
Pinnacle	Pinnacle
Rhino	Rhino
Right Hemisphere	Deep Paint 3D
Sensable	Sensable
Side Effects	Houdini
Softimage	Softimage XSI, Softimage 3D
SolidWorks	SolidWorks
Surfware	SURFCAM
Think3	Think3

By testing new workstation drivers against numerous applications, NVIDIA identifies and fixes 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 user’s workflow and productivity.

These following productivity tools are available for free download from NVIDIA’s Web site, (see reference 12 in the References Appendix), and work only on NVIDIA Quadro workstation products.

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

POWERdraft and MAXtreme significantly improve productivity.

The performance improvements provided by the POWERdraft plug-in driver are significant—they can 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, NVIDIA QuadroView provides significant features and benefits.

Similarly, the performance improvements in 3ds max afforded by NVIDIA's MAXtreme plug-in driver are dramatic compared to Discreet's default 3ds max OpenGL driver. For further details, see reference 14 in the References Appendix. The results show significant productivity benefits on many 3ds max tests and, in some cases, the improvement is nearly double.

Along with the direct performance advantages, the application productivity tools provide a series of additional features and functions. Figures 11 and 12 are screenshots of POWERdraft and NVIDIA QuadroView.

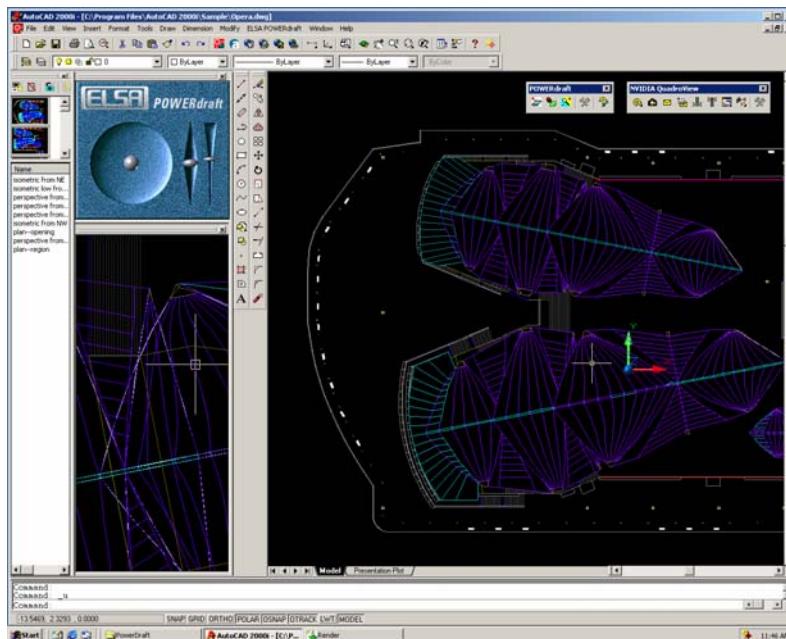


Figure 11. POWERdraft Used with AutoCAD

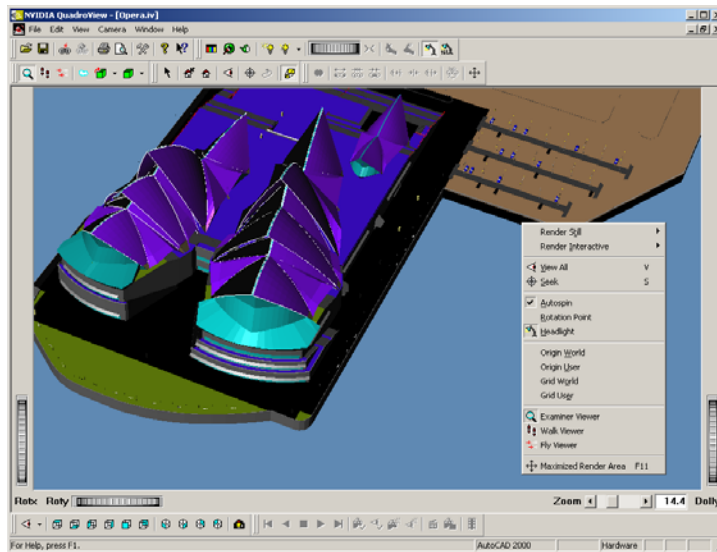


Figure 12. Model Display Automatically Loaded When Read into AutoCAD

As Figure 11 shows, POWERdraft offers 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.

NVIDIA QuadroView lets users view models in stereo. To enable this, the user must turn on stereo in the OpenGL Control Panel (Properties/Settings tab/Advanced/specific product name tab/OpenGL Settings/Performance and Compatibility Options/Enable quadbuffered Stereo API). 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 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 13 is a screenshot of NVIDIA QuadroView with stereo configured.

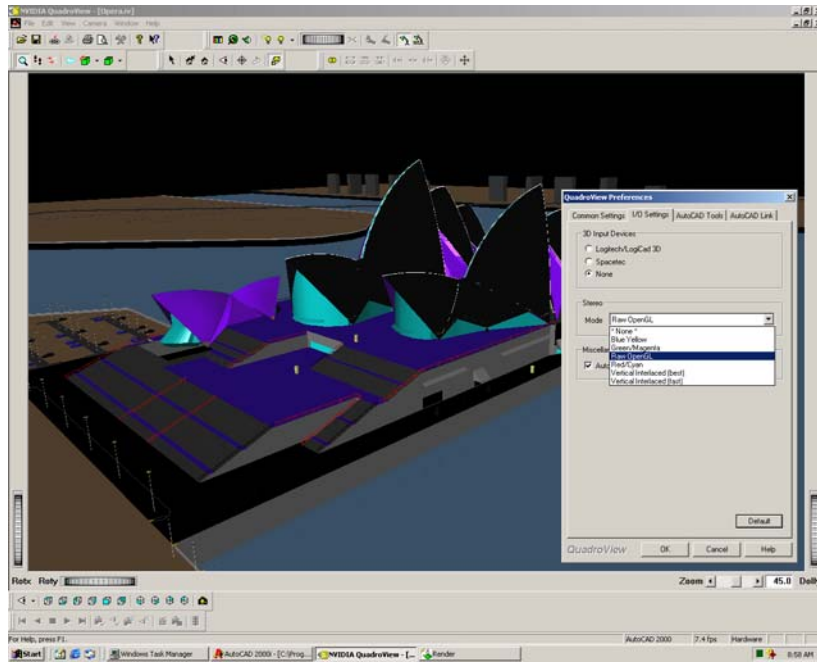
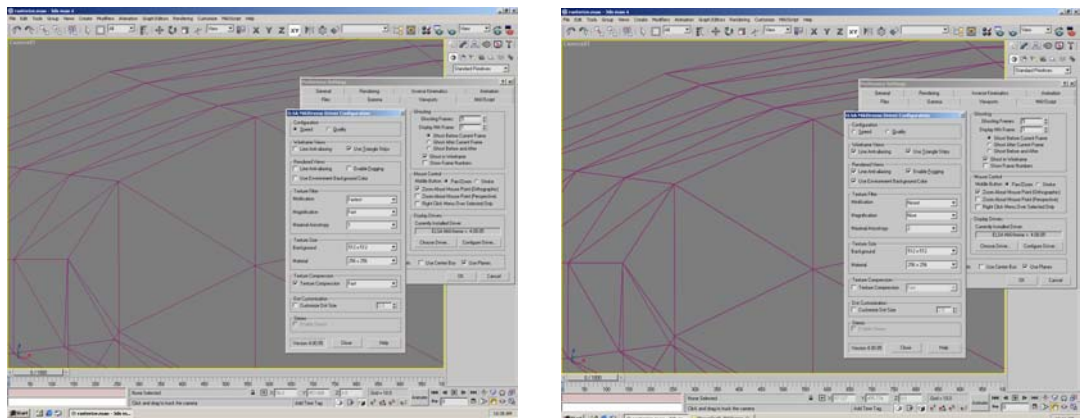


Figure 13. NVIDIA QuadroView Stereo Viewing Configuration

Like POWERdraft, MAXtreme offers 3ds max users many additional features. An important feature is the ability to adjust graphics quality to optimize for quality or performance.

Figure 14 shows a side-by-side comparison of cases optimized for speed and quality.



3D Studio max optimized for speed

3D Studio max optimized for quality

Figure 14. Speed and Quality Optimizations of MAXtreme Plug-In Driver for 3ds max

Along with the speed and quality optimizations, MAXtreme—like NVIDIA QuadroView—lets users view 3ds max scenes in stereo. Again, stereo must first be turned on through the OpenGL Control Panel and then enabled in the MAXtreme driver. Figure 15 is a screenshot of 3ds max with stereo enabled in the MAXtreme driver. As with NVIDIA QuadroView, the eye separation and parallax can be adjusted to ensure optimal viewing.

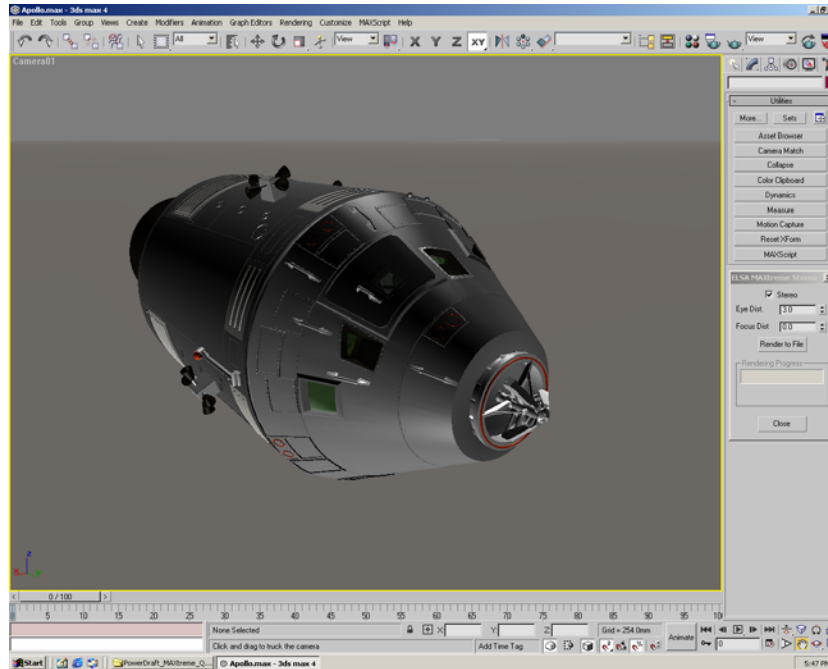


Figure 15. 3ds max Detailing Stereo Viewing with MAXtreme

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

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 document described the features that the NVIDIA Quadro workstation GPU family offers (over and above the consumer GeForce GPU family), which meet these workstation requirements.

Although 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 user’s workflow, this paper has shown how support for these features can increase productivity. The productivity benefits of software features such as workstation application support and the application productivity tools of POWERraft, 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 NVIDIA Quadro GPU families, the advantages move to a new level. In examples like the real-time preview, NVIDIA workstation GPUs enable step changes in workflow.

The competitive advantage and market leadership benefits from these step changes are often large, and 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 NVIDIA Quadro GPU family, it lays the foundation for these business opportunities. And it allows those who capitalize on them to reap substantial financial rewards.

References Appendix

1. OpenGL Programming Guide, 3rd Edition, Mason Woo, Jackie Neider, Tom Davis, Dave Shreiner. Addison-Wesley, ISBN 0-201-60458-2.
2. OpenGL Programming for the X Window System, Mark J. Kilgard. Addison-Wesley, ISBN 0-201-48359-9.
3. NVIDIA's Stereoscopic 3D Development Guide, <http://partners.nvidia.com/developer.nsf>
4. "StereoGraphics developers hand book", StereoGraphics Corporation, <http://www.stereographics.com/support/developers/handbook.pdf>
5. "3D Stereo Rendering Using OpenGL (and GLUT)", Paul Bourke, November 1999
6. <http://astronomy.swin.edu.au/pbourke/opengl/stereogl/>
7. "Calculating Stereo Pairs", Paul Bourke, July 1999. <http://astronomy.swin.edu.au/pbourke/stereographics/stereorender/>
8. Intel Pentium 4 Processor Optimization Reference manual,
9. <http://developer.intel.com/design/Pentium4/papers/>
10. IA-32 Intel Architecture Software Developer's Manual Volume 1: Basic Architecture, <http://developer.intel.com/design/Pentium4/manuals/>
11. IA-32 Intel Architecture Software Developer's Manual Volume 2: Instruction Set Reference Manual, <http://developer.intel.com/design/Pentium4/manuals/>
12. IA-32 Intel Architecture Software Developer's Manual Volume 3: System Programming Guide, <http://developer.intel.com/design/Pentium4/manuals/>
13. To download POWERdraft, MAXtreme and NVIDIA QuadroView go to <http://www.nvidia.com/>, look under "Download Drivers", and then select "Workstation Applications".
14. Information on the CADALYST C2001 benchmark can be found under: <http://www.cadalyt.com/reviews/cadbench/>
15. Information on the 3D Studio max test suite can be found under:
16. <http://www.discreet.com/support/max/index.html>, and then look under "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



Notice

ALL NVIDIA DESIGN SPECIFICATIONS, REFERENCE BOARDS, FILES, DRAWINGS, DIAGNOSTICS, LISTS, AND OTHER DOCUMENTS (TOGETHER AND SEPARATELY, "MATERIALS") ARE BEING PROVIDED "AS IS." NVIDIA MAKES NO WARRANTIES, EXPRESSED, IMPLIED, STATUTORY, OR OTHERWISE WITH RESPECT TO THE MATERIALS, AND EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY, AND FITNESS FOR A PARTICULAR PURPOSE.

Information furnished is believed to be accurate and reliable. However, NVIDIA Corporation assumes no responsibility for the consequences of use of such information or for any infringement of patents or other rights of third parties that may result from its use. No license is granted by implication or otherwise under any patent or patent rights of NVIDIA Corporation. Specifications mentioned in this publication are subject to change without notice. This publication supersedes and replaces all information previously supplied. NVIDIA Corporation products are not authorized for use as critical components in life support devices or systems without express written approval of NVIDIA Corporation.

Trademarks

NVIDIA, the NVIDIA logo, GeForce, and NVIDIA Quadro are trademarks or registered trademarks of NVIDIA Corporation. Other company and product names may be trademarks of the respective companies with which they are associated.

Copyright

© 2003 by NVIDIA Corporation. All rights reserved



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