Entertainment Industry White Paper

New Computer Architectures for the Entertainment Industry

With the introduction of the Silicon Graphics® O2® and Silicon Graphics® Octane® families of workstations, SGI 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 entertainment industry will demand capabilities such as uncompressed nonlinear editing and animation of extremely complex and rich 3D sequences. These new architectures will be the vehicles to deliver outstanding capabilities to creative 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 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 the graphics accelerator's ability to process, but by the system's ability to manage large amounts of data movement. Many of the current generation of graphic accelerator processors have impressive stand-alone performance metrics, but the performance of these devices in a low-bandwidth system leaves users feeling underserved.

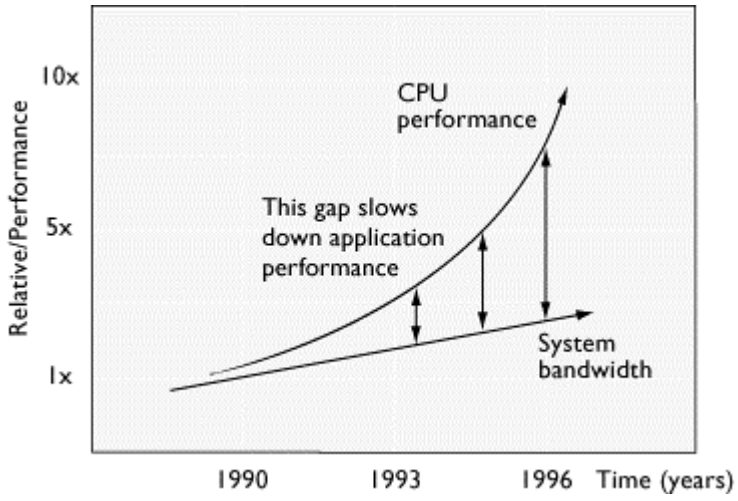

Figure 1: The growing performance gap between processors and system buses

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 allows every element of the computer to directly communicate with every other element, using a private line that only runs between those two elements. This allows the rate of data transfer to be dramatically increased. It also makes the data transfer extremely predictable, since the connection between processing elements is not shared. This predictability allows a stream of data, such as video playing off a disk, 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 from one point to another completely independently; they will not interfere with or block each other. The Octane workstation from SGI incorporates this kind of a crossbar switch in its system architecture. The heart of the Octane workstation is built around an eight-port non-blocking crossbar. As the intelligence of peripheral devices grows, the crossbar architecture allows devices to communicate peer-to-peer. This means that devices such as a video I/O card will stream data directly to a disk interface, and not consume memory bandwidth. Peer to peer transfers are a feature that will become increasingly important as end user data sets grow--such as the transition from video resolution to HDTV resolution. The Octane workstation has the capacity to support these peer-to-peer capable devices as they become available.

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 2.

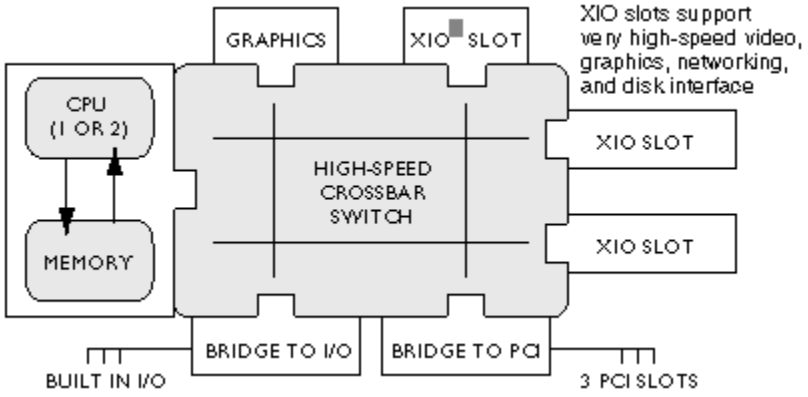

Figure 2: The Octane XIO Architecture

This allows the application software to break up the necessary tasks and assign each to the appropriate processing element (graphics, CPU, compression, etc.) 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. This is especially important for tasks such as rendering multiple frames of an animated sequence. 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.

Unified Memory

The Octane system described above uses dedicated hardware that is optimized to perform a specific function such as special texture memory. To complement this approach, the entry-level O2 workstation uses more flexible 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, imaging engine, or 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, thus greatly reducing the performance penalty imposed by copying data. This architecture is illustrated in Figure 3.

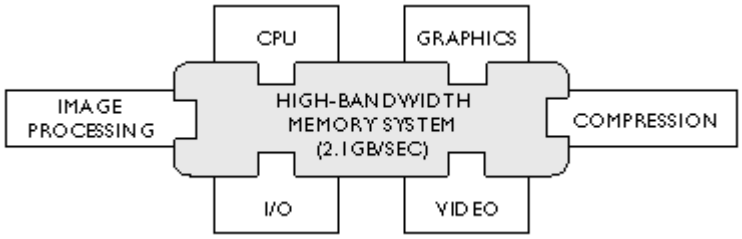

Figure 3: The O2 Unified Memory Architecture

The UMA architecture provides the one-to-one connectivity and major increase in system bandwidth that is necessary to run the radical emerging new entertainment software applications. In addition to the performance benefits of this approach, the UMA architecture brings a new level of flexibility to digital media computers with its ability to handle many different media types and perform many different functions. An O2 workstation can handle media streams of virtually any size and resolution, ranging from Web-based video to uncompressed real-time CCIR-601 to film resolution images. An O2 workstation can process JPEG and MPEG video. The graphics texture memory is limited only by the amount of system memory available, and the texture system can accept video as a texture for true 3D video effects rendered in real time.

The combination of performance and flexibility is what makes the O2 workstation so powerful.

New Capabilities in Editing

Octane is the ideal platform for uncompressed, nonlinear editing. Octane provides two channels of 10-bit CCIR-601 video in and out, as well as disk interfaces (such as UltraSCSI and Fibre Channel) fast enough to store two streams to disk in real time. The base system comes equipped with two Ultra-SCSI buses, and three internal 3.5-inch drive bays that can store up to 27GB of data. Additional UltraSCSI or Fibre Channel interfaces can be added through PCI or through the high-speed XIO bus. Video processing can be performed in the Octane graphics subsystem. The graphics engine can treat full-resolution video as a texture and map it onto a polygon for truly unique 3D effects. Now when you tip back a page turn, you will see that it is really 3D, not the simulated 3D that many effects systems deliver today. Also, editors can create innovative new transitions such as shattering glass or animated flying frames. The combination of unique effects and uncompressed image quality creates an end product unlike anything anyone has ever seen.

Alternatively, the O2 workstation is the platform of choice for compressed editing or off-line uncompressed editing. O2 has the ability to process JPEG video in real time. Editors are not limited to a set number of streams; the O2 system can be programmed to process one, two, three, even four or more streams of video. Obviously, as the number of streams increase, the resolution or real-time performance begins to decrease. O2 can also process uncompressed video, so editors can perform off-line creative work and finish rendering on one platform on-line.

New Capabilities in Graphics

The entry O2 system demonstrates pure SGI performance. With 32-bit color, z-buffering, and hardware-accelerated textures, the O2 workstation is the ideal platform for creating graphics pages and moving text. The powerful O2 graphics allow users to bring true 3D to graphics creation for a unique, differ-entiated look, all created in real time. Add a third-party adapter, and you have a single 8-bit stream of CCIR-601 video in and out.

The Octane workstation is the ideal platform for on-air graphics. With two 10-bit CCIR-601 video I/O ports (video and key), on-air graphics can be generated in real time, blended with live video and brought live to air. Use a cluster of O2 workstations connected by the built-in high-speed networking to have multiple graphics creation stations feeding an on-air Octane system. Octane graphics can generate virtual set elements that enhance local programming such as the nightly news.

New Capabilities in Audio

All SGI workstations contain the most advanced audio I/O system in the computer industry. The base Octane workstation has eight tracks of 24-bit ADAT digital optical audio, two channels of AES3-ID serial digital I/O, and two analog I/O ports. The digital I/O capabilities are also available on a PCI card that can be added to the O2 or Octane workstations. Adding

multiple cards will allow users to have 32 or more tracks of audio I/O. Combine this with the native DSP processing power of the R10000 processor, and you have the most advanced digital audio workstations in the industry.

New Capabilities in 3D Animation

An Octane workstation with the high-powered Octane/MXI graphics system, dual R10000 processors, Octane digital video and compression, and high-speed networking is a 3D animator's dream machine. This collection of technology is powering the next generation of software applications from the leading animation developers. Octane is the right choice for doing complex work such as full scene animations and modeling or animation of complex, organic characters. The dual processors allow animators to work with wireframe models and see photo-realistic final rendered images previewed in real time. This is achieved by using one CPU for the interactive animation, while the other CPU is free to update the photorealistic rendered image. The second CPU only needs to re-render the final image that is changed by the animator, and as a result the photo-realistic preview window can be updated in real time or near-real time. This capability eliminates the need for multiple iterations and waiting for test renders.

The Octane Compression card allows animators to play back a motion test sequence at full resolution on an interlaced composite or S-Video monitor without adding an external RAID or DDR. Motion tests can also be played back at full video resolution on the graphics monitor. Perform 2:3 pull down to preview film sequences at video frame rates. Play back a video sequence in loop mode for synchronizing the animation, all without compromising the system performance. Compressed or full-resolution uncompressed video can also be texture mapped in real time and then inserted into a scene to provide extremely rich detail without extensive animation. Items such as blinking control panels, video billboards, and background settings can now be included as video segments textured onto polygons--all rendered in real time.

The O2 workstation offers many of these capabilities at a new low price point. With an available R10000 processor for high-speed rendering and interaction, hardware-accelerated textures, and integrated video I/O, the O2 workstation delivers high-end desktop features to every user. As the next generation of animation software becomes available, the O2 workstation with its powerful processors will allow animators to perform advanced functions such as particle animation or inverse kinematics on a machine that costs the same as a high-end PC. The built-in compression engine on the O2 workstation can be used for motion tests and previewing video clips. Both O2 and Octane have third-party software solutions that respond to RS-422 deck control commands. This allows animators to use their computers as virtual tape decks, further integrating the SGI solution into the production process.

With an Octane or O2 workstation, an animator can create an animated sequence, composite into a video shot, and edit the final piece using an uncompressed nonlinear editor--all on one platform.

New Capabilities in Game Development

The graphics and processing performance of both the Octane and O2 workstations are allowing game developers to change their workflows, bringing new titles to market faster. Now games can be prototyped and played with fully textured, real-time 3D characters. This allows developers to thoroughly evaluate the game concept without writing any game code. Once a concept has been approved, 3D models can be created using the leading-edge modeling and animation tools described above.

As artists create scenes and models, games can be automatically coded and tested using next-generation development tools from our third-party software providers. This allows game developers to bridge the gap between artists and programmers. In addition, these advanced new tools can be used to take a single game and compile code for multiple game platforms, further reducing time-to-market.

As game developers do more and more film work, the high-bandwidth architectures of the O2 and Octane workstations allows users to readily manipulate film resolution images. The Octane graphics system can support a 1Kx2K image, making real-time film resolution preview an emerging capability on the desktop. Also, game developers can use the powerful compositing, paint, and editing capabilities to create compelling lead-in videos.

Great Performance Requires Great Balance

To achieve optimum performance, system designers must create an architecture that balances the performance of the processors, graphics accelerators, and data I/O systems. SGI has always built well-balanced systems. This is why application performance is always best on SGI® machines.

Tom Gillis
Octane Product Manager