1994 COMPUTER PRESS ASSOCIATION WINNER ♦ BEST COMPUTER NEWSLETTER

MICROPROCESSOR REPORT

THE INSIDERS' GUIDE TO MICROPROCESSOR HARDWARE

VOLUME 9 NUMBER 8 JUNE 19, 1995

Unified Memory Architecture Cuts PC Cost

Defines the Platform for Next Generation of Mainstream Computers

by Yong Yao

A new wave of PC design, named unified memory architecture (UMA), is quietly sweeping the PC industry. With the unified memory architecture, a PC combines its main memory and its frame buffer in a single physical DRAM array, eliminating the traditional standalone frame buffer—a $30 to $60 saving. Core-logic vendors are rushing to get on the UMA bandwagon, and companies that build video/graphics controllers are either seeking partners or diving into the already crowded core-logic business. System manufacturers are enthusiastic about the new architecture and are eager for the cost advantage of upcoming UMA chip sets.

The potential impact of UMA, however, is far more than just cost savings. A paradigm shift from the 16-year-old PC architecture originated by IBM, UMA will redefine the business model of some chip makers and trigger another round of consolidation in the core-logic and graphics markets. Coupled with a move to get rid of the ISA bus, the unified memory architecture leads to a new level of integration. More functions and features will be packed into the PC without significant price increases, ultimately benefiting the consumer.

UMA Approach Brings Cost Savings

Let us examine the memory requirement for today's mainstream PC displays. An 800 × 600 screen with 8-bit color requires 480K of memory. A 1024 × 768 screen with 8-bit color requires 768K, and the same 1024 × 768 resolution requires 1.5M for 16-bit color. On the other hand, the granularity offered by today's commodity DRAM technology is 256K × 16. Thus, to satisfy the requirement of 480K of memory in a 32-bit-wide graphics subsystem (using two DRAMs), a standalone frame buffer has to be 1M in size, wasting 0.52M of valuable DRAM. By the same token, a half megabyte of DRAM will be wasted if the frame buffer requires 1.5M.

In the unified memory architecture, however, the system can allocate the exact amount of DRAM required for the frame buffer. During its booting sequence, the machine assigns a block of DRAM for its frame buffer, and whatever remains is used for main memory. The size for the frame buffer is dynamically configurable, depending on the choice of display resolution and color depth. For instance, 480K will be assigned for 800 × 600 resolution and 8-bit color.

The memory granularity using the UMA approach is determined by the operating system. For DOS and Windows, the granularity is 64K. This is the maximum possible amount of DRAM wasted. In the above example, 32K of DRAM is wasted. Nothing will be wasted if the display resolution is 1024 × 768, whether it is 8-bit, 16-bit, or 24-bit color.

The cost savings do not come for free. Assume that a PC has 8M of main memory, a 1M separate frame buffer, and a 1024 × 768 display with 8-bit color. Now let its graphics subsystem share the 8M DRAM with its main memory subsystem, eliminating the 1M frame buffer. Since 768K of the 8M must be used for the graphics subsystem, only 7.25M will be available for the OS and applications. If the OS and applications need more than 7.25M of main memory, the 9% shortage of available memory will increase the amount of swapping between main memory and the hard disk, which degrades performance. For a true-color high-resolution display, the memory loss is more significant.

Fortunately, most commercially available graphics applications are written for 8-bit color, and most of today's screens are either 800 × 600 or 1024 × 768. To steal one-half or three-quarters of a megabyte from the 8M main memory generally causes no performance problems. But the $30 savings is substantial in today's mainstream desktop market, where system OEMs negotiate over pennies. This explains why most OEMs are enthusiastic about UMA.

The exact savings are implementation dependent. For performance reasons, the amount of memory for the frame buffer may be added to the main memory. For the case above, this would result in 9M of main memory,

with 8.25M available for OS and applications. Another way to trade the cost savings for enhanced performance is to add an L2 cache. The L2 cache will reduce CPU accesses to DRAM, leaving more memory bandwidth for graphics. Therefore, under the UMA umbrella, there will be different approaches for chip-set design.

Weitek Samples 486 UMA Chip Set

Aiming to bring high-performance 64-bit graphics and multimedia to low-end 486 PCs, Weitek has announced its W464 UMA chip set. Historically, Weitek has not been involved in the core-logic business. It does, however, have graphics expertise and PC market experience. This is a good example of how the unified memory architecture can change a company's business outlook.

As Figure 1 shows, the chip set is composed of two 208-pin PQFP devices that integrate complete core logic for a 486 PC (including Cyrix M1sc), a 64-bit DRAM-based GUI accelerator, and a 135-MHz RAMDAC. The graphics accelerator, Weitek's fourth-generation high-performance design, handles up to 1280 × 1024 resolution and 32-bit color, 256 Windows raster operations (ROPs), and full plug-and-play compatibility. In addition, graphics data is stored using a lossless compression algorithm, reducing the amount of memory needed for the frame buffer.

Weitek has seen the first silicon of its W464 chip set, which is currently in debugging. It is now running DOS and Windows applications. The W464 is expected to be in production by September, making it the industry's first PC UMA chip set that offers a PCI interface.

The Weitek implementation has gone one step further. It eliminates not only the standalone frame buffer but also the separate graphics-controller chip. Compared with most 32-bit 486 chip sets, the W464 may actually offer better performance, due to its 64-bit memory and graphics, while lowering overall system cost.

Since the 464 chip set works only with the 486, it might be too late for the desktop market, which is rapidly moving to Pentium-class CPUs. Given the September production date of the W464, OEMs must decide whether to use it based on the system redesign cost and delayed time to market. Even though Weitek will provide a reference design kit (which has become a common business practice in the PC world) for launching a new design, system manufacturers will still need a certain amount of design effort. Weitek's timing would be better for the 486 notebook market, but the W464 is not a 3.3-V chip set. It also lacks enhanced power management and LCD control logic.

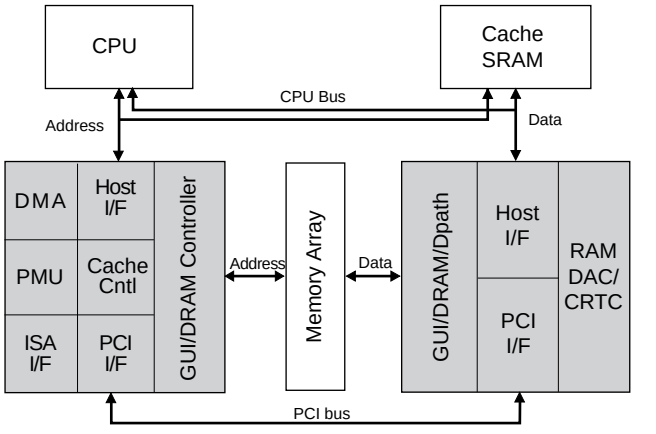

Figure 1. The W464 chip set integrates graphics and core-logic chips that share a single DRAM array.

Weitek is developing a Pentium chip set, named the W564. The experience of the W464 will certainly help Weitek get this Pentium chip set to market more quickly and with less design risk. The W564 will support all Pentium-pinout CPUs *(see **090804.PDF**)*. It will also add several performance enhancements, including Weitek's video-acceleration technology, local-bus EIDE, and advanced DRAM support. The W564 will be a 3.3-V device with 5-V-tolerant I/O.

VLSI Plans a Single-Chip Solution

Like Weitek, VLSI Technology has put the graphics controller into a core-logic UMA chip set. VLSI differentiates its solution by targeting Pentium-class CPUs with a 3.3-V design. VLSI has been designing its desktop UMA chip set, named Coyote, and its notebook UMA chip set, Falcon, in parallel. By incorporating the Intel burst order as well as the linear burst mode, the two-chip sets will support Pentium, the K5, and M1. The projected date for engineering samples is 1Q96 for Coyote and 2Q96 for Falcon. Using a 352-pin BGA package, VLSI designers combine almost all core logic and graphics-control logic into a single chip. Figure 2 shows a system diagram for the Coyote chip set.

The VL82C546 integrates a DRAM controller, PCI interface, graphics accelerator, L2 cache controller, RAMDAC, and distributed DMA controller (see sidebar below). The DRAM control logic supports fast-page-mode DRAM, EDO DRAM, SDRAM, and SGDRAM. By integrating the distributed DMA and serial interrupt, VLSI makes the VL82C542 ISA bridge chip optional. This integration makes it possible to have high-speed yet compatible PCI connections for legacy devices.

VLSI has developed its own graphics technology, called the GraphiCore Architecture, in the past two years. The key elements of GraphiCore are a configurable geometry pipeline and an orthogonal instruction set for single-state execution. The accelerated graphics functions include line draw, clipping, all 256 Windows ROPs, bitBLT, stretchBLT, patternBLT, color expansion, color compression, rectangle fill, and text operations. The graphics accelerator resides on the CPU bus so as to run at the CPU bus frequency.

In contrast to non-UMA approaches, VLSI's solution minimizes chip count, cuts overall system cost, reduces power consumption, and eliminates redundant functions. These features are especially important for notebooks, where board space and power budget are critical. With its parallel development projects, VLSI may well be the first company to bring a UMA notebook chip set to market. This strategy will give VLSI a head start for Pentium-class notebooks.

Opti Partners with Graphics Vendors

Opti has taken a different approach, defining a 20-page specification for the interface between the memory controller and the graphics controller. Figure 3 shows the proposed system. The interface is straightforward. Essentially, it consists of only two signals: MREQ# and MGNT#. Under Opti's specification, the memory controller is the default owner of the DRAM array. The graphics controller must send a request (MREQ#=0) to the memory controller every time it needs to access the DRAM array. It will not gain access until the memory controller responds with an acknowledge (MGNT#=0).

Opti presented this proposal to the VESA committee last month to stimulate a standardization process. Besides the two-signal interface, Opti's specification also includes recommendations on system BIOS, device drivers, and future system compatibility.

Because of the slow process of standardization, Opti has joined with a couple of unnamed graphics vendors to provide a UMA core-logic and graphics chip set. It has also been working with several system houses for early design-ins. Opti expects that its UMA core-logic chip set will be on the market by year-end. The supporting graphics chips should arrive at about the same time.

Compared with VLSI's approach, Opti's solution will have some performance disadvantages due to its interchip arbitration, although some of the arbitration can be overlapped with the DRAM's recharge time. It does, however, have the advantage of being flexible. Designers have flexibility in architecture trade-offs, while OEMs have flexibility when choosing graphics solutions. This approach will be appreciated by those companies having design expertise with either core logic or graphics but not both.

Opti is a major player in the core-logic business. Its business model has always been to pursue volume markets. A fundamental question to be asked before adopting an Opti-like approach, however, is why such an interface protocol is needed in the first place. If the graphics/video controller is integrated into the core-logic chip set, this protocol becomes unnecessary.

PCs are a commodity business: the bottom line for a system OEM is to lower its manufacturing cost while maximizing system performance. To be adopted by PC makers, every new technology must add real value and

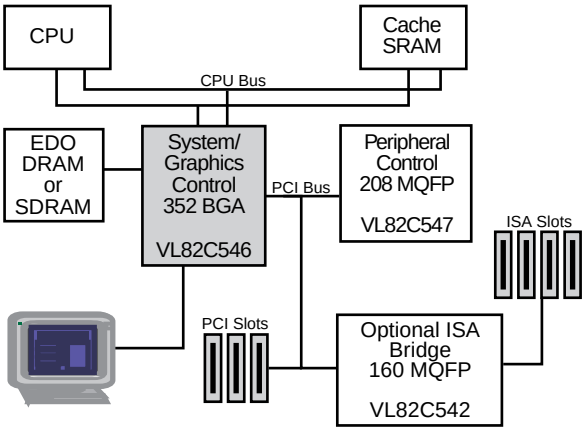

Figure 2. VLSI's Coyote chip set is a single-chip solution for the core logic and graphics functions.

improve the bottom line. For example, a few years ago Intel and ATI tried to establish a standard for interfacing separate graphics and video chips. The standard did not take off because there was no incentive for separating graphics and video, which should naturally fit together. Therefore, any separate UMA graphics approach may be short-lived for the mainstream PCs; the integrated approach will ultimately prevail.

Everyone Wants a Common Standard

Although an Opti-like interface protocol might be short-lived for mainstream PCs , a common specification for BIOS, device drivers, and future system compatibility is extremely important. The entire PC industry will benefit if there is only one UMA standard. All the major UMA players have claimed that they would stand behind a common standard if one appears.

In the past year, Opti, VLSI, and Intel have been the three biggest chip-set suppliers. Of these leaders, only Intel has not said whether it is working on UMA solutions. Its latest chip sets, Triton and Orion, are not

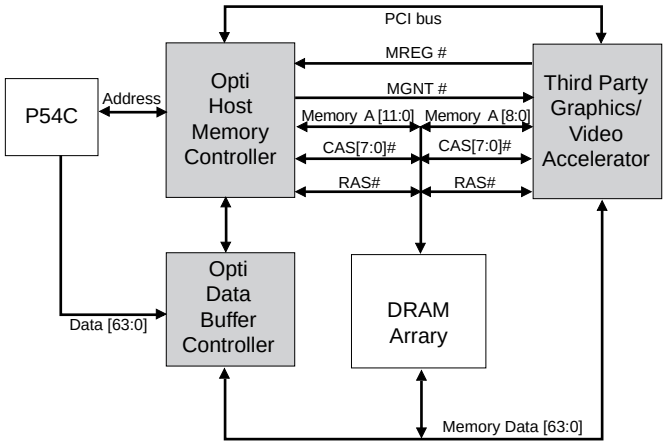

Figure 3. Opti's proposal includes signaling between the memory controller and the graphics controller.

Eliminating the ISA Bus

The ISA bus, introduced in 1981, has long been inadequate for PC performance needs. Most PCs shipped today have two peripheral buses: ISA and PCI. It would be nice if there were only one peripheral bus, but there are two major technical issues—legacy DMA and legacy interrupts—that prevent PC makers from getting rid of the ISA bus.

To support legacy DMA, the PCI bus must address three problems. First, PCI does not have the necessary seven DRQ# and seven DACK# signals. Second, on the ISA bus, the I/O and memory commands are separate signals. Thus, simultaneous I/O and memory transfers are possible on the ISA bus but not on the PCI bus, according to its current definition. The third technical difficulty is that the ISA bus allows long-length I/O transfers (on the order of one microsecond), which will create a major performance limitation for PCI.

In terms of legacy interrupts, the PCI bus has only four sharable IRQs, while the ISA bus supports 11 edge-sensitive IRQs. The common solution is to introduce sideband signals from PCI to ISA. This solution does not allow the elimination of the ISA bus.

Aiming to introduce products without an ISA bus, eight companies, including VLSI, formed a consortium in late 1994 to define a standard to support ISA devices on PCI. The standard, called distributed DMA, is now in its final legal review and will soon be released to the public. Following the distributed DMA standard, chip vendors can integrate ISA support into PCI components. PC system manufacturers will soon be able to build PCs without the ISA bus, resulting in higher integration, better performance, and yet lower cost PCs.

based on the unified memory architecture, but Intel is interested in UMA. No company can ignore the potential UMA impact, including the microprocessor giant. Intel is known to be collaborating with Cirrus Logic on UMA chip sets; Cirrus plans to roll out its UMA graphics controller in the first half of 1996.

Companies like Intel and Cirrus have been taking cautious steps, carefully examining the issues associated with different UMA approaches. PC graphics has some unique characteristics. For instance, more than 90% of CPU cycles to the frame buffer are write cycles; the CPU rarely reads from it. Furthermore, when both the memory subsystem and the graphics subsystem share the same bus bandwidth, what will be the overall performance impact? In this design, it becomes critical to intelligently use the bus bandwidth, and the right kinds of FIFO schemes are helpful. Designers have to be careful that their machines will not hang up when end users switch display resolutions.

Chips and Technologies also is working on its own UMA solution, which may be available before the end of the year. C&T's advantage is its in-house expertise on both graphics and core logic. Even so, C&T would like to work with others to drive one UMA standard, so all UMA chip sets can share the same BIOS and device drivers.

Separate Frame Buffer Has Its Niche

In the PC industry, most technologies or new architectures start at the high end and then migrate to the low end. It seems, however, that the unified memory architecture is evolving the other way around. It will be adopted by low-end PCs first, then by midrange PCs, and may eventually be incorporated into some high-end machines. The unified memory architecture is driven primarily by cost reduction, whereas most other PC architecture changes are initiated to enhance performance.

Even though the unified memory architecture will soon become a dominant PC approach, a graphics controller with a dedicated frame buffer will still have a role for years to come. The desktop market alone is expected to exceed 50 million units in 1996, according to Computer Intelligence InfoCorp. There is enough room for multiple solutions, including the conventional separate frame buffer.

By using new DRAM technologies, such as MDRAM from MoSys, graphics companies such as Tseng Labs and S3 may design their video/graphics accelerators to deliver more performance or to embed specific features that serve power users. For those users, features such as video playback, 3D graphics, and image rendering will soon be mandatory. In addition, video and graphics are still emerging technologies. More advanced approaches are under development. It will be much easier for the standalone graphics controller to take advantage of the latest developments.

With its dedicated frame buffer and specialty memory, a high-end graphics controller should have performance superior to its equivalent UMA solution. For instance, in the UMA design, CPU memory accesses have to be held while the memory bus is busy serving graphics functions. This alone can cause a 10% performance degradation for certain applications.

We recommend that graphics chip vendors give their standalone controllers both a dedicated frame buffer and hooks for interfacing with some of the upcoming UMA chip sets. This is a safe way for a graphics company to reduce risk and maintain its market share.

Technology Revives the UMA Approach

The idea of the unified memory architecture is nothing new. Apple's early Macs and some workstations have been designed this way for years. Perhaps there were no particular reasons why IBM defined its early PCs using a separate frame buffer, although IBM had certain constraints on the availability of off-the-shelf components. In 1979, PCs were using text-mode monochrome dis-

plays; the amount of memory needed for those displays was very small. It would have been easier just to put the frame buffer into main memory.

The non-UMA architecture, however, has been the PC platform of choice since day one. A few companies, such as C&T and Tandy, tried UMA once before but did not go far with it. Why, all of a sudden, has the unified memory architecture become so hot? Various technologies are converging to revive the unified memory architecture in the PC market.

The increasing density of commodity DRAMs has made non-UMA solutions more costly. On the other hand, this same trend will make systems with at least 16M of main memory more popular. Therefore, it makes more sense to steal main memory for the frame buffer.

In addition, main-memory and frame-buffer bandwidths are converging. Exotic high-bandwidth memories for frame buffers are not getting into volume PCs. These low volumes have kept their prices high. For example, VRAM chips have always been two or more times as expensive as commodity DRAMs.

In contrast, affordable new DRAM technologies, such as EDO DRAMs, are substantially increasing main-memory bandwidth, while advanced caching and buffering schemes have kept most CPU memory references off the main-memory bus. In addition, 64-bit data paths, data bursting, and high bus frequencies have greatly reduced the time needed to transfer a given amount of data. Together, these changes have created more bus bandwidth for graphics/video functions.

Today's CPUs have enough computing power to handle jobs like line drawing, image decompression, and graphics rendering. These functions previously relied on special accelerators. A unified memory actually increases CPU graphics performance by making it more convenient for the CPU to access the frame buffer.

BGA packages and high integration offer designers an efficient way of balancing signal I/O and chip gate count. It has become economical to pack the core logic and graphics control into one chip. Although the BGA package costs about one cent more per pin than the PQFP, the BGA package is relatively new, and there is room for BGA prices to drop.

Price & Availability

Samples of the W464 UMA chip set are available now from Weitek in two 208-pin PQFPs. Production is scheduled to begin in September. The 10,000-piece price for the chip set is $43.50. For more information, contact Weitek (Sunnyvale, Calif.) at 408.738.8400; fax 408.739.4374.

Samples of the Coyote chip set will be available in 1Q96 from VLSI Technology, in a single 352-pin BGA. Production is scheduled to begin in 2Q96. Pricing of the chip set will be available in 4Q95. For more information, contact VLSI (Tempe, Ariz.) at 602.752.6481; fax 602.752.6014. For information on the distributed DMA standard, send e-mail to *devoy@tempe.vlsi.com*.

Samples of the Opti UMA chip set will be available in 4Q95 at a $40 sample price. Production is scheduled to begin in 1Q96. For more information on Opti's UMA specification and its chip set, contact Opti (Santa Clara, Calif.) at 408.486.8605; fax 408.980.8860.

Because of the price tag for big monitors, mainstream PCs may still be using 17" or smaller displays for some time to come. Given the limitations of human vision, going beyond the resolution of 1280×1024 does not make much sense for these monitors. Also, human eyes can distinguish only a certain number of colors: 24-bit color space is probably enough. These three facts tell us that the amount of memory needed for the graphics/video subsystem is no more than 4M for mainstream PCs. They also show that the memory bandwidth for supporting the graphics subsystem will not increase dramatically.

In short, technology has arrived at the point where it is inevitable for future mainstream PCs to incorporate the unified memory architecture. Some OEMs will show UMA systems at Comdex next spring. Almost all vendors now working on UMA solutions have plans to extend this architecture to P6-class machines. We project that, 18 months from now, the UMA approach will be dominant for all PCs but high-end desktops. ♦

Yong Yao is the director of MicroDesign Resources' Technology Roadmap service (see ***0908MSB.PDF****).*

