COMPAQ

Technology Brief

April 1999
0054-0499-A

Prepared by Workstation Division

Compaq Computer Corporation

Contents

Graphics: The Workstation Difference

Abstract: Graphics are the most visible and most critical part of workstations. Focusing on 3D graphics, this paper explores the role and importance of graphics and recommends graphics products for specific needs.

A discussion of graphics implementation issues is included, which examines when graphics operations make sense to implement in hardware and which should remain in software. Finally, key graphics interfaces -- both software and hardware -- are explored.

This paper should increase understanding of workstation graphics and help in choosing graphics devices.

Help us improve our technical communication. Let us know what you think about the technical information in this document. Your feedback is valuable and will help us structure future communications. Please send your comments to: Russ.Doty@compaq.com

Notice

Graphics: The Workstation Difference
Technology Brief prepared by Workstation Division

First Edition (April 1999)
Document Number 0054-0499-A

Workstations and Graphics

Graphics and multimedia are driving the evolution of workstations. Graphics have always been a fundamental part of a workstation. Today, they are one of the major differences between a workstation and a personal computer.

User demands for higher resolution, more performance, greater realism, and new types of interaction -- such as video conferencing and voice input -- require the utmost performance from all parts of a workstation. To address these ever changing needs, Compaq offers a family of graphics products for Professional Workstations to meet diverse application requirements ranging from 2D and multiple display 2D to entry-level 3D to high-end professional 3D.

Support for multiple displays and very fast 2D, including window and menu performance, are critical to applications in the financial markets as well as electronic design automation parts of the geographic information systems market. Compaq offers the Matrox Millennium G200 and Productiva 100 Quad Multi Monitor Series to meet the needs of professionals in these areas.

Many applications require text, menus, and simple graphics. Applications in this area would include software development, personal productivity, databases, and networking. Users of these applications require low cost combined with good performance and high display quality. Compaq offers a variety of low cost, high-performance workstations that support the ELSA Gloria Synergy to meet these users needs.

Other applications are graphics intensive, and demand highest graphics performance. These are often technical, Computer Aided Design (CAD), Digital Content Creation (DCC) or analysis applications. As these applications are often used by people in workgroups, cost effectiveness is essential. In this space, we encounter mechanical and electrical CAD users, Geographic Information Systems, and many science and research users. Compaq PowerStorm 300 and 350 graphics accelerators are designed to provide high performance and high value for these users.

The high-end consists of applications where the highest level of performance and realism is required and cost is secondary. These systems are typically used for various types of visual simulation and scientific visualization. Perhaps the most familiar examples are flight simulation and virtual reality, but other examples include scientific visualization, assembly modeling, styling, and architectural walkthrough. Other high-end applications include assembly modeling, or electronic mockup, the technique of designing a product as a single unit -- an entire automobile, airplane, or ship, for example. This approach allows checking for mechanical interference and seeing how well parts fit together. It is a revolutionary approach for detecting many problems early in the design cycle. For example, if the entire engine compartment of a car is in a single model, you can see if you will be able to get a wrench on the bolts of an alternator bracket, or if the water pump or serpentine belt are in the way! In fact, you can see if you will be able to get the wrench (with your hand and arm) to the bolt and then turn the wrench by simulating this movement in the computer.

Assembly modeling requires massive computing power to work with detailed solid or surface models, plus the graphics power to interactively work with these large datasets. High end Compaq PowerStorm graphics products are well suited to these needs. The combination of a Compaq Professional Workstation XP1000 with PowerStorm 300 graphics or Compaq Professional Workstation SP700 with PowerStorm 600 graphics delivers high performance rendering and texture mapping to meet the most rigorous demands.

Like assembly modeling, styling uses large geometric models, but its goal is to produce visual realism by creating an image that is indistinguishable from a photograph of the finished product.

Styling requirements include the manipulation of large datasets, and the use of sophisticated graphics algorithms -- like advanced lighting models and texture mapping -- for the highest possible level of realism. In many ways, styling is as much an artistic endeavor as it is an engineering problem. Scientific visualization has been one of the driving forces in high-end graphics. Many types of scientific analysis can best be understood when the results are shown graphically. An example that may be familiar is molecular modeling. Most people have seen computer images of molecules containing a few dozen or so atoms represented as balls. Much more challenging for both researchers and computers is molecular biology, where the goal might be to show one molecule of 50,000 atoms (I think this goes too deep)docking with another molecule of 30,000 atoms! In addition to manipulating this much raw data, the researcher needs to rotate and view the molecules and animate the motion of docking. Delivering smooth animation of this amount of data is critical to enable researchers to understand the interaction and behavior of the molecules. Another type of scientific visualization used in the petroleum industry involves determining the shape, volume, and characteristics of underground structures. This *reservoir modeling* processes massive amounts of seismic data to produce a complex 3D dataset. Display of this data requires manipulation of 3D volumetric data and use of *texture mapping* to show the characteristics of underground structures. With costs for drilling wells ranging up to several million dollars, an investment in tools which improve analysis of underground structures is extremely easy to justify.

High energy physics, computational fluid dynamics, large scale weather forecasting, multi-spectral satellite data analysis and many other scientific fields demand the utmost computer power for analysis and the highest levels of graphics performance to convey results.

These demanding applications are best supported by on multiprocessor or AlphaPowered workstations with the fastest available processors are recommended, plenty of memory and high performance storage and PowerStorm 300 or 350 graphics are recommended. On Intel based Compaq Professional Workstations where shaded image performance is critical, the geometry accelerated PowerStorm 600 is the best choice.

To effectively support the requirements of these demanding applications, graphics needs to be considered as part of a complete workstation system -- not as an isolated piece of hardware. The next section takes a closer look at the total system approach to application demands.

Graphics: The Total System Approach.

Key to understanding the role of graphics hardware is a simple observation: *all graphics operations can be done in software*. Any picture that can be produced by high end graphics hardware can be produced entirely in software with just a simple frame buffer for display. Through OpenGL, applications use the same software interface and graphics functions, independent of graphics hardware in use.

While all graphics can be done in software, this is not a solution to the broad needs of the graphics community as it does not deal with *high performance 3D*. Though the same image can be produced with any of the graphics options, there are major differences in how fast images can be produced, and in *graphics movement* that can be produced. A specialized type of hardware called a *graphics accelerator* speeds up graphics operations. Compaq offers a family of graphics accelerators, providing similar graphics features and functions at several levels of performance, thus matching user requirements.

For many years, the graphics community has debated the role of graphics accelerators. Some think specialized graphics hardware is the performance winner. Others believe that general purpose processors are on a more rapid performance growth curve, and that this performance growth, coupled with increased flexibility and the ability to support new algorithms, makes using software implementations on extremely fast processors the best choice.

Compaq's graphics provide support for both views. The Alpha processor in the Professional Workstation XP1000 is powerful enough to deliver application performance exceeding that of high-end geometry accelerator/graphics accelerator combinations on other, less high performance platforms.

On the other hand, specialized hardware *can* speed many graphics operations and provide greater system performance, although at a greater cost. An example of this is the PowerStorm 600, which accelerates both graphics rendering and geometry and lighting calculations. Used on the Intel processor based SP700 workstation, the PowerStorm 600 delivers higher graphics performance with full Intel processor software compatibility.

Industry Standard Interfaces

By using common interfaces, Compaq graphics insulate applications software from underlying hardware, providing flexibility to match graphics performance to user needs. In fact, the implementation of these key interfaces allows the applications to not only run on a variety of graphics hardware platforms, but to automatically take advantage of the full graphics acceleration potential of the hardware -- with no changes to the software.

Industry standards address 2D graphics, 3D graphics, video, and hardware requirements. The 2D graphics are included in the base windowing system, which is part of the operating system. Thus, 2D graphics are provided by X-Windows on UNIX systems, and by Windows on Windows NT systems. 3D graphics are provided through OpenGL, the leading open systems interface for high end graphics.

Graphics Hardware

Graphics system design requires optimization of graphics, software, computer systems, user requirements, market demands, and the interactions between these factors. An optimal, high performance, cost effective *total system* balances them all.

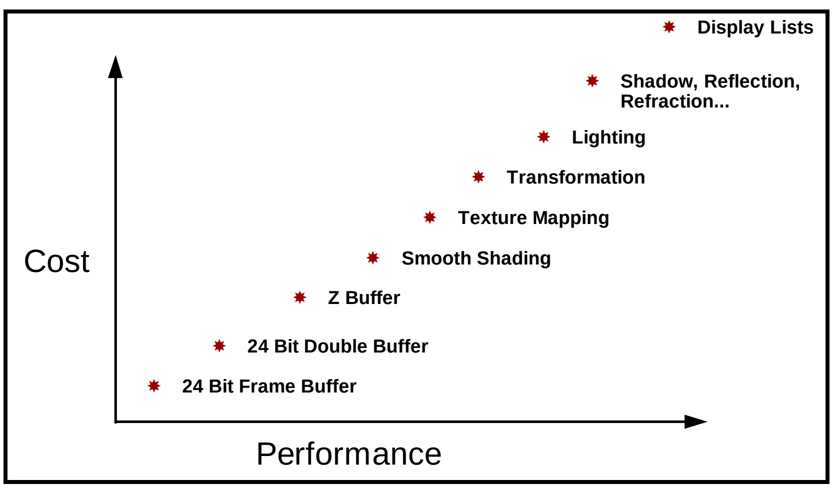

Figure 1: Moving graphics functions into hardware

Figure 1 shows the general relationship between performance and cost for a range of graphics functions that are often implemented in hardware. Graphics hardware options range from 8-bit, 256-color displays and 24-bit frame buffers that display true color images, to high end graphics subsystems that execute virtually all graphics in hardware.

While it is neither appropriate nor possible to move *all* graphics operations into graphics hardware, there are some operations for which specialized hardware makes sense. To determine which operations can be efficiently accelerated in hardware, you need to understand the impact of two parameters on graphics algorithms: *complexity* and *regularity*.

Complexity is a measure of the number of operations that must be executed in a graphics operation. Examples include a simple pixel write, which requires nothing more than "write this pixel value to this location in the frame buffer," and the slightly more complex Z-buffer algorithm, which requires "read current Z value from the frame buffer, compare with new Z value, and write the pixel with the lowest Z value to the frame buffer." These operations are very simple, easy to implement in hardware, and relatively easy to make faster than a pure software implementation.

At the other end of the complexity spectrum are operations such as "read application data and process it." This operation is entirely dependent on the application -- a flight simulator terrain model, a genetic engineering molecular model, and a mechanical CAD model of the front suspension of a car each would be dramatically different.

Regularity measures the amount of variation or choice in an algorithm. Some algorithms are extremely regular -- the previously mentioned pixel write requires no choices and has no variations, and the Z-buffer algorithm requires a single comparison and choice. But the manipulation of application objects like models involves extreme variation and, thus provides very little regularity.

An example of a graphics operation that can vary in both complexity and regularity is lighting. Some lighting, such as ambient or simple directional light, is relatively simple. These two lighting models can be placed in hardware fairly easily. On the other hand, lights can be directional, have "barn doors" (controls that allow screening the light, much like the flaps on professional theater lights), fall off with distance, have color, cast shadows, and other even more complex functions. These lighting models are extremely complex, very irregular, and very difficult to put into hardware. They are invariably done in software.

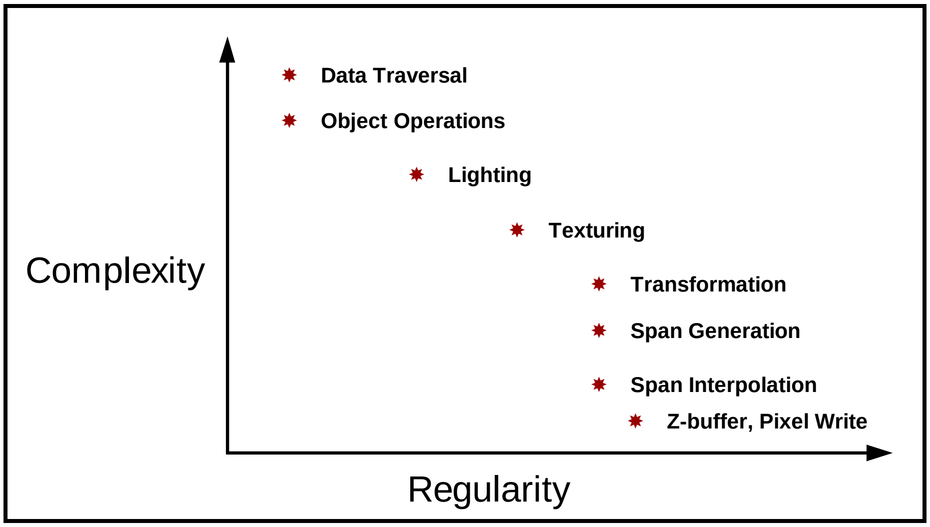

Figure 2: Complexity vs. Regularity

Figure 2 shows the complexity and regularity of several key graphics operations. They range from simple, regular operations such as pixel write and Z-buffer operations, through a set of operations with increasing complexity and decreasing regularity. This chart also shows operations can increase in complexity without decreasing in regularity; this is an important point, as these operations can be efficiently accelerated in hardware.

PowerStorm graphics integrate hardware and software components supporting a broad range of hardware graphics accelerators through a common interface. This insulates applications from underlying hardware, and allows you to make graphics purchasing decisions based upon performance requirements rather than arbitrary combinations of "which graphics hardware goes on which workstation platform and supports which applications."

OpenGL

OpenGL is quite simple in many ways. It can be described as a programming interface for three dimensional graphics, much like Direct3D, PHIGS or a number of proprietary interfaces. But it is also much more than that -- it is hardware, operating system, and window system independent. OpenGL is available from dozens of computer companies. It is a standard part of Microsoft's Windows NT, and is supported on virtually all versions of UNIX. OpenGL is the 3D interface for graphics hardware ranging from a simple 8-bit frame buffer to ultra high end graphics subsystems costing several hundreds of thousands of dollars.

There are several key aspects to OpenGL. The first is that it ***works***. Unlike some standards that were developed in committees, OpenGL was developed to support applications development. OpenGL has evolved and proven itself as an interface that meets the needs of software developers. The capabilities and functions of OpenGL deliver what programmers need. Because of this, OpenGL has rapidly become the preferred interface for applications requiring 3D graphics.

Second is that OpenGL is an open industry standard. The OpenGL specification is controlled by the OpenGL Architecture Review Board (ARB), which has representatives from Silicon Graphics, Microsoft, Compaq Computer Corporation, IBM, Evans & Sutherland, and a number of other companies. The ARB is responsible for maintaining the OpenGL specification and for guiding the evolution of OpenGL in a public forum. Any changes to the OpenGL specification must be formally approved by the ARB. The fact that the OpenGL specification is controlled by the leading workstation companies and is available on all leading workstations makes it the preferred interface for all application developers who need to have portable software.

Third, OpenGL spans both hardware and operating systems. OpenGL can be implemented entirely in software, supporting a simple frame buffer. Many types of hardware acceleration can be applied to OpenGL functions without impacting the interface. This means that workstation companies can develop special hardware to make OpenGL go faster -- without affecting the applications that call OpenGL. The fact that OpenGL graphics can be cleanly accelerated with specialized graphics hardware -- with no impact to application software -- is extremely advantageous to computer companies, application developers, and customers who purchase systems and applications. Compaq uses OpenGL to transparently integrate graphics hardware and software, providing a foundation for a complete family of graphics accelerators.

The ability of OpenGL to span operating systems is based on a single design decision: OpenGL does nothing but 3D rendering. Open GL extends the capabilities of any underlying windowing system. A windowing system provides 2D graphics functions, window and screen management, user interaction mechanisms, text display and manipulation, and similar capabilities. OpenGL simply adds 3D rendering support to the 2D windowing system. This allows OpenGL to support both X-Windows and Windows NT. These two windowing systems are radically different in the way they work, but the same OpenGL interface is available on both for applications running in either environment. Of course, all 2D operations performed by an application are dependent on the windowing system, but the 3D OpenGL interface will be identical across the operating systems.

AGP and PCI

Two hardware interfaces are commonly used for graphics accelerators AGP and PCI. AGP is the newer bus; it is essentially a customization and extension of the PCI bus to support graphics.

The AGP bus is up to 8 times as fast as the PCI bus, which offers clear advantages for graphics. However, as the analysis in the following sections shows, the PCI bus is adequate for today's 3D graphics workloads. This is supported by graphics benchmark results when are almost identical when AGP and PCI versions of the same graphics chip are tested in the same system.

AGP does have several advantages, and will begin to show its performance edge as graphics and system performance continue to improve.

PCI Bus

The PCI bus is a true industry standard. Unlike proprietary buses commonly used on traditional UNIX workstations, the PCI bus is used by many vendors, on many platforms, supporting many different processors and operating systems. Millions of PCI based systems from hundreds of companies are shipped each year, supported by thousands of high performance peripherals.

As a hardware specification, PCI has many advantages. It is a modern design, it runs at a very fast 33 MHz, and it is a full 32-bit interface. The base PCI bus delivers 133 MB/second of bandwidth, meeting the needs of today's systems.

This performance is well matched to the needs of 3D graphics systems. A 3D vertex used in rendering will contain between 12 bytes and 40+ bytes of data. The simplest case would be Gouraud shading with compressed X and Y coordinates. The data structure might look like this:

<table>
<tr><td colspan="2">X coordinate</td><td colspan="2">Y coordinate</td></tr>
<tr><td colspan="4">Z value</td></tr>
<tr><td>Red</td><td>Green</td><td>Blue</td><td></td></tr>
</table>

This example is using common computer conventions, where each row is a 32-bit (4 byte) "word", and the series of rows represent continuous memory.

A more complex case would support non-compressed coordinate values, shading, and perspective-corrected texture mapping. Non-compressed coordinate values eliminate the requirement for the CPU to perform data compression, thus reducing CPU load at the expense of additional bandwidth for data transfer. Texture mapping adds additional data to each vertex: U and V coordinates and, perhaps, perspective correction. Such a data structure might look like this:

X screen coordinate

Y screen coordinate			
Z value			
Red	Green	Blue	Alpha
Texture U coordinate			
Texture V coordinate			
Texture W (perspective) value			

When used in a triangle strip (common for high performance graphics), the PCI bus can handle roughly 10 million of the simple triangles per second. Using the full OpenGL interface, the PCI bus can handle 3-4 M triangles in the real world.

Today's graphics accelerators can display 3-5 M triangles per second, which is a reasonable match for the PCI bus. Further, few applications can produce more than 0.5-1.0 M triangles per second to feed into the graphics system. Thus, the PCI bus is seldom a bottleneck for graphics on today's systems.

The PCI bus supports multiple devices on a single bus. All of the devices share the bandwidth of the PCI bus. Several types of peripherals are beginning to approach or exceed the bandwidth of the PCI bus with a single device. These include graphics accelerators, network interfaces (such as Gigabit Ethernet) and storage controllers (Ultra2 SCSI and Fibre Channel). With multiple high performance devices, PCI bandwidth becomes a bottleneck.

This is addressed in two ways: The first is to provide multiple PCI buses in a single system. This is done with the Professional Workstation SP100 and XP1000, which both offer two fully independent PCI buses, thus doubling the PCI bandwidth of the system. In addition, the XP1000 incorporates a 64 bit PCI bus, which is fully compatible with 32 bit PCI devices and, when used with 64 bit PCI devices, doubles the available bandwidth.

The second approach is to use special purpose buses. The leading example of this is the use of AGP for graphics.

AGP

The AGP bus is the newest interface for graphics, and has received considerable publicity. In many ways, the AGP bus is an evolutionary growth of the PCI bus. In fact, the AGP specification notes that AGP is based on the 66 MHz PCI bus specification and adds a set of extensions to that specification.

AGP differs from PCI in two major ways. First is performance. The AGP bus allows up to 1 GB/s of bandwidth (with AGP4X). Second, AGP is a point to point bus. This means that only two devices can be on the AGP -- the system controller (usually the same chip that functions as memory controller) and the graphics device.

There are three AGP specifications: AGP 1X, AGP 2X and AGP 4X. They differ in performance, with AGP 1X providing 266 MB/s of bandwidth, AGP 2X providing 512 MB/s bandwidth and AGP 4X providing 1024 MB/s. AGP 1x is essentially an extension of the PCI 66 MHz interface. AGP 2X and 4X retain the 66 MHz bus of AGP 1X, but allow multiple data transfers on a single bus clock cycle. AGP 4X allows four data transfers on each clock cycle.

The implementation of AGP as a point to point bus enables much of AGP's performance. By eliminating multiple devices on the bus, signal integrity problems are greatly reduced and bus clock rates can be increased. Likewise, eliminating the need to determine which of several devices has control of the bus simplifies the software interface and improves performance there.

Another advantage of point to point connection is that the graphics device has the full bandwidth of the AGP bus -- it doesn't share this bandwidth with any other devices.

AGP is implemented to provide a high bandwidth, low overhead connection between the graphics device and system memory. This connection, combined with DMA transfers, allows effective and efficient communication between the system and graphics.

The major disadvantage of the point to point connection of AGP is that it only allows a single graphics device. You can't implement multihead displays with AGP based graphics. System vendors have overcome this problem by using both AGP and PCI graphics devices in a single system.

The major impetus for the creation of AGP was to allow textures for 3D graphics to be stored in system memory. As stated in the AGP specification:

> *"In general, 3D rendering has a voracious appetite for memory bandwidth, and continues to put upward pressure on memory footprint as well. As 3D hardware and software become more pervasive, these two trends are likely to accelerate, requiring high speed access to ever larger amounts of memory, thus raising the bill of material costs for 3D enables platforms. Containing these costs while enabling performance improvements is the primary motivation for the A.G.P. By providing up to an order of magnitude bandwidth improvement between the graphics accelerator and system memory, some of the 3D rendering data structures may be effectively shifted into main memory, relieving the pressure to increase the cost of the local graphics memory. Texture data are the first structures targeted for shifting to system memory…"*

Using main memory for textures is referred to as DIME texturing -- Direct In Memory Execution of Textures. DIME texturing supports large textures, but has two significant costs compared to dedicated texture memory on the graphics accelerator: lower graphics performance and system performance impacts caused by sharing main memory between graphics operations and CPU operations.

At the time AGP was created, memory prices were in the $40-$50 per MB range. Memory prices today are in the $1-$2 range. Thus, some of the original economic justification for AGP is greatly reduced. Low cost graphics cards with 16 MB to 32 MB of memory are becoming common.

Conclusion

This paper introduces the importance of graphics to workstations and some of the applications that high performance graphics enables. It explored several factors impacting the design and implementation of graphics hardware and the interface between graphics and the workstation.

Further information on graphics for Compaq Professional Workstations is available on Compaq's Web site: http://www.compaq/products/workstations

Graphics related information available from the Web site includes:

Compaq Professional Workstations Graphics Product Positioning, document 0053-0499-A

Professional3D: a New Graphics Classification, document 0054-0499-A

Graphics Performance: Measures, Metrics and Meaning, document 0054-0499-A