December 1, 1998

The GXT3000P Graphics Accelerator

Introduction

The GXT3000P graphics accelerator brings a significant leap in performance over previous accelerators from IBM and, in combination with the RS/6000* Model 260 and its POWER3 microprocessor, leads the pack in workstation graphics. It offers a number of advanced visualization features such as two-sided hardware lighting, 3D texture mapping, video scaling, and stereo-in-a-window. It continues IBM's leadership in native hardware support for multiple application program interfaces (APIs) - OpenGL** 1.2, graPHIGS*, and X11R6.1. The GXT3000P provides outstanding performance and function executing advanced design, modeling and walk-through applications in a cost effective graphics accelerator design. The design takes advantage of IBM's leadership in manufacturing advanced integrated circuits to deliver a high performance graphics solution at a low cost.

GXT3000P Performance

The GXT3000P graphics accelerator represents a major advancement in IBM's offerings in the workstation graphics market. The table below shows how the GXT3000P compares to earlier IBM graphics products and to some competitors in the engineering workstation market.

Workstation	IBM RS/6000 43P-260	IBM RS/6000 43P-150	IBM RS/6000 43P-140 332 MHz	HP** C240	SGI** Octane** 250 MHz	Sun** Ultra 60 mod 1360
	uni- : multi-processor	uniprocessor	uniprocessor	uniprocessor	uniprocessor	uniprocessor
Graphics	GXT3000P	GXT3000P	GXT800P(T)	fx6	Maximum Impact	Elite3D m6
Benchmark[a]						
PLBwire93	436.4 : 625.3	257.7	157.4	427.3	n/a[b]	412.0
PLBsurf 93	598.1 : 843.0	467.0	264.9	839.6	n/a	658.4
CDRS-03	248.99	94.50	40.97	200.00	74.11	138.92
CDRS-04	248.05	96.72	n/a	n/a	n/a	n/a
GLperf line	7.05M	4.9M	1.5M	n/a	3.80M	n/a
GLperf triangle	5.19M	3.7M	924K	n/a	2.01M	n/a

a. Notes for benchmark data are shown at the end of the paper
b. n/a: Results are not available

The graphics performance of a system depends on both the capabilities of the graphics accelerator as well as those of the system processor and its associated memory and bus subsystems. In the case of the GXT3000P, the system processor is responsible for calculating the geometry of a scene while the graphics accelerator does the lighting calculations and the rasterization. Because of this split in the work load, the GXT3000P has the potential to support future, more powerful systems since today's systems do not exploit all of the accelerator's capability when geometric calculations are the limiting factor.

The numbers in the table below show the performance of one pixel, flat shaded lines with and without Z-buffering (shown in the table as "Peak 3D Lines" and "Peak 2D Lines" respectively) and one pixel, smooth shaded triangles with Z-buffering ("Peak Triangles") measured on the 43P-260. These numbers are compared with measurements taken through a play-back program that is designed to keep system performance from being a bottleneck. The difference between the two sets of measurements is an indication that the GXT3000P has the potential to improve its performance on these small primitives as more powerful systems become available in the future.

Millions of primitives per second	Peak 2D Lines	Peak 3D Lines	Peak Triangles
Graphics subsystem limit	17.6	10.4	10.3
RS/6000 43P-260 200MHz w/ GXT3000P	10.3	8.3	8.26

The GXT3000P Frame Buffer

The frame buffer of the GXT3000P uses Mitsubishi's 3D-RAM to provide high performance 3D rendering. The buffer is composed of sixteen chips organized as 1280x1024 pixels with a depth of 128 bits per pixel. The bits of a pixel may be organized in one of three ways to support different applications. These three formats are shown in the table below. The 32-bit color mode provides for double buffered, true color applications. The 16-bit color mode organizes each pixel into four buffers - front/back and left-eye/right-eye for use with stereo-in-a-window animation. The 8-bit index mode is also double buffered and supports OpenGL's color index mode applications.

Frame Buffer Pixel Formats

Pixel Format	Color Buffer A				Color Buffer B				Stencil/Depth		Auxiliary		
32-bit color	A	R	G	B	A	R	G	B	S	Z	W	O	U
16-bit color	$a_l a_r$	$r_l r_r$	$g_l g_r$	$b_l b_r$	$a_l a_r$	$r_l r_r$	$g_l g_r$	$b_l b_r$	S	Z	W	O	U
8-bit index	x	x	x	I	x	x	x	I	S	Z	W	O	U

Key to the Frame Buffer Mode Table:
ARGB = 8-bits each of Alpha, Red, Green, Blue
S: 8-bit Stencil, Z: 24-bit depth buffer, W: 8-bit Window ID, O: 8-bit Overlay, U: 16-bit Utility buffer
$a_l a_r r_l r_r g_l g_r b_l b_r$=Stereo: 4-bits each left/right of Alpha, Red, Green, Blue
I: 8-bit Index, x: reserved

In all color modes, a 24-bit depth buffer and eight bits of stencil buffer are available. In addition, there is an 8-bit overlay buffer which is used for the user interface. Since such user interface functions as pop-up menus have their own buffer on the GXT3000P, they do not displace data in the primary buffer. This makes the user interface respond more quickly and makes animation in the primary buffer smoother.

Along with the primary color buffers and overlay buffer, the window ID bits of the pixel are scanned out as per-pixel tags to the palette DAC (Digital to Analog Converter). The palette DAC uses this information, to display a mixture of pixel types concurrently on the screen. Thus, there can be a mix of animated stereo, static true color, and other windows all displayed simultaneously. The remaining bits are for utility functions. For example, these are used to store clipping planes so that odd shaped windows are supported with minimal performance degradation.

GXT3000P Interfaces

The GXT3000P uses industry standard interconnects for both the card's interface to the system and for the monitor and other output interfaces.

System Interface

The accelerator is a 64-bit PCI revision 2.1 compliant card (target and master) but occupies two 20.3 mm pitched slots with a height of 122.5 mm. The PCI bus interface chip is designed to operate at up to 66 MHz.

The Monitor Interface

The GXT3000P graphics accelerator drives a standard DDC-2B monitor interface. Screen sizes of 1280x1024 at up to 85 Hz and 1024x768 at up to 120 Hz are supported. The 1024x768 at 120 Hz mode is well suited to stereo display with minimum flicker. In addition to the monitor interface, the card has a VESA** standard stereo display output.

GXT3000P Supported Software Standards

The GXT3000P runs under AIX* 4.3.2 and supports a number of graphics application programming interfaces including X11R6.1, graPHIGS, and OpenGL 1.2. It was the first graphics accelerator in the industry to support OpenGL 1.2.

In addition to the base OpenGL function, a number of OpenGL extensions are supported. For example, OpenGL extensions for multi_draw_arrays and compiled_vertex_array allow the GXT3000P to render with much higher speeds than otherwise. Other currently supported OpenGL extensions are listed below.

EXT_abgr	EXT_bgra	EXT_blend_color
EXT_blend_logic_op	EXT_multi_draw_arrays	EXT_packed_pixels
EXT_polygon_offset	EXT_rescale_normal	EXT_subtexture
EXT_texture3D	EXT_texture_object	EXT_vertex_array
IBM_rasterpos_clip	IBM_vertex_cull	IBM_texture_mirrored_repeat

GXT3000P Hardware Performance Features

The GXT3000P provides a broad range of hardware performance assistance to the various software APIs used on IBM RS/6000 workstations. While some of the hardware performance features are intended for use by a specific API, others are used by several of the APIs or serve to speed the transition from one API to another during system context switches.

General Performance Features

For example, the chips incorporate special control mechanisms for moving the on-chip graphics context data to and from the system to allow rapid swaps between 3D graphics applications and the 2D graphical user interface (GUI) that provides the environment in which the 3D applications run. The graphics context data includes such information as the current color, drawing modes, line styles and stipples, and so on that controls rendering.

The bus interface chip also provides dual command input FIFOs, one for the current 3D API and one for the 2D GUI, so that the need to synchronize the accelerator during context switches is minimized. In addition to these two command input FIFOs, there is an asynchronous frame buffer interface that bypasses the rendering pipeline and allows the frame buffer to be read from and written to directly without a context swap.

The bus interface also provides a high function DMA (Direct Memory Access) controller that is exploited by the APIs. This DMA controller can retrieve commands from system memory and feed them to either the 2D or 3D command FIFO. Commands to the DMA controller itself may be imbedded into this data stream causing it to automatically switch between the two. In addition, the controller will perform sub-image DMA blits where it extracts a small image from inside a larger image, striding across unneeded sections of the larger image as required. The controller uses scatter-gather techniques to access system memory so that the commands in memory or blitted images may be in non-contiguous memory locations.

Performance Features for the User Interface

A number of the performance features of the GXT3000P are oriented toward the 2D GUI. For example, five sets of hardware rectangular clippers are provided to allow the GUI manager to protect regions of the frame buffer from updates by the 3D APIs. In addition, the frame buffer provides utility planes that are used to mask off regions of the buffer when the five clippers are not sufficient.

Hardware directly supports elements of the X11 interface such as the styled lines and stipple masks used by the API. Also, the XY coordinate interface to the accelerator is dual mode, supporting both the packed, 16-bit integer format native to the GUI as well as the unpacked, fractional XY format used by the 3D APIs. The hardware provides window offset registers that allow applications, both 2D and 3D, to work directly in window coordinates rather than in screen coordinates.

In order to maximize image text performance with minimal bus overhead, hardware performs a one-bit-per-pixel to index or to true color mode expansion when blitting. This unique blit function takes advantage of the 3D-RAM's special modes to speed this operation.

Performance Features for OpenGL

While many of the hardware performance assistance features are equally useful to the 2D GUI and to other 3D APIs, it is convenient to think of the features of the GXT3000P as they apply to OpenGL.

In the areas of data movement and image processing, the hardware provides not just the high function DMA controller and one-bit-per-pixel conversion mentioned before, but it can accept a wide variety of input pixel formats and sizes and automatically convert them to any of the three native formats of the frame buffer. It converts back from the native to the original pixel format when the data is read from the frame buffer. OpenGL color processing is also performed on-the-fly as blit data is sent to the frame buffer. And, the data may be passed through scaling hardware to grow or shrink the image.

In applications using OpenGL, the XY coordinates are handled in sub-pixel format. Not only the window offsets mentioned earlier, but the view port offset is added in hardware and a scissor rectangle is provided in addition to the five clippers used by the 2D GUI.

The GXT3000P provides a high function lighting engine in hardware. This engine supports up to eight light sources positioned at infinity with the full set of ambient, diffuse, and specular components for each light source. The lighting hardware supports per-vertex ambient, diffuse, specular, and emissive material parameters independently for both the front and back sides of objects as well as model ambient lighting. The specular light component may be added either before or after texturing as desired.

The rasterization hardware provides for both flat and Gouraud shading. Polygon depth offset is supported. Line and point rasterization includes both wide and antialiased modes. Hardware stippling support for both lines and surfaces is provided. In the GXT3000P, fog factors are interpolated over an extended range thus providing more realistic effects when polygons are clipped.

The accelerator hardware supports the full set of OpenGL per-fragment operations including alpha test, pixel ownership, stencil operations, depth (Z) testing, dithering, blend and logic operations. The implementation of these operations is split between the rasterization chips and the Mitsubishi 3D-RAM.

Performance Features for graPHIGS

While almost all of the GXT3000P performance features mentioned above in regard to OpenGL are equally applicable to graPHIGS, IBM's experience with this API has shown that some unique hardware support is needed to maximize its performance.

For example, the specularity in the graPHIGS lighting equation is slightly different than in the OpenGL equation so the accelerator hardware allows either to be used. Also, PHIGS supports the definition of "accurate" line styles, commonly used by drafting applications, and this too is supported directly in hardware. The PHIGS color processing function is another area where unique hardware is provided that is not required by OpenGL. Color comparison functions are also supported by the GXT3000P. Finally, PHIGS defines a depth test function where the display of an entire polymarker or text string is conditioned on the marker's reference point passing a depth test. This function is supported by the hardware.

Performance Features for Texture

The GXT3000P has a number of features that boost its texture performance well above that of previous generations of IBM workstation graphics accelerators. It has a separate set of SDRAM memory chips attached to each of the rendering pipelines. As discussed below, this four-way redundant memory allows each pipe to fetch texture data independently thus greatly enhancing performance. This texture memory can support up to four texture maps of 4 megabytes each.

The GXT3000P texture hardware supports the broad range of texture magnification and minification filter modes defined by OpenGL. These include 1D and 2D filters for point, linear, bilinear, and trilinear mipmap modes. The GXT3000P is one of the few graphics accelerators in the industry to support 3D textures in hardware. For 3D textures, modes of point, linear, and bilinear are supported. The texture maps may be stored as true color at four or eight bits per color band, as four or eight bits per band of luminance-alpha, or as 8 bits of luminance data.

Texture with borders is supported using any of the standard clamping modes defined by OpenGL as well as several of the clamping modes defined in some of the OpenGL extensions. The complete set of OpenGL texture blending functions and modes are supported. Finally, as mentioned earlier, it is possible to add the specular component of the lighting either before or after the texture application, a rare feature but highly desirable for realism.

Other performance features

In addition to the hardware performance features of the GXT3000P that are associated with the APIs discussed above, there are some other capabilities that are available but are not exposed by those APIs. These include a color keyed blit that will block writing incoming pixels in a selectable range of colors and support for up and down scaling of both RGB and YUV format pixels. This improves the ability to handle such functions as video data display.

With future extensions to texture functional capability in mind, the hardware supports two sided textures by calculating the direction of the facet normals to select between front and back facing texture maps. Thus any object can be rendered using two independent texture maps that will be selected automatically depending on the orientation of the object.

Hardware overview

The figure below shows a high level overview of the GXT3000P graphics accelerator. In addition to the 3D-RAM frame buffer and SDRAM texture memory, the accelerator has four large application specific integrated circuits (ASICS): a bus interface chip, a lighting and blit management chip, a rasterizer chip, and the palette DAC.

Bus Interface Chip

The bus interface chip, as mentioned earlier, provides the dual input command FIFOs and a high function DMA controller for the accelerator.

Lighting and Blit Chip

The lighting and blit chip accepts commands from the bus interface chip. It processes blit commands by converting the incoming pixel data stream into one of the native frame buffer formats, performs any scaling required, and then distributes the data to the appropriate rasterizer chip.

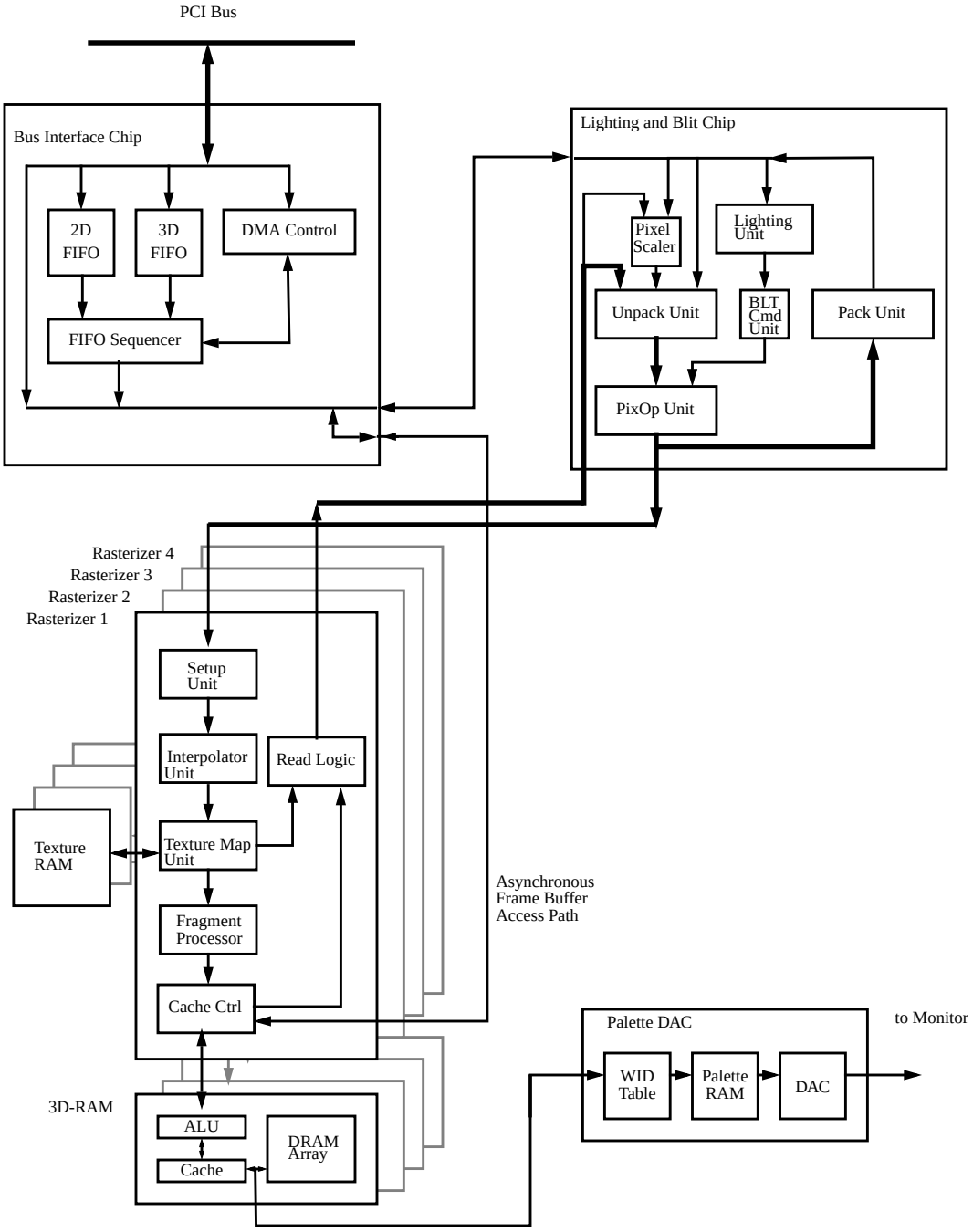
In addition to the blit commands, this chip also accepts drawing commands and vertex data from the bus interface chip. Vertex data can contain per-vertex material definitions and normals that are used to calculate the lit color of each vertex. The results of these lighting calculations are passed as vertex colors to the rasterizer chips. In contrast to blit commands and pixel data, drawing commands and the associated vertex data are broadcast to all four rasterizers at the same time.

Rasterizer Chip

To maximize performance, the four rasterizer chips operate asynchronously during most operations. Each rasterizer chip drives one fourth of the frame buffer interleaved on a pixel column by pixel column basis. These chips contain most of the setup and interpolation logic for rasterizing the drawing primitives. And, each processes textures from its own copy of the texture memory, so that there is four times as much texture memory physically on the card as the amount seen by the user. They also contain memory controllers for the 3D-RAM and texture SDRAM. Pixel fragment processing is split between these chips and the 3D-RAMs.

3D-RAM

The 3D-RAMs are an intrinsic part of the rasterizer data path. They handle much of the OpenGL per-fragment operations as well as the pixel logic operations for the 2D GUI. The cache and ALU inside the 3D-RAMs can be thought of as a special purpose SIMD (single instruction, multiple data) computer with the rasterizer chip acting as the tag RAM for the cache and as the controller for the data path.



Each of the 3D-RAM chips has a 256-bit wide path between its cache and the DRAM array. The sixteen 3D-RAMS thus have a total bus width of 4096-bits. This provides a much greater band-width to memory than conventional designs.

The Palette DAC

The palette DAC is a customized version of the IBM RGB640 palette DAC modified to accommodate the 3D-RAM interface and to support advanced features used in engineering workstations. It has a 2Kx8 palette RAM for each color band. This RAM can be used as eight independent 256 entry palettes or partially subdivided into smaller 64 entry palettes for even more independent palettes. The palette DAC also has a gamma correction RAM table that can be activated on a pixel-by-pixel basis. The window ID byte that is scanned in from the frame buffer is allocated as five bits to control the primary buffer and the remaining three to control the overlay buffer. Thus there can be up to 32 primary and 8 overlay window IDs concurrently on the screen. Each window ID can independently select a pixel format and palette, enable gamma correction, and select from A/B animation or left-eye/right-eye buffers. In typical usage, many windows will share a common window ID but, a few will make use of unique IDs. The large number of independent window IDs and palettes minimizes the annoyance of color flashing when several applications that use large color palettes are run concurrently.

Other features of this chip include a 10-bit monotonic DAC that is capable of up to 160 MHz operation in this application and of even higher speeds in other applications. The chip also provides both cross hair and 64x64 sprite hardware cursors.

Technology

The GXT3000P takes advantage of key IBM integrated circuit technology. The integrated circuits are implemented in IBM's 5se (0.27 micron) and SA12 (0.18 micron) silicon. There are over 40 million transistors in the seven chips. The chips are mounted in IBM's industry leading flip-chip ceramic ball grid array (CBGA) packages which provide unparalleled performance and density as required by today's.

Development Process

The GXT3000P is a major step in IBM's long history of offering graphics accelerators for RS/6000 workstations. It was developed by IBM's Visual Systems Group in Austin, Texas with assistance from IBM groups around the globe including Israel, India and IBM Microelectronics Division.

The design points of this graphics accelerator were created from requirements gathered from customers and end users of a variety of application suites mainly in the Automotive and Aerospace industries.

Key to the design is the Visual Systems Group's extensive experience with OpenGL, PHIGS, and X11 standards. Experts in these APIs worked closely with the ASIC designers to help ensure that performance critical areas were supported in hardware.

In designing the ASICs, the group applied an innovative design methodology using a combination of industry standard design languages and IBM proprietary tools. Extensive simulation was done before the chip designs were released to manufacturing using both software simulation and hardware emulation to insure a rapid development cycle.

Summary

With the introduction of the 43P-260 utilizing the POWER3 microprocessor in either uniprocessor or multiprocessor configurations, the GXT3000P provides a single workstation solution capable of both performing advanced design and modeling as well as the execution of compute intensive analysis and simulation applications on the same platform. This enables the advanced engineering community to increase productivity and reduce multiple platform requirements. This will result in reduced development cycle times and reduced capital overhead costs for each and every advanced user.

Also offered with the IBM Model 43P-150, the GXT3000P provides significantly improved graphics rendering capabilities over today's model 43P-140 and the GXT800P in a cost effective desktop package. The GXT3000P provides industry leadership performance in many wireframe, solids modeling and walkthrough applications and is offered at a truly outstanding cost to our engineering customers.

Notes:

* The following terms are registered trademarks of International Business Machines Corporation in the United States and/or other countries: AIX, IBM, RS/6000. The following terms are trademarks of International Business Machines Corporation in the United States and/or other countries: graPHIGS.

** Octane and OpenGL are trademarks or registered trademarks of Silicon Graphics, Inc. in the United States and/or other countries. VESA is a registered trademark of the Video Electronics Standards Association in the United States and/or other countries. Other company, product and service names, which may be denoted by a double asterisk may be trademarks or service marks of others.

IBM performance figures were measured in a development-level system environment and are presented for illustrative purposes only. While these values should be indicative of the performance of generally available systems, this cannot be guaranteed. Other performance figures were taken from the following references:

- PLB benchmarks are from http://www.spec.org/gpc/plb/plb.summary.html on December1, 1998. These benchmarks are geometric means of literal and optimized Picture Level Benchmark (PLB) tests for 3D wireframe and 3D surface tests. The benchmark and tests were developed by the Graphics performance Characterization (GPC) Committee. The results shown used the graPHIGS API. Larger values indicate better performance.
- CDRS-03 benchmark is from http://www.spec.org/gpc/opc/opc.cdrs.summary.html on December 1, 1998. This benchmark is the weighted geometric mean of individual viewset metrics. The viewsets were developed by ISVs (Independent Software Vendors) with the assistance of OPC (OpenGL Performance Characterization) member companies. Larger values indicate better performance.
- The GLperf application measures the performance of OpenGL 2D and 3D graphics operations. These operations are low-level primitives (points, lines, triangles, pixels, etc.) The GLperf program was developed by the OpenGL Performance Characterization (OPC) group. More

information on GLperf can be found at http://www.spec.org/gpc/opc.static/index.html. Larger values indicate better performance.

- Line Fill Rate - Display list mode, 3D, RGB, flat shade, Z buffer 10 pixel from http://www.spec.org/gpc/opc/glperf_publish/data/results/SGI/octane_250_mxe/summary.html#28 on December 1, 1998.
- Triangulate Fill Rate - Display list mode, 3D, RGB, smooth shade, Z buffer 25 pixel from http://www.spec.org/gpc/opc/glperf_publish/data/results/SGI/octane_250_mxe/summary.html#71 on December 1, 1998.