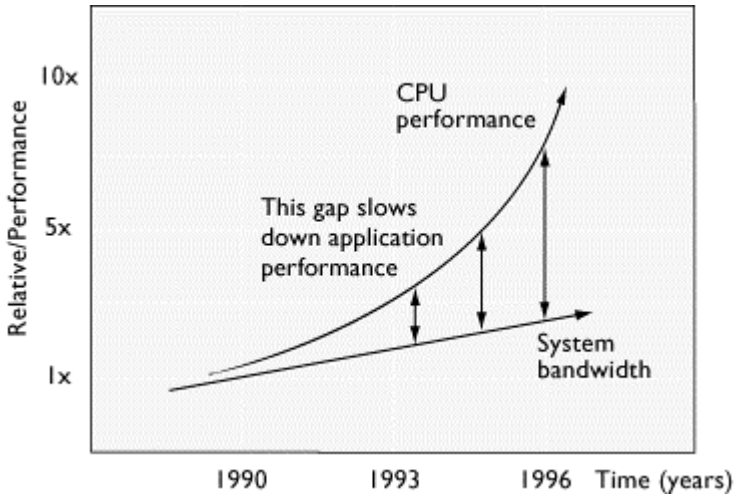
Introducing the Architectures for the Next Generation

With the introduction of the Silicon Graphics®O2® and Silicon Graphics® Octane® families of workstations, Silicon Graphics has defined the architecture for the next generation of desktop visualization. These new high-bandwidth, low-latency systems allow users to perform tasks never before achievable on desktop computers. The future trends of the manufacturing industry will demand fundamental capabilities such as complex solid modeling, as well as emerging needs for concurrent engineering analysis and design, in-context design, and advanced visualization on the desktop. These new architectures will be the vehicles to deliver outsatnding capabilities to engineering professionals, allowing them to perform complex tasks with ease.

The Limitations of a Traditional Computer Architecture

Microprocessor technology has experienced a remarkable and steady increase in performance over the last decade. Approximately every five years we have seen a tenfold increase in processor power. At the same time, the improvements in conventional system bandwidth have increased much more slowly--roughly two times every four years. The growing gap between microprocessor performance and system bandwidth is illustrated below in Figure 1. The result of this performance gap is that the speed and interactivity of the end-user application is limited not by the processor or graphics accelerator's ability to process, but by the systems ability to manage large amounts of data movement. Many of the current generations of processors and graphics accelerators have impressive stand-alone performance metrics, but the performance of these devices in a low-bandwidth system leaves users feeling underserved.

Figure 1: The growing performance gap between processors and system buses



A shared bus has been at the heart of traditional computer architectures since the introduction of the earliest computers at the University of Chicago in the 1950s. The shared bus has two primary drawbacks. The first is raw speed. The approach to increasing traditional bus speed has been to increase the number of data lines from 8- to 16-bits and then from 32- to 64-bits. The next logical step would be to push to 128-bits. The problem with this approach is that a 128-bit bus becomes unnecessarily expensive and impractical to implement. Furthermore, a 128-bit bus only increases performance by two times, which lags behind the tenfold increase in processing power. The other limitation of the shared bus is that all system traffic must take place one at a time on a single line. This is similar to the party telephone line. If more than two people try to have a conversation, the rate at which any two individuals can talk decreases in direct proportion to the number of people on the line. Today's applications have multiple processes taking place at once,

and often two separate processes within the machine will collide, causing both processes to run slower and share the party line.

One-to-One Architecture

Instead of competing for a shared bus, an ideal computer would allow every element of the computer to directly communicate with every other element using a private line that runs only between those two elements. This would allow the data transfer rate to be dramatically increased. It would also make the data transfer extremely predictable, since the connection between processing elements is not shared. This predictability would allow a stream of data, such as video playing off a disk, to avoid the risk of being interrupted by another random process such as the arrival of an e-mail.

The challenge then becomes how to take the individual one-to-one links between components and turn them into a complete system. The answer to this problem is the crossbar switch. A crossbar switch uses advanced packet switching technology to route messages directly from one processing element of the computer, say the CPU, to another element such as the graphics system. A true non-blocking crossbar allows multiple streams of data to flow independently from one point to another; they will not interfere with or block each other. The OCTANE workstation from Silicon Graphics incorporates this kind of crossbar switch in its system architecture. The heart of the OCTANE workstation is built around an eight-port non-blocking crossbar.

The OCTANE workstation makes use of dedicated hardware processing elements to optimize the performance of key computing tasks such as graphics processing or video compression. Each of these dedicated computing engines resides on a different arm of the crossbar switch, as shown in Figure 3.

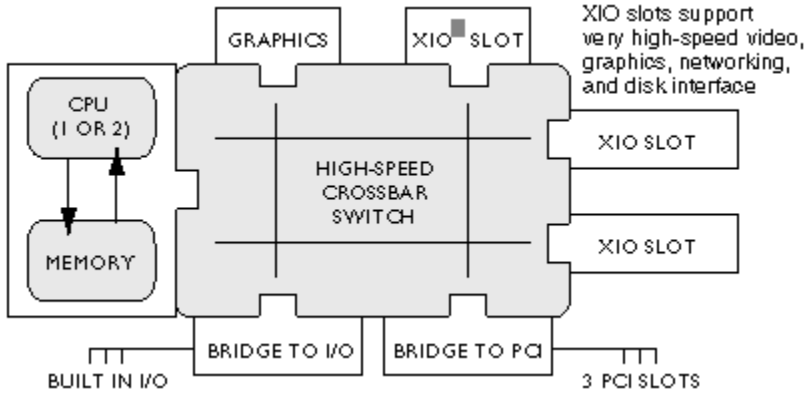


Figure 3: The OCTANE XIO Architecture

This allows the application software to break up the necessary tasks and assign each to the appropriate processing element (graphics, CPU, compression) for parallel execution. OCTANE can actually support two CPUs so the task at hand can be split between the two processors and then be executed twice as fast, or two separate processes can be concurrently run at full speed. This is especially important for tasks such as concurrent design and analysis. One CPU can provide interactivity to the solid modeling, while the other CPU can solve an analysis. For particularly difficult problems, both CPUs can be used to accelerate a single-threaded application. The crossbar architecture allows these various processing elements to communicate as fast as they can process data, and therefore they never have to wait for traffic on the bus to subside. From the perspective of the individual processing elements, the crossbar delivers infinite bandwidth--the processors will never have to wait for the system bus.

Unified Memory

The OCTANE system described above uses dedicated hardware that is optimized to perform a specific function, such as special texture memory. In contrast to this approach, the entry-level O2 workstation uses more flexible, repurposable hardware to bring the high-end desktop features down to unprecedented low price points.
The O2 system is built around a Unified Memory Architecture (UMA). UMA places high-bandwidth memory at the heart of the system. This memory effectively replaces the shared bus of traditional computer systems. Five dedicated processing blocks access this main memory. These processing blocks include the CPU, the imaging engine, the graphics engine, the compression engine, the video system, and O2 I/O. All of these processing elements access data from a single ultra-high-speed unified memory bank. This means that a variety of data types can pass through the system with

ease. The compression engine can process a stream of video, and then be easily accessed by either the CPU, the imaging engine, or the graphics display, all at full resolution and frame rate. Rather than copy data from one subsystem to another, the O2 subsystems can simply exchange pointers, and thus greatly reduce the performance penalty imposed by copying data. This architecture is illustrated in Figure 2.

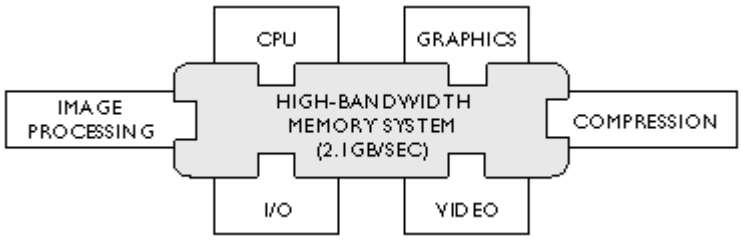


Figure 2: The O2 Unified Memory Architecture

The UMA architecture provides the one-to-one connectivity and major increase in system bandwidth that is necessary to run the radical new software applications emerging in the entertainment industry. In addition to the performance benefits of this approach, the UMA architecture brings a new level of flexibility to personal workstations. This flexibility is especially important in the entry-level machines, because engineering professionals are likely to use the O2 entry systems for many tasks ranging from advanced solid modeling to running a Microsoft® Excel spreadsheet. This type of platform needs to handle many different media types and perform many different functions. UMA provides this flexibility. An O2 workstation can display an engineer's solid model, capture the display as video and compress it to disk, then edit the resulting movie and drop it on the engineer's Web site. The graphics texture memory is limited only by the amount of system memory available, and the O2 workstation can be upgraded to take advantage of the processing power of the MIPS RISC R10000TM CPU.

It is the combination of performance and flexibility that makes O2 workstations so powerful. These unique architectures enable fundamentally new capabilities that will change the way manufacturing companies develop products.

Changing the Way We View Solid Models

There are three key features to the new desktop architectures from Silicon Graphics that make these the most powerful solid modeling machines available. The first is memory bandwidth. Most of the leading solid modeling software packages are very CPU-intensive. The unified memory of the O2 system has a peak bandwidth of 2.1 gigabytes per second. This is more than five-times the bandwidth found in most traditional architectures. This high-bandwidth, low-latency memory translates directly into faster model regeneration and interference checking. This accelerated model regeneration will allow more iterations to create a higher-quality product, or keep the same number of iterations and get to market sooner.

The second feature is high-speed, uninterrupted I/O. The one-to-one connections mean that a user can be loading a model directly from the UltraSCSI disk to memory, without the loading operation being interrupted by other processes taking place in the system such as the arrival of an e-mail. The model load times of the desktop machines from Silicon Graphics are the fastest available from any system.

The third feature is the outstanding graphics performance. Silicon Graphics has always been known for its powerful graphics accelerators, and these new architectures have the highbandwidth needed to keep the graphics pipeline full. The result is fluid interaction with even the most complex assemblies. The OCTANE systems use a hardware Geometry Engine to perform the transformation of the 3D model into an image that appears on the monitor. This hardware geometry engine acts like a powerful co-processor that off-loads the CPU of repetitive geometry calculations. One unique feature to the hardware Geometry Engine® processors of the Octane system is that it can be scaled--adding a second chip yields twice the performance. This is not true of CPU-based geometry. Adding a second CPU will not necessarily yield twice the geometry.

The O2 workstation is the right choice for component modeling and small to medium assembly modeling bringing high-end features and performance to an unprecedented new price point. The more powerful OCTANE systems are the right choice for larger components and assemblies. OCTANE also has enough power at a low enough price to allow component designers to view entire assemblies and see how their components interact with assemblies.

Analysis Becomes Part of the Design Process

Every engineer performs some type of engineering analysis be it on a calculator, spreadsheet, or supercomputer. As the power of desktop systems increases radically, the software tools have matured to the point where a designer can simply push a button and quickly arrive at an optimized design.

Engineering analysis software is computationally intensive, requiring strong floating-point performance, symmetric multiprocessing, high-speed memory and I/O systems, and 64-bit processing. Silicon Graphics has always been a leading provider of engineering analysis solutions because of the powerful and scalable POWER CHALLENGETM and Origin supercomputers. The new O2 and Octane systems bring many of these advanced capabilities to the desktop.

The graphics power of O2 and Octane workstations make them well-suited for pre- and post-processing of models. Most of the leading solid modeling applications have seamless integration into advanced engineering analysis packages. The high memory bandwidth of the new architectures, coupled with the floating-point performance and 64-bit processing of a MIPS R10000 CPU, means that solving complex analyses can now be done right on the desktop. The OCTANE system supports up to two R10000 processors. This means that the user can apply both CPUs to solving a single analysis, or a user can run an analysis on one CPU and continue to interact with and design a model using the second CPU.

Digital Prototyping Becomes Reality

Digital prototyping is an emerging technology pioneered by some of the leaders in the automotive and aircraft industries to optimize designs and reduce costs. With the growing power of desktop computing systems this technology is penetrating into more and more manufacturing companies. Digital prototyping allows engineers to review large assemblies and subassemblies and monitor part movements to digitally check for interferences without building an actual physical prototype.

Digital prototyping requires the fastest graphics available for real-time interaction with large models and texture mapping for realism. The Octane/MXI system is the most powerful desktop visualization machine available. This machine allows engineers to verify a design before the first metal is cut.

New Ways to Share Your Work

Both the O2 and the OCTANE workstations provide innovative multimedia capabilities. Combine these advanced media tools with the Web-infused operating system, and the result is a new way to communicate and share information. Using the screen-capture capabilities, an engineer can rotate and examine designs in real time while capturing a compressed video image and storing it on the system disk. The engineer can then edit the movie sequence using the intuitive media editing tools included with the operating system. Once the movie is done, the engineer drops it onto the OutBox personal Web server. This way a manager can review a design from her PC in the corner office, or a customer can look at a design from a notebook computer on the road.

In addition to screen-capture capabilities, the new Silicon Graphics® desktop machines support InPersonTM videoconferencing software and a shared 3D- whiteboard. Other software available includes SoftWindowsTM 95, a next-generation Windows® emulator that allows technical users to run standard Windows 95 applications right on their workstations, as well as networking software to share data with PC-based or Apple® computers.

Great Performance Requires Great Balance

In order to achieve optimum performance, system designers must create an architecture that balances the performance of the processors, graphics accelerators, and data I/O systems. Silicon Graphics has always built well-balanced systems. This is why application performance is always best on a Silicon Graphics machine.

Tom Gillis
Octane Product Manager